Argument-based validation of the English–Japanese Phrasal Vocabulary Size Test: The role of familiarity in item difficulty

Ayako Aizawa
Rikkyo University, Japan

This study validates the English–Japanese Phrasal Vocabulary Size Test (E–J PVST), a bilingual adaptation of the Phrasal Vocabulary Size Test (Martinez, 2011) assessing receptive knowledge of high-frequency English phrasal expressions among L1 Japanese adults. Using Chapelle et al.'s (2008) argument-based validation framework, evidence is synthesized for the domain definition, evaluation, generalization, and explanation inferences. The test was administered online to 226 adults spanning CEFR A1 to C1. Rasch analyses indicated high item reliability (.97) and person reliability (.88), and provided provisional support for a largely unidimensional structure. However, 12% of items showed underfit and 22% showed overfit, with distractor malfunction identified as a primary source of misfit. Scores increased monotonically across CEFR levels. Learner-reported familiarity was significantly correlated with item difficulty, whereas corpus frequency was not, with more familiar expressions tending to be easier. The study demonstrates that argument-based validation can identify specific strengths and weaknesses in low-stakes vocabulary tests, and highlights the role of subjective familiarity, more than corpus frequency, in predicting item difficulty for multiword expressions.

Keywords: multiword expressions, phrasal expressions, vocabulary testing, argument-based validation

Email address for correspondence: a.aizawa@rikkyo.ac.jp

© The Author(s) 2026. This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits the user to copy, distribute, and transmit the work provided that the original authors and source are credited. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Introduction

Measuring knowledge of multiword expressions

The measurement of vocabulary knowledge serves important purposes for both learners and educators. For learners, vocabulary tests can reveal what they know and do not know, and repeated testing can track vocabulary growth, potentially enhancing motivation (Beglar, 2010). For teachers, vocabulary assessments provide information about learners' proficiency levels that can inform materials selection and instructional planning (Nation & Anthony, 2016). Despite the recognized importance of multiword expressions (MWEs) in second language (L2) proficiency, vocabulary testing research has predominantly focused on single-word knowledge. Standardized measures of MWE knowledge remain relatively scarce, and there is as yet no widely accepted benchmark test for MWEs comparable to the Vocabulary Size Test (VST; Nation & Beglar, 2007) for single words (Gyllstad & Schmitt, 2018).

One notable effort to address this gap is the Phrasal Expressions List (PHRASE List; Martinez & Schmitt, 2012) and its associated Phrasal Vocabulary Size Test (PVST; Martinez, 2011). The PHRASE List is a corpus-derived collection of 505 high-frequency English phrasal expressions, defined as "a fixed or semi-fixed sequence of two or more co-occurring, but not necessarily contiguous, words with a cohesive meaning or function that is not easily discernible by decoding the individual words alone" (Martinez & Schmitt, 2012, p. 304). The list was compiled from the British National Corpus (BNC) and organized into five frequency bands of approximately 100 expressions each, with items spanning six structural types, including adverbial phrases (53% of the list), verb phrases (32%), and others. Examples include expressions such as *as a result* (meaning *consequently*), *to do with* (meaning *related to*), and *give up* (meaning *stop trying*), whose meanings cannot be straightforwardly derived from the individual component words.

The Phrasal Vocabulary Size Test

The PVST was developed by Martinez (2011) to estimate learners' receptive knowledge of the phrasal expressions in the PHRASE List. The test comprises 50 items, 10 from each of the five frequency bands, restricted to adverbial and verb phrases. Each item presents

a target phrasal expression within a minimal, non-defining context sentence, followed by four multiple-choice options written in English. The format parallels that of the VST (Nation & Beglar, 2007), and test scores are intended to represent the proportion of phrasal expressions from the PHRASE List that a learner knows. Because the PVST samples only adverbial and verb phrases, test scores most directly index knowledge of these two phrase types within the PHRASE List rather than knowledge of all MWEs.

Martinez (2011) conducted an initial validation of a 75-item prototype of the PVST with over 2,200 L1 Austrian German learners at an estimated CEFR B2 level of English proficiency. The participants were divided into three groups, each completing a different 30-item form of the test. Classical test theory analyses confirmed that the items measured a common construct, written receptive knowledge of phrasal expressions, and that performance was consistent across the three test forms. Based on difficulty, discrimination, and validity indices, 50 items were selected for the final version of the PVST.

Research on single-word vocabulary tests has demonstrated several advantages of presenting multiple-choice options in the test-takers' L1 rather than their L2. Bilingual test formats eliminate the need for learners to decode options from L2 to L1, which can lead to more accurate measurement of the target construct (Nguyen & Nation, 2011). L1 options can also avoid the inaccuracy and redundancy that arise when answer choices must be written using only high-frequency L2 vocabulary (McLean et al., 2016). Additionally, L1 options have been shown to reduce test anxiety (Elgort, 2013) and shorten response time, thereby reducing fatigue and facilitating test completion (McLean et al., 2016; McLean et al., 2020). Bilingual versions of the VST have been developed for multiple languages and have generally produced reliable results that separate participants according to their English proficiency (e.g., Nguyen & Nation, 2011; Park, 2024). If similar benefits extend to tests of MWEs, a bilingual adaptation of the PVST could provide a more valid and accessible measure of phrasal expression knowledge, particularly for learners at lower proficiency levels. These advantages motivate examining whether the English–Japanese Phrasal Vocabulary Size Test (E–J PVST) can improve accessibility without compromising measurement quality. The present study therefore introduces and

provides validation evidence for the English–Japanese Phrasal Vocabulary Size Test (E–J PVST), a Japanese-translated version of the PVST developed for this research.

Literature review

Factors influencing item difficulty

An important aspect of validating any vocabulary test is understanding what factors drive item difficulty. For size tests organized by frequency bands, the expectation is that higher-frequency items will be easier and lower-frequency items more difficult. Beglar's (2010) Rasch-based validation of the VST found that item difficulty generally mirrored word frequency bands for Japanese learners, supporting the construct validity of the frequency-based organization. However, other research has questioned how strictly frequency predicts difficulty, particularly beyond the most frequent vocabulary. Schmitt et al. (2021) observed that while frequency ranking works relatively well for broad vocabulary bands up to the 5,000 level, it does not reliably predict the knowledge rankings of individual words.

For MWEs, the relationship between frequency and difficulty may be even more complex. Matsuo's (2022) analysis of the PVST revealed notable mismatches between frequency band and item difficulty: the most difficult item in Matsuo's data, *at once*, came from the third 1,000-word band, whereas the easiest item, *come up to*, was from the fifth (least frequent) band. This suggests that factors beyond corpus frequency, such as learner familiarity, semantic transparency, L1 influence, and instructional emphasis, may play substantial roles in determining which phrasal expressions learners know.

Familiarity, a subjective measure of how often learners feel they encounter a given expression, offers a learner-centered perspective on item difficulty. A phrase that is highly frequent in a general English corpus may not be familiar to a particular group of learners if it appears primarily in genres or contexts to which they have not been exposed. Conversely, a moderately frequent expression may be highly familiar if it appears frequently in textbooks or is explicitly taught. Kanazawa (2020) collected familiarity ratings for the phrasal expressions in the PHRASE List (Martinez & Schmitt, 2012) from

Japanese university students, providing data that can be used alongside corpus frequency counts to examine whether objective frequency or subjective familiarity better accounts for variation in item difficulty. The present study takes advantage of these data to investigate the relative contributions of frequency and familiarity to item difficulty on the E–J PVST, an analysis not previously conducted for any version of the PVST, thereby providing evidence relevant to the explanation inference within the argument-based validation framework.

Argument-based validation framework

Previous validation studies of the PVST (Martinez, 2011; Matsuo, 2022) lacked a systematic theoretical framework for organizing and evaluating validity evidence. The present study addresses this limitation by adopting the argument-based validation framework proposed by Chapelle et al. (2008). This framework builds on the work of Messick (1989) and Kane (2006) to structure validity evidence as a chain of six inferences: domain definition, evaluation, generalization, explanation, extrapolation, and utilization. The present study focuses on the first four inferences, which concern the internal validity of the test itself. The extrapolation and utilization inferences, which address the relationship between test scores and real-world language use and the appropriateness of score-based decisions, respectively, are beyond the scope of this initial validation because the E–J PVST is currently used for research purposes, and no criterion measures of naturalistic phrasal expression processing were available. While this approach has seen increasing use in language assessment research (Dursun & Li, 2021), its application to vocabulary testing remains limited.

Applying argument-based validation to MWEs tests raises challenges that differ from those in single-word vocabulary assessment. First, the construct of knowledge of a phrasal expression is often less clearly bounded, because learners may recognize constituent words without knowing the holistic phrasal meaning, creating a continuum of partial knowledge. Second, phrasal expressions vary along multiple dimensions such as frequency, compositionality, semantic transparency, and L1 congruence, which may independently influence item difficulty and reduce the predictability of frequency-banded designs. Third, the domain definition inference requires justification not only for which

expressions are included but for what type of MWEs the test claims to represent, given ongoing taxonomic debates (Wood, 2020). The present study addresses these challenges through Rasch analysis and frequency-familiarity modeling.

The present study

The present study seeks to validate the E–J PVST for measuring receptive knowledge of English phrasal expressions among adult L1 Japanese learners of English. The E–J PVST retains the same 50 items and format as the original PVST but presents the multiple-choice options in Japanese rather than English. Using the argument-based validation framework of Chapelle et al. (2008), the study addresses the following research questions (RQs):

RQ1: To what extent is the E–J PVST a reliable and valid measure of receptive knowledge of English phrasal expressions?

RQ2: To what extent do item difficulty estimates on the E–J PVST correlate with the corpus frequency and the learner-reported familiarity of the phrasal expressions?

Methods

Participants

The test was taken by 226 participants (128 female, 96 male, 2 undeclared), ranging in age from 18 to 63 years (M = 29.0, SD = 11.8). They were recruited through a combination of personal contacts, announcements in university classes, and social media postings. The sample was intentionally diverse in occupation: 102 were office workers, 81 were students (high school: n = 2, undergraduate: n = 65, postgraduate: n = 14), 12 were schoolteachers, and 31 held other occupations (e.g., public employees, part-time workers). This diversity was a deliberate design choice, as previous PVST validation studies (Martinez, 2011; Matsuo, 2022) relied exclusively on university student samples, which may not represent the broader population of English learners.

Participants self-reported their English proficiency level on the Common European Framework of Reference (CEFR) scale, with corresponding scores on well-known standardized tests (TOEFL iBT, IELTS, TOEIC Listening & Reading, and Eiken) provided as reference points (see Table 1). The distribution of self-reported CEFR levels was as follows: A1 (1%), A2 (8%), B1 (28%), B2 (28%), C1 (26%), and unspecified (9%). The relatively small proportion of beginner-level participants likely reflects a self-selection effect: volunteers who agree to take an English test tend to have at least some English learning background. Those who had never taken any standardized English test were given the option of selecting *uncertain*.

Table 1. CEFR bands and the corresponding English language test scores

CEFR level	Eiken	TOEIC Listening & Reading	TOEFL iBT	IELTS
C2			114 and above	9.0
C1	Grade 1	945 and above	95–113	7.0–8.5
B2	Grade Pre-1	785–940	72–94	5.5–6.5
B1	Grade 2	550–780	42–71	4.5–5.0
A2	Grade Pre-2	225–545		4.0
A1	Grade 3–5	120–220		

Note. Eiken (Dunlea, n.d.), TOEIC (Tannenbaum & Wylie, 2008), TOEFL (TOEFL, n.d.), IELTS (IELTS, n.d.).

Comparison with previous studies

The participant profile in the present study differs from those in Martinez (2011) and Matsuo (2022) in several important respects. Martinez's sample consisted exclusively of L1 Austrian German university students at an estimated CEFR B2 level. Matsuo's participants were L1 Japanese female English majors in their first or second year of university (mean TOEIC scores corresponding to CEFR A2–B1), along with 11 international students who were also L2 English users. The present study includes both students and non-students, spans a wider range of English proficiency (CEFR A1 to C1), and controls the L1 variable by including only L1 Japanese participants. This broader sampling addresses the concern that vocabulary testing research often over-relies on convenience samples of university students (Schmitt, 2010), whose test performance may not generalize to the wider population of language learners.

Instruments

Personal background questionnaire

A short questionnaire collected demographic information and self-assessed English proficiency. Participants provided their name (or a pseudonym), age, gender, occupation, and English proficiency level. The CEFR scale was used as the proficiency index, with corresponding scores on TOEFL iBT, IELTS, TOEIC Listening & Reading, and Eiken provided to help participants identify their approximate level (Table 1). All questionnaire materials were presented in Japanese, as using participants' L1 for test instructions has been shown to increase the probability of clear understanding (McLean & Kramer, 2015).

The English–Japanese Phrasal Vocabulary Size Test

The E–J PVST was created by translating the four English multiple-choice options for each of the 50 PVST items into Japanese. The translation was carried out independently by two L1 Japanese postgraduate students majoring in applied linguistics. To ensure accuracy, corpus-based English–Japanese dictionaries (Minamide & Nakamura, 2022; Takebayashi, 2002) were used as primary references. The inter-translator agreement rate was .88, and all disagreements were resolved through discussion. The resulting translations were reviewed by a professor of applied linguistics, who suggested minor adjustments. The translated test was pilot-tested with 11 participants: seven undergraduate and four postgraduate students with varying English proficiency levels (A2: n = 1; B1: n = 3; B2: n =4; C1: n = 3). No translation problems were identified during piloting. Each item on the E–J PVST presents a target phrasal expression within a minimal English context sentence (identical to the original PVST), followed by four multiple-choice options written in Japanese. The following example illustrates the format of the original PVST alongside the E–J PVST. The complete E–J PVST is provided in the Appendix.

Original PVST:
1. go on: It will go on.
a. sleep
b. repeat
c. be fast
d. continue

E–J PVST:
1. go on: It will go on.
a. 寝る
b. 繰り返す
c. 速い
d. 続く

Frequency and familiarity data

Two types of data were used for the analysis of factors influencing item difficulty. First, corpus frequency data, the number of occurrences of each phrasal expression in the BNC, were obtained from the PHRASE List (Martinez & Schmitt, 2012). Second, learner-reported familiarity ratings were drawn from Kanazawa (2020), who collected subjective judgments for the phrasal expressions in the PHRASE List. These ratings were obtained from a separate sample of L1 Japanese university students using a 7-point scale ranging from 1 (*I do not encounter the item at all*) to 7 (*I encounter the item very often*). These data reflect how frequently learners perceived themselves to encounter each expression, regardless of objective corpus frequency. Because these ratings were based on a specific student population, they may not fully represent the familiarity patterns of the non-student participants in the present study.

Familiarity ratings from a separate population were used rather than collected from the same participants for two reasons. First, asking participants to rate the familiarity of each phrase immediately before or after taking the test would risk reactivity effects, with the rating task plausibly influencing test responses or vice versa. Second, Kanazawa's (2020) ratings were obtained from a substantially larger sample (n = 501) than was feasible in the present study, providing more stable estimates of population-level familiarity. The trade-off is that Kanazawa's sample consisted exclusively of university students, whereas the present sample is more occupationally diverse, a limitation noted in the Discussion.

Procedure

Participants received a link to a Google Forms survey that presented the research information sheet, a consent form, the background questionnaire, and the 50 E–J PVST

items. Participants could take the test at a time and place of their choosing, using their smartphone, tablet, or personal computer. Because the test was administered remotely and without supervision, participants were instructed to complete it in a single sitting without consulting dictionaries or other external resources. While unsupervised administration raises potential concerns about data quality (Drown et al., 2023), it also reflects realistic conditions for low-stakes diagnostic assessment and enabled recruitment of a more diverse sample than would have been possible with in-person administration alone. Participants were instructed to skip items if they did not know the answer.

Data analysis

Rasch analysis (RQ1)

For RQ1, Rasch analysis was performed using Winsteps 5.6.0 (Linacre, 2023a) to evaluate the measurement properties of the E–J PVST. Rasch analysis was chosen because it provides a unified framework for examining multiple aspects of test quality simultaneously: item and person reliability, separation indices, item fit, dimensionality, and test precision. Unlike classical test theory, which yields sample-dependent item statistics, Rasch analysis produces item difficulty and person ability estimates on a common interval scale, enabling direct comparison between items and test-takers (Bond et al., 2021).

The analysis evaluated four aspects of the E–J PVST using Rasch measurement:

(a) *Reliability and separation*. Following Linacre (2023b), thresholds for item reliability (> .90) and separation (> 3.00) were used to confirm the stability of the item difficulty hierarchy. For person reliability (> .80) and separation (> 2.00), these values indicated sufficient discrimination across ability levels.

(b) *Item fit*. Infit and Outfit mean square (MNSQ) statistics were assessed using an acceptable range of 0.75 to 1.30 (Bond et al., 2021; McNamara, 1996). Both statistics were examined to ensure comparability with Matsuo (2022) and to follow recommendations by Mizumoto and Takeuchi (2023). Item-level misfit was reported on the basis of Outfit MNSQ, which is more sensitive to unexpected responses far from an item's difficulty (e.g.,

guessing on items above ability, or slips on easy items; Linacre, 2023b). For a low-stakes vocabulary test with possible guessing among lower-ability test-takers and possible slips among higher-ability test-takers, prioritizing Outfit MNSQ was therefore considered diagnostically informative. Item-measure correlations were also monitored for negative values.

(c) *Dimensionality*. Unidimensionality was evaluated via a principal component analysis (PCA) of standardized residuals. A disattenuated correlation above .70 between item clusters was interpreted as evidence of a single underlying construct (Boone & Staver, 2020).

(d) *Test precision*. A Wright map was used to inspect the alignment between item difficulty and person ability, identifying any gaps or potential floor and ceiling effects.

Finally, a one-way ANOVA with Bonferroni-corrected post hoc comparisons examined score differences across self-reported CEFR levels.

Correlational analysis (RQ2)

For RQ2, Spearman's rank correlation coefficients and Pearson's product–moment correlation coefficients were calculated to examine the relationships between item difficulty (Rasch logit measures, centered at the item mean), BNC corpus frequency (Martinez & Schmitt, 2012), and learner-reported familiarity (Kanazawa, 2020). Spearman's rho is reported for comparison with prior research that used proportion correct as the difficulty index and because familiarity ratings were obtained on a 7-point Likert scale; Pearson's r is reported because Rasch logits provide interval-scaled estimates suitable for parametric analysis (Bond et al., 2021).

To further explore the predictive value of frequency and familiarity, simple regression analyses were performed with item difficulty as the dependent variable and, separately, frequency and familiarity as predictors. Assumptions of linearity, normality of residuals, homoscedasticity, and independence were examined for each regression. Items with large positive residuals (i.e., harder than predicted) and large negative residuals (i.e., easier than predicted) were identified and examined qualitatively. Large residuals were defined

as those with absolute standardized residual exceeding 2. Table 2 summarizes the alignment between each validity inference of the argument-based framework and the Rasch indices used as supporting evidence in the present study.

Table 2. Mapping of argument-based validity inferences to Rasch indices in the present study

Inference	Rasch indices and evidence used in the present study
Domain definition	Item selection from the PHRASE List frequency bands; coverage of adverbial and verb phrases
Evaluation	Item fit (Infit/Outfit MNSQ); distractor analysis (mean person ability per option); item difficulty hierarchy; dimensionality (PCA of residuals); local independence (Yen's Q3)
Generalization	Item reliability and separation; person reliability and separation; sample diversity across CEFR levels
Explanation	Frequency–familiarity correlations with item logits; regression of logits on frequency and familiarity

Results

Overall test performance

Descriptive statistics for the E–J PVST by self-reported CEFR level are presented in Table 3. The overall mean score was 69% (*SD* = 17%), indicating that the test was of moderate difficulty for this sample. Scores ranged from 30% to 98%, suggesting that the test captured a wide range of phrasal expression knowledge. The A1 and A2 groups were combined for analysis due to the small number of A1 participants (n = 2).

Table 3. Descriptive statistics for the overall results

CEFR Level	n	M [95%CI]	SD	Min	Max
C1	59	86% [84%, 88%]	7%	66%	96%
B2	64	73% [69%, 77%]	13%	36%	98%
B1	63	59% [55%, 62%]	12%	36%	88%
A1 and A2	19	48% [44%, 50%]	6%	38%	70%
Total	226	69% [66%, 71%]	17%	30%	98%

Note. 21 participants did not specify a CEFR level and are included in the Total row only.

Mean scores showed a clear ascending pattern across CEFR levels: A1+A2 (*M* = 48%), B1 (*M* = 59%), B2 (*M* = 73%), and C1 (*M* = 86%). A one-way ANOVA confirmed that the effect of CEFR level on E–J PVST scores was statistically significant, $F(3, 201) = 99.88$, $p < .001$, $\eta^2 = .60$. Post hoc comparisons using the Bonferroni correction indicated that all pairwise

differences between CEFR levels were statistically significant (all *p*s < .001), with large effect sizes (Cohen's d ranging from 1.10 to 3.67; see Table 4). The large effect size for the overall ANOVA (η^2 = .60) and the monotonic increase in scores across proficiency levels provide preliminary evidence that the E–J PVST is sensitive to differences in English proficiency.

Table 4. *Post hoc pairwise comparisons of E–J PVST scores across CEFR levels*

Level	Comparison	*T*	*d* [95%CI]	*p*
C1	B2	4.18	1.21 [0.70, 1.71]	< .001
	B1	14.19	2.57 [1.98, 3.16]	< .001
	A1 and A2	13.90	3.67 [2.81, 4.52]	< .001
B2	B1	7.70	1.37 [0.86, 1.87]	< .001
	A1 and A2	9.42	2.46 [1.69, 3.23]	< .001
B1	A1 and A2	4.18	1.10 [0.38, 1.81]	< .001

The item reliability estimate was .97, indicating very high reproducibility of the item difficulty ordering across different samples of comparable test-takers. The person reliability estimate was .88, indicating high reproducibility of person ability estimates across different sets of comparable items. As a reliability estimate of 1.00 represents perfect reproducibility, both values can be considered excellent.

The item separation index was 5.98. Because the item separation index exceeded 3.00 and item reliability exceeded .90, the criteria proposed by Linacre (2023b) were met, confirming that the sample size was sufficient to establish a stable item difficulty hierarchy. The value of 5.98 further suggests that the items can be statistically separated into approximately five or six distinct difficulty levels**.** Although this number coincides with the five frequency bands from which the items were drawn, the empirical alignment between Rasch difficulty and frequency band was weak (see RQ2); the separable strata should therefore not be interpreted as reflecting the frequency-band structure. The person separation index was 2.67, exceeding the threshold of 2.00 recommended by Linacre (2023b) for adequate discrimination among test-takers. This value indicates that the E–J PVST can distinguish approximately two to three distinct proficiency strata among test-takers.

Overview of misfit

Although 34% of items met at least one misfit criterion, the most consequential concern was underfit, observed in 12% of items and indicative of potential measurement disturbance. The larger set of overfitting items primarily suggests redundancy in a 50-item test and points to opportunities for efficiency gains rather than invalid measurement. Item-measure correlations were positive for all 50 items, indicating that no item's response pattern conflicted with the overall latent variable. This finding provides basic evidence that all items contribute positively to the measurement of the construct.

The 50-item analysis included one extreme item, Q21 (*give up*), which was answered correctly by all 226 test-takers. This extreme score pattern means that conventional fit and discrimination statistics for this item should be interpreted with caution. Nevertheless, Q21 was retained because it represents a core, very high-frequency phrasal expression and helps preserve content coverage and test-taker engagement.

To provide a more nuanced understanding of the misfit items, they were classified into distinct categories based on the nature and probable cause of the misfit. Five items showed underfit on the Outfit MNSQ. Underfit indicates that the response patterns for these items were more random than the Rasch model predicts, which can occur when items are ambiguous, when distractors do not function properly, or when the item measures something other than the intended construct. Two items showed unexpected response patterns when non-response was examined. For Q17 (*as soon as*), one test-taker who skipped the item had a person ability estimate of 2.11 logits, higher than the mean ability (1.22 logits) of the 223 test-takers who answered correctly. For Q32 (*next door*), two test-takers who skipped the item had a mean ability of 1.88 logits, compared with 1.41 logits for the 181 test-takers selecting the correct option. Among the three active distractors for Q32, mean abilities were 0.72 (n = 28), 0.52 (n = 11), and 0.01 (n = 4), all below the key, indicating that the distractors themselves functioned in the expected direction. These patterns therefore reflect honest non-response by a small number of higher-ability participants, who were explicitly instructed to skip items when uncertain (see Procedure), rather than distractors that attracted higher-ability test-takers away

from the correct answer. Q4 (*at all*), Q15 (*for instance*), and Q50 (*over time*) showed relatively moderate underfit and may reflect minor ambiguity in the English–Japanese translations or context-dependent interpretation of the distractors. For instance, in Q4 (*at all*), the correct option 全然 (*zenzen*, "(not) at all") functions as a negative-polarity intensifier in Japanese, whereas all three distractors are frequency or time adverbs, namely いつも (*itsumo*, "always"), はじめは (*hajime wa*, "at first"), and 時々 (*tokidoki*, "sometimes"). Test-takers who read "at all" as carrying a frequency or temporal nuance (e.g., interpreting "I don't like it at all" as "I never like it") may have been drawn to one of these distractors rather than to the intensifier 全然 (*zenzen*).

The distractor efficiency analysis revealed several patterns relevant to item quality. Eight items (Q4, Q6, Q10, Q14, Q17, Q21, Q34, Q48) had an actual equivalent number of effective options less than four, indicating that at least one distractor was never selected. Items with non-functional distractors are effectively operating with fewer options than intended, which reduces the precision of measurement and may inflate correct response rates for low-ability test-takers.

When test-takers were divided into three proficiency groups (upper, middle, lower) based on their E–J PVST total scores, the proportion answering each item correctly was consistently in the expected order (upper > middle > lower) for all 50 items. This ordering provides evidence that the items discriminate in the intended direction. However, for 12 items, all test-takers in the upper group answered correctly, which is consistent with two interpretations: the distractors did not function effectively for high-proficiency test-takers, or, because the PVST samples from a curated list of high-frequency phrasal expressions, these items were genuinely known by all higher-ability participants. Because the PVST is designed as a size test estimating the proportion of phrasal expressions a learner knows, a ceiling pattern on very easy items is partly expected and reflects the intended content coverage. For two items, Q16 (*as a result*) and Q17 (*as soon as*), the distractors drew very few selections across all proficiency groups. For Q16, two of 226 test-takers selected an active distractor (one each on two of the three distractors); for Q17, two test-takers selected one distractor and one additional test-taker skipped the item. Given the small numbers involved and the high frequency of these expressions, the

limited distractor selection is consistent with genuine knowledge of the items by most of the sample rather than with systematic distractor failure.

Comparison of overfit and underfit: Implications

It is important to note that overfit and underfit have different implications for test validity. Underfit items are more detrimental because they introduce measurement noise: they may be measuring something other than the intended construct, or their scoring may be compromised by ambiguous options. The five underfit items identified in the present study may warrant revision, with particular attention to distractor plausibility and the accuracy of Japanese translations. Overfit items, while less harmful, suggest redundancy that could be addressed by either removing items with overlapping difficulty levels or replacing them with items that provide more unique measurement information. Both types of misfit point to the need for targeted item revision rather than wholesale reconstruction of the test.

Distractor functioning

A systematic check of distractor mean ability was conducted across all 50 items, comparing the mean Rasch logit of test-takers selecting each option with that of the correct key (Linacre, 2023b). For every item, the expected Rasch ordering held: the correct key showed the highest mean ability, and all three active distractors showed lower mean abilities. The few unexpected reversals reported in the *Overview of Misfit* (Q17, Q32) involved test-takers who skipped the item rather than those who selected an active distractor, and therefore they do not represent violations of the Rasch principle, which concerns distractor selection rather than non-response. For Q16 (*as a result*), the two test-takers who each selected an active distractor had a mean ability of 1.46 logits, only marginally above the mean for correct-option selectors (1.27 logits), a difference unlikely to indicate a systematic distractor flaw. Across the test as a whole, distractors functioned in the expected direction, providing supporting evidence for the evaluation inference.

Dimensionality

The disattenuated correlation between person measures on items in clusters 1 and 3 of the first PCA contrast was .73. The Rasch dimension accounted for 31.0% of the total raw variance (against 31.6% expected by the model), and the eigenvalue of the first contrast was 2.19, accounting for 3% of the total variance and 5% of the unexplained variance. This pattern provides provisional support for unidimensionality, although the first-contrast eigenvalue slightly exceeds the commonly cited benchmark of 2.0. According to the criteria proposed by Boone and Staver (2020), a disattenuated correlation above .70 indicates a strong correlation between the contrasting item clusters, suggesting that the clusters are not measuring distinct constructs. This finding supports the conclusion that the E–J PVST measures a single construct, written receptive knowledge of English phrasal expressions, despite the presence of misfit items.

The relatively modest value of .73 (compared to a perfect correlation of 1.00) may reflect the influence of the misfit items or minor secondary dimensions such as knowledge of constituent words, familiarity with specific phrase types, or sensitivity to distractor plausibility. Nevertheless, the value exceeds the threshold for unidimensionality, and no strong evidence of multidimensionality was found.

Local independence

Local independence was further assessed using Yen's (1993) Q3 statistic, calculated as the Pearson correlation between standardized residuals across all item pairs. Of the 1,225 item pairs, the mean |Q3| was .064, indicating very low residual association overall. The maximum |Q3| was .305 (Q14 *in particular* / Q15 *for instance*), and 21 pairs exceeded the conventional threshold of |Q3| > .20 (Christensen et al., 2017). Inspection of these pairs showed that a substantial portion involved Q21 (*give up*), the extreme item with a perfect score, for which residual correlations are inflated as an artifact of zero response variance. Setting Q21 aside, the remaining flagged pairs typically involved semantically or structurally adjacent items (e.g., *in particular* and *for instance* as analogous discourse markers). On balance, the pattern is consistent with substantial but not perfect local independence, with semantic overlap among a small subset of translated items as the

most plausible source of the small remaining residual association. This is consistent with the recommendation of careful distractor design discussed earlier.

Test precision

The Wright map (Figure 1) provides a visual representation of the alignment between person abilities (left side) and item difficulties (right side) on a common logit scale. Several observations are relevant to evaluating test precision.

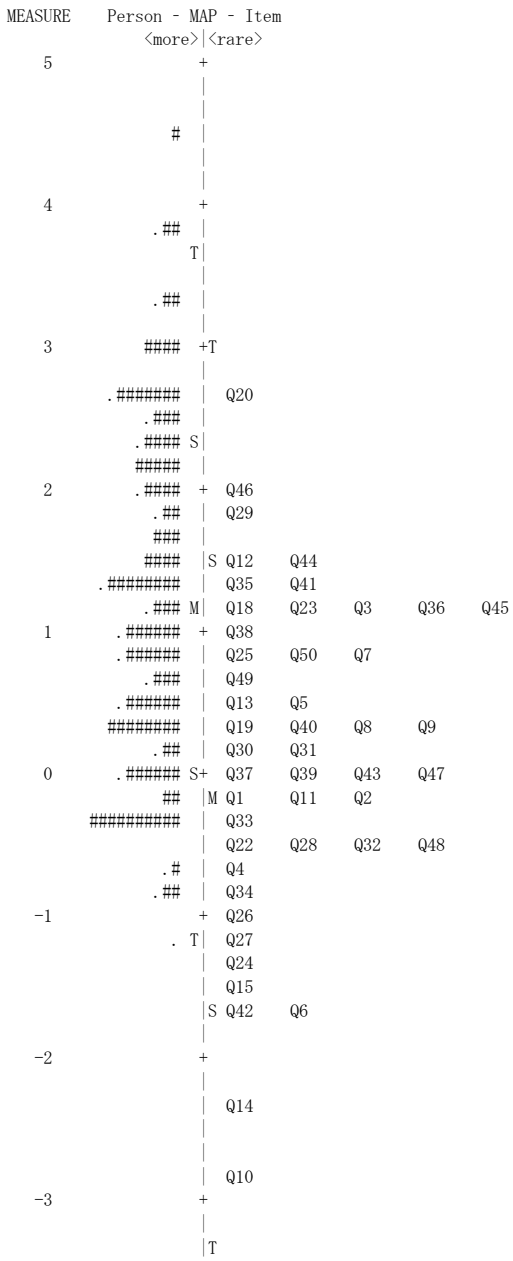

Figure 1. Wright map of the E–J PVST

First, regarding the distribution of items, the most difficult item was Q20 (*be expected to*), with a logit difficulty of approximately 2.75, which only 24% of test-takers answered correctly. The easiest item was Q21 (*give up*), which was answered correctly by all test-takers. Item difficulties ranged from approximately −5.00 to 2.75 logits, spanning a substantial range. Second, the distribution of persons shows that most test-takers clustered in the upper portion of the map (roughly 0 to 3 logits), while most items were located in the lower portion (roughly −2 to +1 logits). This mismatch indicates that the test was somewhat easy for this sample on average: many test-takers had ability estimates that exceeded the difficulty of most items. Seven test-takers had person measures more than two standard deviations above the mean**.** For these test-takers, the test does not provide precise psychometric measurement, in the sense that few items are sufficiently difficult to distinguish among them. With respect to coverage of the PHRASE List itself, however, the result is interpretable as evidence that these test-takers know most items in the list, which is consistent with the test's intended use as an estimate of phrasal expression knowledge. Third, in the medium difficulty range, multiple items overlapped in difficulty at six difficulty levels, with three or more items clustering at similar logit values. While some overlap is normal, excessive clustering means that these items are providing redundant measurement information. Fourth, no major gaps in item difficulty were observed in the range where test-takers were located. Gaps were visible at the bottom of the map (below −2 logits), but because no test-takers had ability estimates in this range, these gaps had no practical impact on measurement precision. Fifth, no floor or ceiling effects were observed: no test-taker scored 0% or 100% on the E–J PVST. This indicates that the test contains items of sufficient difficulty to challenge even the most proficient test-takers (at least minimally) and items easy enough that even the least proficient test-takers can demonstrate some knowledge.

Comparison with previous studies

The overall results were compared descriptively with Martinez (2011) and Matsuo (2022) to provide context for the observed score levels. Because raw data were unavailable for both studies and the instruments and samples differed in multiple respects, formal statistical comparisons were not conducted.

Martinez (2011) reported that Austrian German university learners at an estimated CEFR B2 level (n = 2,204 across three pilot versions of the PVST prototype) achieved mean scores in the range of 67–76% (Version A: M = 75.6%, SD = 17.7%), similar to the present study's CEFR B2 group (M = 73%, SD = 13%), despite differences in L1 background and response options. This suggests that providing L1 options does not necessarily inflate scores at this proficiency level, although the comparison remains tentative given design differences. Matsuo (2022) validated the 50-item PVST with female university students majoring in English (n = 119; mean TOEIC scores corresponding to CEFR A2–B1) and reported good Rasch model fit and unidimensionality, though total raw scores were not reported in percentage terms. The present study's overall mean (69%) should be interpreted alongside differences in sample composition and proficiency distribution. These cross-study comparisons are intended as contextual benchmarks rather than confirmatory extrapolation evidence.

Correlations between difficulty, frequency, and familiarity

Before examining the predictors of item difficulty, the relationship between corpus frequency and learner familiarity was assessed. Across the full 505-item PHRASE List, there was a moderate positive correlation between familiarity rank and BNC frequency rank (Spearman's ρ = .47, 95% CI [.40, .54], p < .001). When restricted to the 50 PVST items, the correlation was slightly stronger (ρ = .53, 95% CI [.33, .73], p < .001). These findings indicate that frequency and familiarity are related, with more frequent phrases tending to be more familiar, but the two measures still capture distinct aspects of learner experience with phrasal expressions, justifying their separate examination as predictors of item difficulty. Having confirmed that frequency and familiarity are related but distinct, the next analyses examined how each predicted item difficulty on the E–J PVST.

Across the 50 PVST items, item difficulty (Rasch logit measures, centered at the item mean) correlated weakly and non-significantly with BNC frequency count (Spearman's ρ = −.21, 95% CI [−.48, +.09], p = .143; Pearson r = −.11, p = .450) and weakly to moderately and significantly with learner-reported familiarity (ρ = −.31, 95% CI [−.59, −.03], p = .031; r = −.38, p = .007). The negative correlations are in the expected direction: items associated with more frequent and more familiar expressions tended to be easier (lower

logits). For comparison, parallel analyses using proportion correct as the difficulty index yielded essentially the same Spearman ρ in magnitude (+.21 and +.31, respectively, with the positive sign reflecting the direction of proportion correct as easiness). The difference between the two correlations was not statistically significant (p = .620); thus, although only the familiarity correlation reached statistical significance, these results do not establish that familiarity is a stronger predictor of item difficulty than corpus frequency. See Figures 2 and 3 for the scatterplots. These figures plot proportion correct (item easiness) for visual comparability with prior research that used this index, whereas the coefficients reported above are based on Rasch logit difficulty; because Spearman's ρ is rank-based, its magnitude is identical under either index, and the upward trend in the figures corresponds to the negative logit-based coefficients (more frequent and more familiar expressions are easier).

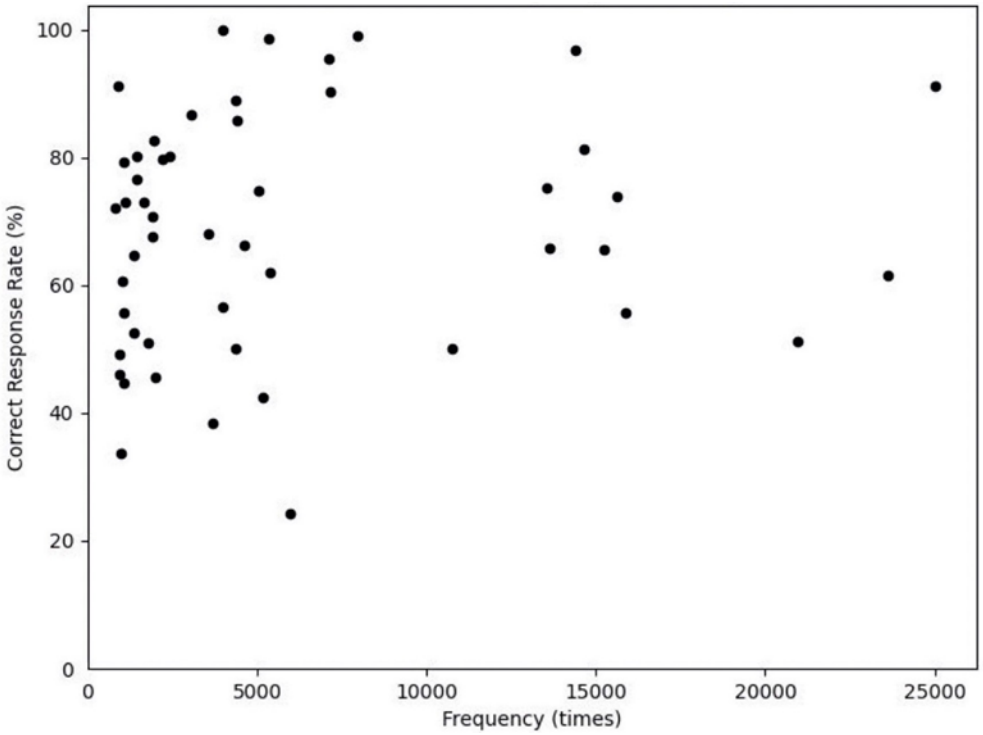

Figure 2. Correlation between item difficulty (proportion correct) and BNC frequency

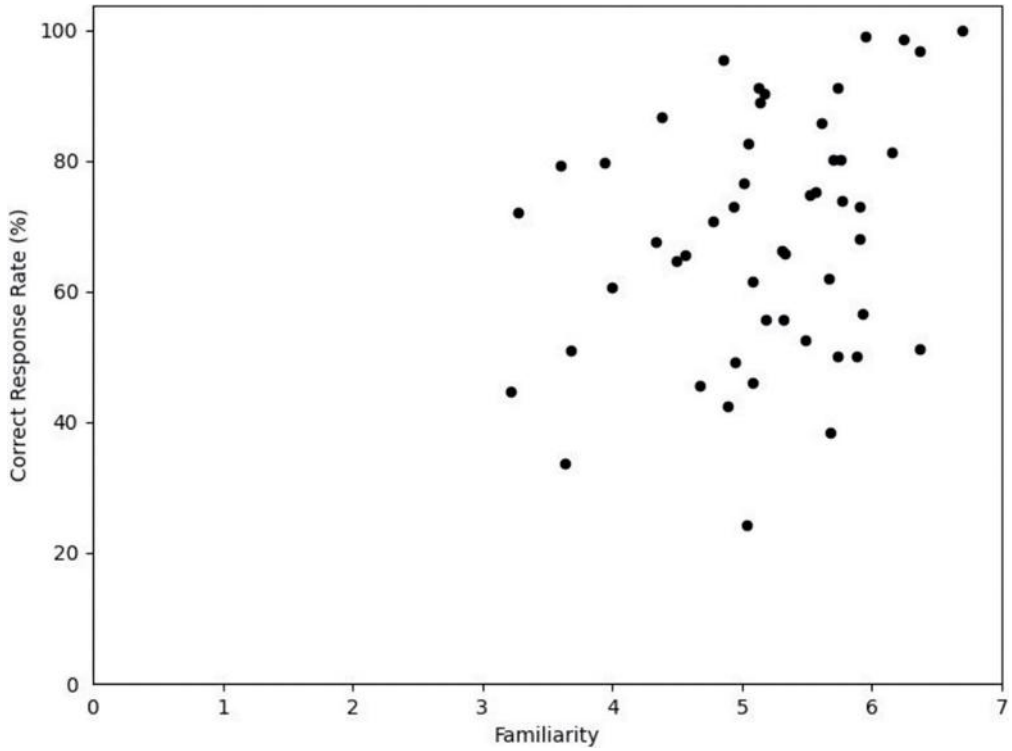

Figure 3. Correlation between item difficulty (proportion correct) and Japanese learners' familiarity ratings (Kanazawa, 2020)

Regression analysis and residual examination

Simple regression analyses were performed to identify items whose difficulty deviated most from what would be predicted by frequency or familiarity alone. With Rasch logit difficulty as the dependent variable, BNC frequency was not a significant predictor ($\beta = -2.57 \times 10^{-5}$, $t = -0.76$, $p = .451$, $R^2 = .01$), whereas familiarity was significant ($\beta = -0.68$, $t = -2.83$, $p = .007$, $R^2 = .14$). In a multiple regression including both predictors, familiarity remained significant ($\beta = -0.72$, $t = -2.70$, $p = .012$) while frequency did not contribute ($\beta = 1.19 \times 10^{-5}$, $t = 0.34$, $p = .731$), and the combined model explained 14.5% of the variance in item difficulty ($R^2 = .145$, $F(2, 47) = 3.99$, $p = .034$).

Linear regression assumptions were examined for both simple regressions (logit difficulty on BNC frequency, logit difficulty on familiarity). The frequency model showed evidence of non-normality of residuals (Shapiro–Wilk $W = .92$, $p = .002$), driven by extreme items at the easy end of the scale; homoscedasticity (Breusch–Pagan $p = .860$) and independence (Durbin–Watson = 1.90) were not violated. The familiarity model met the

assumption of residual normality (W = .96, p = .110) and independence (Durbin–Watson = 1.97), but showed evidence of heteroscedasticity (Breusch–Pagan p = .008). Given these violations at the tails, results are interpreted with appropriate caution, and Spearman correlations (which do not require these assumptions) are reported in the preceding correlation section and are treated as the primary evidence for RQ2. Large residuals were defined as those with |standardized residual| > 2 (Cohen et al., 2003).

Because BNC frequency was not a significant predictor and accounted for almost none of the variance in item difficulty, residuals from the frequency-based model were not interpreted as deviations from meaningful frequency-based predictions; the residual examination therefore focused on the familiarity-based model.

In the familiarity-based regression, items that were substantially easier than predicted included Q14 (*in particular*), Q16 (*as a result*), Q27 (*in touch*), Q17 (*as soon as*), and Q42 (*come up to*). Items that were substantially more difficult than predicted included Q20 (*be expected to*), Q29 (*at once*), Q3 (*so that*), Q46 (*to blame*), and Q12 (*to do with*).

Discussion

This section is organized according to the four inferences of the argument-based validation framework (Chapelle et al., 2008). For each of the domain definition, evaluation, generalization, and explanation inferences, the evidence obtained in this study is synthesized and its implications for the validity argument are considered.

Domain definition inference

The domain definition inference concerns whether the test content adequately represents the target language use domain. For the E–J PVST, this inference is well supported. The test items are drawn from the PHRASE List (Martinez & Schmitt, 2012), a corpus-derived and expert-vetted collection of high-frequency English phrasal expressions. The PHRASE List was compiled from the British National Corpus and organized into five frequency bands, providing systematic coverage of the phrasal expression domain. The 50 PVST items sample from two structural types, adverbial phrases and verb phrases, that together

account for over 80% of the PHRASE List, ensuring representativeness of the most common phrase types.

The Rasch analysis showed that item difficulties spanned a substantial range (approximately −5.00 to +2.75 logits), and the Wright map confirmed that the distribution of item difficulties was broadly aligned with the distribution of person abilities, with no floor or ceiling effects. Although the test does not assess productive knowledge, oral/aural recognition, or all subtypes of MWEs (e.g., collocations, idioms, lexical bundles), its scope is clearly defined and appropriate for its intended purpose: measuring written receptive knowledge of phrasal expressions.

One consideration relevant to domain definition is the age of the source corpus. The BNC, compiled primarily from texts produced in the early 1990s, may not fully reflect contemporary English usage. However, because phrasal expressions tend to be highly stable over time (Wray, 2002), with the probability of newly invented phrasal expressions of the type included in the PHRASE List being low, the BNC-based frequency data remain a reasonable basis for item selection and difficulty prediction.

A further consideration is that the BNC is predominantly composed of British English texts, whereas many Japanese EFL learners are primarily exposed to American English through textbooks and media. Although many high-frequency phrasal expressions are common across varieties, frequency rankings can differ across corpora, and some expressions may be more or less salient depending on the learner's input variety. Future work could examine whether alternative and more recent corpora, such as large American English corpora, yield different frequency rankings for these same phrasal expressions. Because predictions of item difficulty in the present study were based on corpus frequency, frequency rankings that better reflect the English input Japanese learners actually receive (e.g., from textbooks and US-origin media) might produce closer alignment between predicted and observed difficulty.

Evaluation inference

The evaluation inference, corresponding to Kane's (2006) scoring inference, concerns whether test performances are scored accurately and consistently and whether items

function as intended. This inference was largely supported: overall item fit was acceptable and the Rasch analyses yielded clear evidence of reliability and targeting, although the fit and distractor results identified specific items where revision or replacement would strengthen measurement.

However, the item fit results indicate that 12% of items showed underfit to the Rasch model, suggesting potential measurement disturbance, while an additional 22% showed overfit, suggesting item redundancy. Although overfit is generally considered less harmful than underfit (Bond et al., 2021), eleven overfitting items represent 22% of a 50-item test and therefore reduce the test's informational efficiency. In a short test where each item's contribution matters, this level of redundancy suggests that the test could be shortened without a meaningful loss of measurement precision, or that redundant items could be replaced with items targeting underrepresented difficulty levels, particularly in the upper range where the Wright map indicated person abilities exceeding item difficulties. Underfitting items remain the highest priority for revision, especially where distractor malfunction was identified as a likely source of misfit.

The distractor analysis also suggests possible L1-specific sources of construct-irrelevant difficulty introduced through translation. For Q29 (*at once*), the distractor 早く (*hayaku*, "early") and the correct answer すぐに (*sugu ni*, "immediately") share semantic overlap in Japanese, both conveying a sense of promptness. For Q30 (*in time*), two options, 速く (*hayaku*, "fast") and 早く (*hayaku*, "early"), are homophones in Japanese but written with different kanji, potentially creating a visual discrimination task unrelated to knowledge of the English phrasal expression. These cases illustrate that while L1 options can reduce construct-irrelevant difficulty, careful translation and distractor design remain essential to avoid introducing new sources of confusion.

A further translation-related observation concerns part-of-speech alignment between English and Japanese options. In the example item Q1 (*go on*) shown in the Methods section, the four English options were verb forms (*sleep*, *repeat*, *be fast*, *continue*) translated as 寝る (*neru*, "sleep"), 繰り返す (*kurikaesu*, "repeat"), 速い (*hayai*, "be fast"), and 続く (*tsuzuku*, "continue") respectively. Option (c) is an adjective (速い, *hayai*, "be

fast") rather than a verb, reflecting that the underlying English option *be fast* is a copular construction (copula + adjective), which the Japanese translator rendered with the adjective alone. A small number of other items showed similar grammatical asymmetries. This inconsistency was not flagged during piloting and did not produce systematic misfit in the present sample. Future revisions of the E–J PVST will aim for category-aligned distractors within each item.

Where direct translation introduces semantic overlap or homophone-induced ambiguity, three approaches may help. First, distractors can be paraphrased rather than directly translated, prioritising clear separation in L1 meaning over surface fidelity to the L2 option. Second, when an L1 lexical gap or homograph is unavoidable, the source-language item can be supplemented with a brief disambiguating context in L1 (e.g., a short collocate or domain marker) to make the intended distractor sense unambiguous. Third, when a particular distractor remains problematic despite these measures, it can simply be replaced with a different distractor that avoids the semantic overlap or homophony altogether.

Q21 (*give up*) was retained for two reasons. First, in the context of a size test designed to estimate the proportion of phrasal expressions a learner knows, items that most learners know are expected and necessary because they confirm that the test captures the floor of the knowledge domain. Second, removing an item answered correctly by all participants would reduce content coverage of the highest-frequency band without improving measurement precision for any part of the ability range. If the test is used primarily with advanced populations, replacement with a more discriminating item from the same frequency band would be advisable.

Generalization inference

The generalization inference addresses whether the observed test scores generalize to the broader target domain sampled by the test, namely the high-frequency adverbial and verb phrases on the PHRASE List (Martinez & Schmitt, 2012). This inference is strongly supported. The E–J PVST demonstrated high internal consistency, with a person reliability estimate of .88 and an item reliability estimate of .97. The person separation

index (2.67) confirmed that the test can discriminate among at least two to three distinct ability levels, and the item separation index (5.98) confirmed that the items span approximately five to six distinct difficulty levels. While the number of separable levels coincides numerically with the five frequency bands from which items were drawn, the empirical alignment between Rasch difficulty and BNC frequency band was weak (see RQ2), so the separable difficulty strata should not be taken as direct evidence of frequency-band-based difficulty structure. Completion times were not systematically recorded in the present study; future iterations could include this metric to document the test's practical efficiency for low-stakes assessment contexts.

The absence of floor or ceiling effects indicates that the test provides informative measurement across the full range of proficiency represented in the sample. The 50-item test length balances measurement precision with practical efficiency: it is long enough to yield reliable scores but short enough to be administered in a single sitting without excessive fatigue, particularly when L1 options reduce response time.

One qualification to the generalization inference concerns the effect of sample heterogeneity on person reliability. The wide proficiency range in the present sample (self-reported CEFR A1 to C1, SD of total scores = 17.2 percentage points) likely inflated the person reliability estimate relative to what would be observed with a more homogeneous group. If the test were administered to a single CEFR level, such as B1 learners only, person reliability would likely be lower than .88. This is a statistical property of reliability coefficients, which increase with score variability, and it does not indicate a flaw in the test. However, it does mean that the reported person reliability should be interpreted as an upper bound for relatively homogeneous testing contexts, and users should consider estimating reliability within their target population.

Explanation inference

The dimensionality analysis showed a disattenuated correlation of .73 between contrasting item clusters, supporting the conclusion that the test measures a single construct. While this value is above the .70 threshold for unidimensionality, it is not overwhelmingly high, leaving open the possibility that minor secondary dimensions (e.g.,

knowledge of constituent words, sensitivity to contextual cues) may influence performance to some degree.

The frequency–familiarity analysis (RQ2) provided evidence relevant to the explanation inference. Item difficulty correlated weakly with corpus frequency (Spearman's $\rho = -.21$, $p = .143$) and weakly to moderately with learner-reported familiarity ($\rho = -.31$, $p = .031$), suggesting that the construct measured by the E–J PVST is shaped by learners' experience with phrasal expressions and cannot be fully explained by corpus statistics alone. The moderate size of these correlations also indicates that item difficulty is multiply determined, likely reflecting additional factors such as semantic transparency and item construction.

The distractor efficiency results provide complementary, indirect evidence. Items with functional distractors that attracted lower-ability test-takers more than higher-ability test-takers show response patterns consistent with a gradient of phrasal knowledge, whereas items with non-functional distractors may allow correct responses through elimination of implausible options, reducing construct-relevant discrimination. A limitation of the present explanation evidence is the absence of process-level data that could directly confirm whether responses are primarily driven by holistic phrasal knowledge rather than compositional inference or test-taking strategies. Future validation work incorporating process-oriented methods would strengthen this inference.

Limitations and future research

Several limitations should be considered when interpreting the results. First, English proficiency was based on self-reported CEFR levels. Without concurrent standardized test scores for all participants, this classification may lack precision. Future studies should include an independent proficiency measure to provide stronger criterion-related evidence. Second, the unsupervised remote administration allowed for a diverse sample but introduced the possibility of participants consulting external resources. While this is a common challenge in online research, comparing supervised and unsupervised modes would clarify the impact of the testing context on validity. Third, the absence of detailed learner backgrounds, such as learning history or length of residence in English-speaking

countries, limits the understanding of individual differences. Future research should also incorporate process-oriented methods, such as think-aloud protocols, to verify whether correct responses reflect holistic knowledge or test-taking strategies. This would specifically strengthen the explanation inference. Finally, the E–J PVST is designed for L1 Japanese learners, and the results may not generalize to other L1 populations or different types of MWEs like collocations or idioms. Expanding the validation to diverse languages and formulaic sequences will extend the evidence base for the argument-based framework in vocabulary assessment.

Conclusion

This study provided validation evidence for the E–J PVST using the argument-based validation framework. By aligning empirical evidence with four key inferences (domain definition, evaluation, generalization, and explanation), the findings provide largely supportive validation evidence for the E–J PVST as a measure of written receptive knowledge of the adverbial and verb phrases on the PHRASE List (Martinez & Schmitt, 2012) for L1 Japanese learners.

The domain definition inference was supported by the systematic selection of items from the PHRASE List. The evaluation inference received partial support, although item misfit and distractor malfunction were identified as areas requiring revision. Evidence for generalization was provided by the high reliability and separation indices, which confirmed the stability of the item difficulty hierarchy. Finally, the explanation inference was supported as performance reflected knowledge of phrasal expressions, with familiarity emerging as a stronger predictor of item difficulty than corpus frequency.

This research contributes to the limited literature applying argument-based validation to vocabulary assessment. It demonstrates that the framework is applicable to researcher-developed instruments rather than being restricted to high-stakes commercial tests. The findings also highlight the role of subjective familiarity alongside corpus frequency in predicting item difficulty. Identifying specific issues with distractor quality and L1 translation provides a clear path for improving the instrument.

On the basis of the present findings, the E–J PVST is a promising tool for assessing phrasal expression knowledge and for raising awareness of MWEs in educational settings. Targeted revisions addressing the identified issues will further strengthen the validity argument for the E–J PVST across diverse populations and testing contexts.

Acknowledgements

I am deeply grateful to Drs. Tatsuya Nakata, Ron Martin, and Satomi Mori for their guidance, support, and valuable feedback throughout this study. I also thank Taiga Kaitsu, Sayane Soma, Yumi Fujita, Kyoko Hitomi, Daisuke Izumi, Noriaki Mikajiri, and Ryuhei Mizoguchi for their helpful comments and advice on earlier drafts of the manuscript.

Author disclosures

The author reported no conflict of interest.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. This paper is based on the author's master's thesis submitted to Rikkyo University.

ChatGPT and Claude were used to check the grammar and vocabulary of the manuscript. The author reviewed all suggestions and takes full responsibility for the content of the article.

ORCID iD

Ayako Aizawa https://orcid.org/0009-0009-0670-2488

References

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge. https://doi.org/10.4324/9780429030499

Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch analyses in the human sciences*. Springer. https://doi.org/10.1007/978-3-030-43420-5

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge. https://doi.org/10.4324/9780203937891

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates.

Drown, L., Giovannone, N., Pisoni, D. B., & Theodore, R. M. (2023). Validation of two measures for assessing English vocabulary knowledge on web-based testing platforms: Long-form assessments. *Linguistics Vanguard, 9*(1), 113–124. https://doi.org/10.1515/lingvan-2022-0115

Dunlea, J. (n.d.). *Eiken to CEFR to no kankeisei ni tsuite* [The relationship between Eiken and CEFR]. Eiken Foundation of Japan. Retrieved February 25, 2026, from https://www.eiken.or.jp/eiken/group/result/pdf/report_02.pdf

Dursun, A., & Li, Z. (2021). A systematic review of argument-based validation studies in the field of language testing (2000–2018). In C. A. Chapelle & E. Voss (Eds.), *Validity argument in language testing: Case studies of validation research* (pp. 45–70). Cambridge University Press. https://doi.org/10.1017/9781108669849

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing, 30*(2), 253–272. https://doi.org/10.1177/0265532212459028

Gyllstad, H., & Schmitt, N. (2018). Testing formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 174–191). Routledge. https://doi.org/10.4324/9781315206615

IELTS (n.d.). *IELTS and the CEFR (Common European Framework of Reference)*. Retrieved February 25, 2026, from https://www.ielts.org/about-ielts/ielts-in-cefr-scale

Kanazawa, Y. (Ed.). (2020). *Formula to gaikokugo gakushu, kyouiku: Teikei hyogen kenkyu nyumon* [Formulaic sequences in foreign language learning and education: An introduction]. Kurosio.

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

Linacre, J. M. (2023a). *Winsteps® Rasch measurement computer program* (Version 5.6.0). Winsteps.com.

Linacre, J. M. (2023b). *Winsteps® Rasch measurement computer program user's guide* (Version 5.6.0). Winsteps.com.

Martinez, R. (2011). *The development of a corpus-informed list of formulaic sequences for language pedagogy* [Unpublished doctoral dissertation]. University of Nottingham.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299–320. https://doi.org/10.1093/applin/ams010

Matsuo, T. (2022). A Rasch-based validation of the Phrasal Vocabulary Size Test. *Journal of Osaka Jogakuin University, 18*, 93–109.

McLean, S., Ishii, T., Stoeckel, T., Bennett, P., & Matsumoto, Y. (2016). An edited version of the first eight 1,000-word frequency bands of the Japanese–English version of the Vocabulary Size Test. *The Language Teacher, 40*(4), 3–7.

McLean, S., & Kramer, B. (2015). The creation of a new vocabulary levels test. *Shiken, 19*(2), 1–11.

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing, 37*(3), 389–411. https://doi.org/10.1177/0265532219898380

McNamara, T. F. (1996). *Measuring second language performance*. Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

Minamide, Y., & Nakamura, M. (2022). *Genius English–Japanese dictionary* (6th ed.). Taishukan Publishing.

Mizumoto, A., & Takeuchi, O. (2023). *Gaikokugo kyoiku kenkyu handobukku* [Handbook of research on foreign language teaching] (Expanded ed.). Shohakusha.

Nation, I. S. P., & Anthony, L. (2016). Measuring vocabulary size. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. 3, pp. 355–368). Routledge.

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

Nguyen, L. T. C., & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal, 42*(1), 86–99. https://doi.org/10.1177/0033688210390264

Park, H. I. (2024). Validation of the Korean bilingual version of the Vocabulary Size Test. *English Teaching, 79*(2), 139–162. https://doi.org/10.15858/engtea.79.2.202406.139

Schmitt, N. (2010). Researching vocabulary: A vocabulary research manual. Palgrave Macmillan. http://dx.doi.org/10.1057/9780230293977

Schmitt, N., Dunn, K., O'Sullivan, B., Anthony, L., & Kremmel, B. (2021). Introducing Knowledge-based Vocabulary Lists (KVL). *TESOL Journal*, *12*(4). e.622. https://doi.org/10.1002/tesj.622

Takebayashi, S. (Ed.). (2002). *Kenkyusha New English–Japanese dictionary* (6th ed.). Kenkyusha.

Tannenbaum, J. R., & Wylie, E. C. (2008). Linking English-language test scores onto the Common European Framework of Reference. *TOEFL iBT Research Report, 6*. https://doi.org/10.1002/j.2333-8504.2008.tb02120.x

TOEFL (n.d.). *Comparing TOEFL iBT Scores*. Retrieved February 25, 2026, from https://www.ets.org/toefl/score-users/ibt/compare-scores.html

Wood, D. (2020). Classifying and identifying formulaic language. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 30–45). Routledge.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press. https://doi.org/10.1017/CBO9780511519772

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Appendix

PVST Items and Multiple Choice Options

First 1,000

Q1	go on: It will **go on**.	
	a.	寝る
	b.	繰り返す
	c.	速い
	d.	続く

Q2	lead to: No one knows what it will **lead to**.	
	a.	～がほしい
	b.	内部に～を持っている
	c.	将来～を引き起こす
	d.	～を見つける

Q3	so that: He sat **so that** they could do it.	
	a.	～ができるように
	b.	～なので
	c.	とてもゆっくりと、そしてそれから
	d.	～する前に

Q4	at all: I don't like it **at all**.	
	a.	いつも
	b.	全然
	c.	はじめは
	d.	時々

Q5	I mean: Two, **I mean**, three.	
	a.	私は推測している
	b.	もしかして
	c.	そしてそのあと
	d.	いや、その

Q6	at least: **At least** it is warm.	
	a.	少なくとも
	b.	何日も経って、そして今
	c.	私は～だと信じられない
	d.	最も重要でないことは

Q7	be likely to: He is **likely to** go.	
	a.	～するのが好きだ
	b.	～できる
	c.	～したい
	d.	おそらく～するだろう

Q8	be to: He **is to** speak this afternoon.	
	a.	～することになっている
	b.	～できる
	c.	～したい
	d.	～するかもしれない

Q9	deal with: I can **deal with** it.	
	a.	～を解決する
	b.	～を試す
	c.	～を見つける
	d.	～を見る

Q10	used to: I **used to** go.	
	a.	～したい
	b.	以前は～した
	c.	たいてい
	d.	いつも

Second 1,000

Q11	so far: It's good **so far**.	
	a.	今までのところは
	b.	そういうわけではない
	c.	時々
	d.	日中は

Q12	to do with: It is **to do with** money.	
	a.	～を作ること
	b.	～のために
	c.	～に関する
	d.	私たちの

Q13	take over: They will **take over**.	
	a.	終わっている
	b.	引き継ぐ
	c.	後で来る
	d.	それについて考える

Q14	in particular: I want that **in particular**.	
	a.	特に
	b.	内緒で
	c.	それは違うから
	d.	おそらく

Q15	for instance: **For instance**, it is cheaper.	
	a.	おそらく
	b.	短い間
	c.	私の考えでは
	d.	例えば

Q16	as a result: **As a result,** it was done.	
	a.	～かどうか誰も知らない
	b.	長い時間を経（へ）て
	c.	その前に
	d.	結果として

Q17	as soon as: I'll go **as soon as** I can.	
	a.	～するとすぐに
	b.	～する場合にのみ
	c.	～した後に
	d.	～する前に

Q18	carry out: It was **carried out** yesterday.	
	a.	始められた
	b.	見つけられた
	c.	読まれた
	d.	実行（じっこう）された

Q19	be about to: I **am about to** read the newspaper.	
	a.	～するのが待ちきれない
	b.	もうすぐ～するつもりだ
	c.	～するのがとても好きだ
	d.	～しようと努（つと）めている

Q20	be expected to: We **are expected to** do this.	
	a.	～するのを待っている
	b.	～するのを望んでいる
	c.	～しなければならない
	d.	～することができる

Third 1,000

Q21	give up: I **give up**.	
	a.	精一杯努力する
	b.	出発する
	c.	あきらめる
	d.	練習する

Q22	feel like: I just did not **feel like** it.	
	a.	～を愛する
	b.	～したいと思う
	c.	～について考える
	d.	～しようとする

Q23	turn out: It **turned out** different.	
	a.	～で始まった
	b.	～のように思われた
	c.	～になった
	d.	～に見えなかった

Q24	other than: **Other than** that, it's good.	
	a.	～を除いて
	b.	～を含むならば
	c.	～のおかげで
	d.	～の後に

Q25	get to: She **got to** the car.	
	a.	～に着いた
	b.	～を運転した
	c.	～を受け取った
	d.	～に入った

Q26	all over: It is **all over** the bed.	
	a.	～を覆（おお）って
	b.	～の上にない
	c.	～の上に
	d.	～の近くに

Q27	in touch: Keep **in touch**.	
	a.	それを感じて
	b.	連絡して
	c.	それを押して
	d.	考えて

Q28	get rid of: They **got rid of** it.	
	a.	～を捨てた
	b.	～を受け取った
	c.	～にうんざりした
	d.	～を選んだ

Q29	at once: I did it **at once**.	
	a.	一回
	b.	何回も
	c.	早めに
	d.	直ちに

Q30	in time: **In time** they bought a house.	
	a.	速く
	b.	早く
	c.	やがて
	d.	代わりに

Fourth 1,000

Q31	prove to be: It has **proved to be** important.	
	a.	きっと～になるだろう
	b.	それ自体が～であることを示した
	c.	～であり続けた
	d.	決して～ではなかった

Q32	next door: It's just **next door**.	
	a.	もうすぐ来る
	b.	一般的な
	c.	完璧な
	d.	とても近い

Q33	run out: I think we **ran out** of it.	
	a.	なくなった
	b.	うんざりした
	c.	考えた
	d.	外に移動した

Q34	take advantage: You must **take advantage**.	
	a.	ゆっくり進む
	b.	利用する
	c.	注意を払う
	d.	くつろぐ

Q35	in effect: It is **in effect** the reason.	
	a.	おそらく
	b.	～ではない
	c.	今
	d.	実際には

Q36	in the light of: It was accepted **in the light of** the money.	
	a.	～にもかかわらず
	b.	～の理由で
	c.	～に加えて
	d.	～の代わりに

Q37	by no means: He is **by no means** tall.	
	a.	とても
	b.	全く～ではない
	c.	いくぶん
	d.	～だと考えられている

Q38	come across: They **came across** a hotel.	
	a.	～に泊まった
	b.	～を開いた
	c.	～に近かった
	d.	～を見つけた

Q39	happen to: She **happened to** call.	
	a.	～のふりをした
	b.	～を一生懸命（いっしょうけんめい）試みた
	c.	～したくなかった
	d.	たまたま～した

Q40	even so: **Even so** it's better.	
	a.	それにもかかわらず
	b.	その方が
	c.	同様で、さらに
	d.	おそらく

Fifth 1,000

Q41	by far: She is **by far** the most intelligent.	
	a.	～になろうとする
	b.	全く～ではない
	c.	本当に
	d.	時々

Q42	come up to: He just **came up to** me.	
	a.	～に近づいた
	b.	～を断った
	c.	～が好きではなかった
	d.	～を真似した

Q43	straight away: They did it **straight away**.	
	a.	すぐに
	b.	正しい方法で
	c.	ゆっくりと
	d.	彼らがしたかったから

Q44	would appear: It **would appear** it's true.	
	a.	～であるはずがない
	b.	確かに～である
	c.	～だと思われる
	d.	～だと考えられている

Q45	turn down: She **turned down** the money.	
	a.	～を隠した
	b.	～をなくした
	c.	～を作った
	d.	～を断った

Q46	to blame: We are not **to blame**.	
	a.	全く同じ意見である
	b.	興味がある
	c.	誰かを責（せ）めている
	d.	責任がある

Q47	take for granted: She **took it for granted**.	
	a.	それを保（たも）った
	b.	当然のことと思った
	c.	それがとても欲しかった
	d.	それを慎重（しんちょう）に考えた

Q48	as of: It changes **as of** today.	
	a.	～の時点で
	b.	時々
	c.	もしかしたら
	d.	～のおかげで

Q49	can tell: You **can tell**.	
	a.	話すかもしれない
	b.	賢い
	c.	分かる
	d.	するかもしれない

Q50	over time: **Over time** it was cheaper.	
	a.	昔は
	b.	結局は
	c.	手遅（ておく）れになった時に
	d.	ちょうどいい時に