Article
Shiken: A Journal of Language Testing and Evaluation in Japan
2025, Vol. 29(2) 1–26
DOI: 10.37546/JALTSIG.TEVAL29.2-1
ISSN: 1881-553

A validation study of a test of vocabulary depth using many-facet Rasch analysis

Chihiro Komiya[1], James Sick[2] and Stuart McLean[3]
chihirokomiya19@gmail.com
[1]*Temple University, Japan Campus*
[2]*Temple University, Japan Campus*
[3]*Kindai University*

Abstract

Vocabulary knowledge can be classified as having two components, size and depth. Size refers to the number of words one recognizes at a basic degree, a quantitative aspect, whereas vocabulary depth emphasizes the richness of understanding each word, a qualitative aspect. While vocabulary size has been widely researched, less attention has been paid to depth of vocabulary knowledge. Few tests have been developed that attempt to measure the multiple components that make up depth of vocabulary knowledge, nor has validity evidence been provided for such tests, especially using Many-facet Rasch analysis. This study investigated the reliability, validity, and dimensionality of four tests targeting four components of vocabulary depth: form-meaning link, collocation, word parts, and polysemy. Data from 120 learners of English were examined using many-facet Rasch analysis. The results showed that the four tests can reasonably be combined to form a single measure labeled depth of vocabulary knowledge, with adequate reliability and validity. However, revisions on some problematic items might be necessary for greater validity.

Keywords: vocabulary depth, many-facet Rasch analysis, validity, reliability, unidimensionality

Many scholars have emphasized the importance of second language (L2) vocabulary knowledge in English as a second or foreign language literature (Nation, 2013; Read, 2000; Schmitt & Schmitt, 2020). Vocabulary has been researched widely as a language component that plays a significant role in the overall language proficiency of L2 learners (Nation, 2013). However, vocabulary knowledge is a complicated construct because multiple word knowledge components are involved (Nation, 2013; Read, 2000), and it is challenging to measure this multidimensional construct.

In L2 vocabulary research, it is generally accepted that vocabulary knowledge is comprised of size (breadth) and depth, concerning quantitative and qualitative aspects, respectively. When measuring knowledge of these two dimensions, the construct of vocabulary size has generally been simplified to mean passive recognition of a single word meaning, whereas the construct of vocabulary depth has been treated as multifaceted, with various components that indicate a deeper or more nuanced understanding of a word. Depending on the researcher, tests designed to measure vocabulary depth have differed both in which components are targeted as well as in how they are measured.

Literature Review

The Components Approach to Vocabulary Knowledge

Among attempts to conceptualize vocabulary knowledge, Nation's (2013) systematic summary of various types of word knowledge is still considered the best model (Schmitt, 2014). Nation classifies vocabulary knowledge into three components—form, meaning, and use—and each component is further divided into three aspects of knowledge. These nine subdivisions—Spoken, Written, Word parts, Form and meaning, Concepts and referents, Associations, Grammatical functions, Collocations, and Constraints on use—are

further divided into receptive and productive aspects. Receptive knowledge is required for comprehension during reading and listening, while productive knowledge is required for speaking and writing. In total, there are 18 aspects of vocabulary knowledge, as shown in Table 1.

Table 1

Overall Statistics for EFQ

Form	Spoken	R	What does the word sound like?
		P	How is the word pronounced?
	Written	R	What does the word look like?
		P	How is the word written and spelled?
	Word parts	R	What parts are recognizable in this word?
		P	What word parts are needed to express this meaning?
Meaning	Form and meaning	R	What meaning does this word form signal?
		P	What word form can be used to express this meaning?
	Concepts and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	Associations	R	What other words does this make us think of?
		P	What other words could we use instead of this one?
Use	Grammatical functions	R	In what patterns does the word occur?
		P	In what patterns must we use this word?
	Collocations	R	What words or types of words occur with this one?
		P	What words or types of words must we use with this one?
	Constraints on use	R	Where, when and how often would we expect to meet this word?
		P	Where, when and how often can we use this word?

Note. R = Receptive; P = Productive. Adapted from "Learning Vocabulary in Another Language" by Nation, 2013, p. 49. Copyright 2013 by Cambridge University Press.

The construct of vocabulary knowledge was first classified into breadth (size) and depth of knowledge by Anderson and Freebody (1981). Vocabulary breadth refers to how many words a person knows, whereas vocabulary depth refers to how well a person knows these words. Vocabulary size is the most tested dimension of vocabulary knowledge (Kremmel & Pellicer-Sánchez, 2020), and there are different types of vocabulary size tests that have been widely employed, such as the Vocabulary Levels Test (Nation, 1983; Schmitt et al., 2001), the Vocabulary Size Test (Nation & Beglar, 2007), and the updated Vocabulary Levels Test (Webb et al., 2017). In contrast, there are no standardized vocabulary tests that measure vocabulary depth. In addition, although there are multiple components of vocabulary knowledge as shown in Table 1, not many studies have been carried out to measure multiple components of vocabulary knowledge concurrently (González-Fernández & Schmitt, 2019).

Multiple Components of Vocabulary Knowledge

Several researchers who have examined the relative difficulty of multiple components of vocabulary knowledge employed their own self-constructed instruments. Chui (2006) measured academic vocabulary

knowledge of 186 university students in Hong Kong. Twenty words were taken from the Academic Word List (Coxhead, 2000) and four vocabulary components were tested: word class recognition, meaning recall, collocation recognition, and derivative form production. The results showed that the knowledge of the recognition of collocation and the productive derivative forms lagged behind the ability to recognize word class (derivational suffixes) and to recall word meaning.

González-Fernández and Schmitt (2019) examined the knowledge of 20 words across 144 Spanish-speaking learners of English for four components of vocabulary knowledge: form-meaning link, derivatives, multiple meanings, and collocations. Both recall and recognition knowledge were tested for the 20 words. The results showed the word knowledge components were highly correlated (.678 to .945) and that recall knowledge of all four components was more difficult than any type of recognition knowledge. In terms of individual components of recognition knowledge, form-meaning link was the easiest, followed by collocations and multiple meanings, and derivatives were the most difficult. In contrast, for recall knowledge, it was collocation that was the easiest, followed by form-meaning link, derivates, and multiple meanings. The researchers suggested that the measurement of only 1 to 3K frequency bands in the recall collocation test might have resulted in a higher score than the form-meaning link test where words from 1K to 9K were tested. In addition, the results of structural equation modeling (SEM) revealed that all word knowledge components significantly contributed to the global vocabulary knowledge construct.

More recently, Read and Dang (2022) measured the high-frequency academic vocabulary knowledge of 222 first-year students in a program of English for academic purposes at a Vietnamese university. Three different kinds of tests were created by the authors to measure the participants' knowledge of synonyms, collocations, and word parts. A selected-response format was used for the synonyms test, while a Yes/No response with a Not Sure response was used for the collocation and word parts tests. The results showed that knowledge of synonyms was the easiest, followed by inflected/derived forms. Collocation was the most difficult.

Although efforts have been made to investigate the relationships among multiple vocabulary components, there are limitations in existing research. For example, the test formats employed were sometimes inconsistent, such as mixing recognition and recall across components (Chui, 2006; Li & Kirby, 2015). Other studies measured multiple components, but for different target words (Li & Kirby, 2015; Read & Dang, 2022). In addition, the use of these researcher-developed tests has not always been supported by sufficient validity evidence (González-Fernández & Schmitt, 2019; Read & Dang, 2022).

Although current approaches to validity focus primarily on score interpretation and the consequences of inferences based on test scores (Messick, 1989), they still recognize the evidential basis of construct validity (see Fulcher & Davidson, 2007). Furthermore, within Rasch measurement theory, establishing the fit of items to a measurement model is an essential part of the validation process (Bond et al., 2020).

Because instrument validation is a multi-step process, a single study cannot provide full validation. Accordingly, this study offers preliminary evidence regarding the unidimensionality of a "depth of vocabulary knowledge" construct. Future research will examine the extent to which score-based inferences about vocabulary knowledge and reading comprehension, such as performance on the TOEIC reading test, are justified.

The following four research questions will be addressed:

Research Questions

1. Do the four subcomponents of vocabulary depth proposed here—form-meaning link, collocation, word parts, and polysemy—form a coherent unidimensional variable?
2. What is the reliability and validity of the combined measure of vocabulary depth?
3. What are the reliability and validity values of the four subtests analyzed individually?
4. Is unidimensionality of the combined test further supported by correlation and factor analyses?

Methods

Participants

Data were collected from a total of 120 participants (97 female, 23 male) enrolled at three institutions in Japan. They were in their late teens or early 20s and had a minimum English education background of three years in junior high school and three years in high school. The participants included 60 (47 female, 13 male) first, second, third, and fourth-year students in various departments from a private university, 10 (five female, five male) second-year students in an English major from a vocational school, and 50 (45 female, five male) first-year students in a non-English department from another private university.

Target Words

Based on the following criteria, 30 words were selected to measure the four aspects of word knowledge components (form-meaning link, collocation, word parts, and polysemy).

- Range of frequency (1K–4K), to ensure that there would be no floor or ceiling effect for participants.
- Different parts of speech, to have a representative sample of words. The final target words list included 15 nouns, nine verbs, and five adjectives.
- Multiple meanings, to measure knowledge of multiple meanings. It was necessary for each target word to have at least two senses. The secondary meaning was chosen so that it could be as different as possible from the core meaning tested in the form-meaning link test.
- Derivative forms, to measure knowledge of word parts. All target words needed to have at least one derivative form for the same meaning as the core meaning measured in the form-meaning link.

The final 30 target words were *price, interest, mind, bank, hard, serve, employ, bright, fresh, poor, season, character, period, fashion, respect, coach, industry, apply, operate, smooth, wealth, charity, province, compose, absorb, implement, appropriate, leaf, adhere,* and *shrink.* The same 30 target words were tested across the four-word knowledge components, specifically so that the relative difficulty of the four components could be examined.

Measures

Form-meaning Link

Participants were asked to translate 30 English sentences into Japanese. We chose to measure meaning-recall knowledge because it has been shown to correlate more strongly with reading comprehension (e.g., Komiya & McLean, 2024; Zhang & Zhang, 2020) than meaning-recognition knowledge such as multiple-

choice questions. The vocabulary tests were developed to examine the relationship with reading comprehension.

Other than the target words, all the surrounding words used in each sentence consisted of words from 1K with the single exception of *desk* (2K). The most frequent meaning of each target word was determined by referring to Corpus of Contemporary American English (COCA; Davies, 2008). For example, *season* can be both a noun and a verb. However, its most frequent meaning was considered to be the noun form, because its frequency as a noun is 217,695, whereas its frequency as a verb is 1,220. COCA, however, does not provide separate word frequency counts for different meaning senses when the meanings are the same part of speech. In such cases, the more frequent meaning was determined by examining the order of the definitions presented in three learner dictionaries: Oxford Advanced Learner's Dictionary 8th edition (2010), Longman Dictionary of Contemporary English 5th Edition (2009), and Cambridge Online English Dictionary. For instance, *hard*, has multiple meanings as an adjective. However, because the three dictionaries put the meaning of 'solid' first, it was tested as the main meaning.

In the sample item below, the target word is price. The target word is not highlighted in any way that would indicate to test-takers which word is the target. Words in parentheses are some of the acceptable answers. Participants were asked to translate the entire sentence into Japanese, but only the target word was marked. Therefore, the answer was marked correct when their translation included any of these acceptable answers.

I don't know the price. (値段、価格、金額).

Collocation

Participants were asked to translate 30 English sentences into Japanese. The collocates were taken from the COCA (Davies, 2008) and were among the most frequent collocates of the target word. The collocations were defined as any two words in combination with a greater corpus frequency than could happen by chance, which have a Mutual Information (MI) score ≧ 3 (Hunston, 2002) and within a span of ± 4 words. All 30 collocations in the test were lexical and from the 1K and 2K frequency bands, with the exception of tennis (3K), baseball (4K), and lettuce (6K), all of which are loanwords in Japanese and thus commonly known to Japanese test takers. In the sample item below, the target word is price and the collocate word is high (i.e., the collocation is high price). Words in parentheses are some of the acceptable answers. Participants were asked to translate the entire sentences into Japanese, but only the target word was marked. Thus, when their translation contained any of these acceptable Japanese words, it was marked as correct.

I don't want to pay a high price for this. (値段、価格、金額)

Word Parts

Participants were asked to translate 30 English sentences into Japanese. Other than the target words, all the surrounding words were taken from the 1K level with the exception of four words from 2K (*restaurant, printed, banana,* and *vegetable*). Out of 30 target derivative words, ten words were nouns, one word was a verb, 15 words were adjectives, and four words were adverbs. According to Bauer and Nation's (1993) levels of affixes, out of 30 target words, 14 words were in Level 3 (*-y, -ly, -able, -er*), nine words were in Level 4 (*-ful, -ness, -al, -ial, -ation*), two words were in Level 5 (*-en, -ence*), and five words were in Level 6 (*-ic, -ition, -ion*). Thus, parts of speech as well as levels are diverse. In the sample item below, the target

word is *price*. Words in parentheses are some examples of acceptable answers. To be scored correctly, test-takers must write any of these acceptable Japanese words.

This is a pricey restaurant.（高い、高級な、高価な）

The whole test sets are reported in Appendix A.

Procedures

The test was administered using Survey Monkey in a classroom under teacher supervision with no time limit. The instruction began with an explanation that an English word could have different meanings depending on the context, using the example sentences “She is kind” and “What kind of house do you live in?” with the word kind in bold. Participants were asked to read each sentence carefully and translate it into Japanese. They were encouraged to translate as much as possible even if they did not understand every word in each sentence. They were told that their participation was voluntary and that the vocabulary tests would not affect their final grades.

Analyses

The analysis of these data presented challenges because the aim was not only to investigate whether vocabulary depth can reasonably be regarded as a unidimensional construct, but also to examine the relative difficulty of four components of depth (form–meaning link, collocation, word parts, and polysemy). In addition, because marking translations from English to Japanese involved subjective judgments, it was necessary to include multiple raters to evaluate whether judgments could be applied consistently based on a scoring guide.

The many-facet Rasch model (MFRM; Linacre, 1989), implemented in FACETS version 3.18.2 (Linacre, 2017), was employed because it allows four facets—person ability, word difficulty, subtest difficulty, and rater severity—to be calibrated concurrently. Person measures and person fit indices can be used to evaluate whether the test is well matched to the ability of the test-takers and whether they respond consistently. The relative difficulty of the subtests provides insight into the construct, while subtest fit indices indicate whether subtest difficulty is consistent across words and test-takers. Rater severity and rater fit indices show whether a scoring guide can be applied consistently and whether multiple raters are required. Finally, relative word difficulty can be examined to determine whether it aligns with expectations, such as lower-frequency words generally being more difficult.

The Rasch model is also robust to missing data, as a perfectly complete data matrix is not required (Bond et al., 2020). In this study, some of the 120 participants did not complete all four vocabulary tests. However, as long as they took at least one test, their data could be included. The scale was oriented so that persons of greater ability received higher logit scores. Higher logits were also assigned to more difficult words and subtests, as well as to more severe raters.

Some alternative approaches were also considered. For example, a dichotomous Rasch analysis would treat the four components of vocabulary knowledge for each word as separate items. A problem with this approach is that because four items are derived from each word, local dependencies arise among the items.

Another option was a testlet model (Wainer & Kiely, 1987), in which a set of four items is scored from 0–4 depending on how many were answered correctly and treated as a polytomous rating scale. A limitation of this approach is that a testlet model does not establish an order of difficulty among items

within a set. In the present study, however, whether the subtests and subtest items exhibit a consistent order of difficulty was of particular interest.

Given the inclusion of raters and four components of vocabulary depth, a many-facet Rasch analysis (MFRA) was judged most appropriate. First, MFRA can estimate rater severity and determine whether multiple raters are necessary. Second, MFRA can control for the lack of local independence among items (Bond et al., 2020). In this study, the primary unit of analysis was knowledge of the "word," examined from the four perspectives of the vocabulary depth tests. This is analogous to a performance test evaluated by four raters. Finally, MFRA allows item difficulty differences—reflecting varying levels of vocabulary depth—to be estimated stochastically, unlike a polytomous rating scale. Item fit estimates can then be examined to assess how closely the items, across four subtests, approximate a unidimensional measurement model.

FACETS produces two fit statistics, Infit and Outfit (Wright & Masters, 1981), to evaluate all facets—Test, Word, Rater, and Persons. Infit mean-square (Infit MNSQ) statistics are information-weighted and are sensitive to unexpected response patterns near the person's ability or the item's difficulty level. Outfit mean-square (Outfit MNSQ) statistics are unweighted standardized residuals and are more influenced by outliers—that is, responses far above or below the subtest's or word's difficulty (Bond & Fox, 2007). Infit and Outfit values of 1.00 indicate perfect model fit. In this study, values above 1.50 or below 0.50 were considered misfitting (Linacre, 2017). Person and word fit were assessed primarily with Infit MNSQ statistics, as they pose a greater threat to measurement quality (Linacre, 2010).

Results

Facets Analysis of the combined four subtests

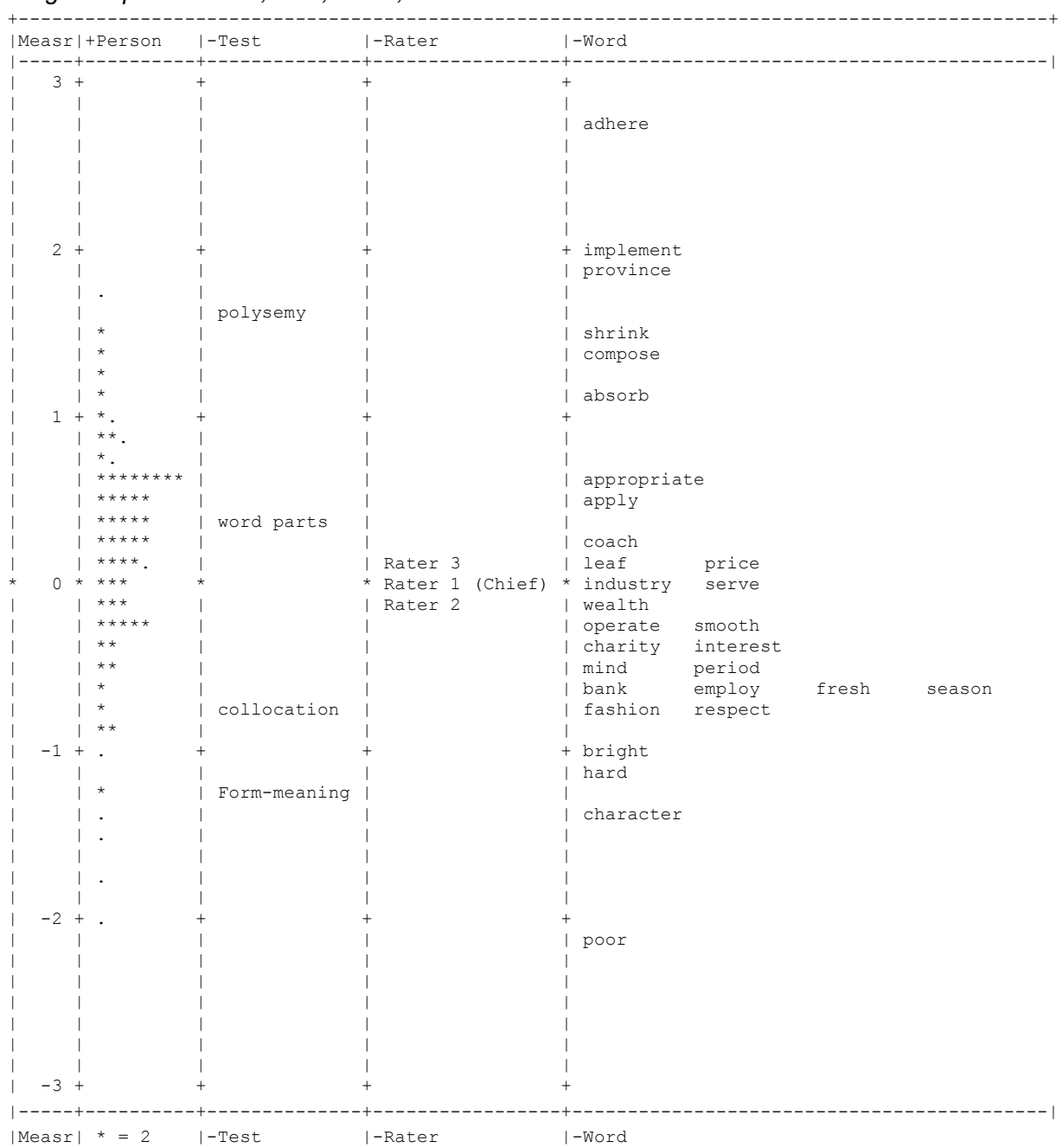
The Wright map in Figure 1 displays the four facets from left to right: person-ability (Person), test difficulty (Test), rater severity (Rater), and word difficulty (Word). The logit scale is shown on the far left-hand column. Positive values indicate high ability for persons, difficulty for tests and words, and severity for raters, whereas negative values indicate low ability, easiness, and leniency. MFRM requires that all facets save one be set to have an average Rasch measure of zero, whereas the remaining facet, which is specified as noncentered, is allowed to vary in relation to the other facets. Following standard practice, in this analysis, person ability was specified to be noncentered, while word difficulty, test difficulty, and rater severity were centered at zero.

Because the four vocabulary tests employed in this study used meaning-recall formats where students translated each English sentence into Japanese, answers could vary depending on individuals. Therefore, a scoring guide was created to guide raters, two additional raters were enlisted to mark a subsample of scores, and the severity and consistency of raters was examined. The first author marked all the answers and the two additional raters were asked to mark the answers of different groups of five students each according to the scoring guide. Each set of five students included two higher-ability, one middle-ability, and two lower-ability students.

As shown in Figure 1, the three raters were approximately equal and located at average severity, which indicates that the scoring guide for the vocabulary tests worked well in terms of providing uniform guidance to raters. The FACETS output also revealed that the three raters had approximately 95% exact agreement and Infit MNSQ statistics close to 1.00, an indication that their ratings were very consistent.

Figure 1

Wright Map for Person, Test, Rater, and Word



In addition, the low separation index of 0.63 and low reliability of .29 of the rater severity facet indicated that the three raters were very similar in their ratings, and the fixed-effects X^2 of the rater facet demonstrated that the three raters were not statistically different in marking ($X^2 = 2.0$, $df = 2$, $p = .36$). Additionally, the results of a follow up interaction analysis between rater severity and test difficulty

showed that the effects were neither measurably different nor statistically significant with *t*-values less than ±2 and *p*-values greater than .05 for each rater on each test, suggesting that the observed differences were most probably due to chance (X^2 = 2.6, *df* = 12, *p* = 1.00). Therefore, subsequent analyses are reported based on the chief rater (first author), who marked all of the participants' answers.

Person Measurement

Figure 1 shows that person abilities are approximately normally distributed, with the majority of persons falling within +1 to -1 logit range. Table 2 displays the summary of the person ability measures. The mean person measure is slightly higher than zero (0.16), indicating that tests and items were somewhat easy from the students' perspective, but overall, the four tests were well-targeted for the sample of learners. Person fit statistics showed that all persons' Infit MNSQ values were within the accepted range of 0.50 to 1.50, which means that the person measures in this sample had good fit to the Rasch model, supporting the hypothesis that the four tests together can reasonably be regarded as a unidimensional variable (RQ 1). The separation index of 2.36 and the reliability score of .85 mean that the four tests combined have good reliability and that similar results could be expected if these persons took a test comprised of similar items. The tests are capable of distinguishing between two groups of person abilities. The fixed-effects X^2 test shows that these persons were statistically different (X^2 = 851.3, *df* = 119, *p* < .001).

Table 2

Summary of the Person Ability Measures

M	*SD*	*Min*	*Max*
0.16	0.68	-1.98	1.76

Note. Reliability = .85. Separation = 2.36

Test Measurement

Table 3 shows the test measurement report. As can also be seen in Figure 1, the results showed that the form-meaning link test was the easiest (logits = -1.30). The collocation test was the second easiest (logits = -0.77), followed by the word parts test (logits = 0.41). The polysemy test was the most difficult (logits = 1.67).

Table 3

Test Measurement Report

Test	Observed Average	Difficulty Measure	*SE*	Infit MNSQ	Outfit MNSQ	Estim Discm	Point measure *r*
Polysemy	.24	1.67	0.04	1.05	1.28	0.87	.33
Word parts	.47	0.41	0.04	1.11	1.13	0.71	.39
Collocation	.69	-0.77	0.05	0.91	0.90	1.14	.55
Form-meaning link	.76	-1.30	0.05	0.84	0.78	1.19	.58

Note. SE = standard error; MNSQ = mean square; Estim Discm = Estimated Discrimination.
Because items were scored dichotomously (0 or 1), the observed average of a test is equivalent to the percentage of items answered correctly. Difficulty was measured in logits.

As for test difficulty estimates, as shown in Table 3, Infit and Outfit MNSQ values for the four subtests were all close to 1.0. This indicates that the four subtests, which consist of different items based on the

same words, show consistency in their behavior. Overall, for each word, the Polysemy item is consistently more difficult than the Word parts item, which is more difficult than the Collocation item, which is in turn more difficult than the Form-meaning link item, which is the easiest. Exceptions will be discussed below. The separation index of 29.50 and the reliability score of 1.00 were very high, which indicates that variance in test difficulty was adequately large to locate the test along the Rasch measure continuum. The fixed-effects X^2 test shows that these four vocabulary tests were statistically different (X^2 = 2523.2, df = 3, $p < .001$).

Word Measurement

In terms of word difficulty estimates, Table 4 shows that *adhere* (Item 29) was the most difficult word at 2.79 logits, whereas *poor* (Item 10) was the easiest at -2.25 logits. Infit MNSQ values were all within the accepted range of 0.50 to 1.50 with the exception of item 5, *hard*, which had an Infit MNSQ of 1.76. When Outfit MNSQ values were examined, four words were out of the accepted range of 0.50 to 1.50: *implement* (Item 26), *province* (Item 23), *apply* (Item 18), and *hard* (Item 5), which had an Outfit MNSQ of 1.74, 2.61, 1.58, and 1.90, respectively. The point measure correlation, which is the correlation between the word and the total score, is very low for *province* (r = .08), which implies that correct translation on *province* seems unrelated to correct translation on other items. This low correlation may be related to the high Outfit MNSQ value for this item. The separation index of 8.90 and the reliability score of .99 mean that variance in word difficulty was adequately large to precisely locate the words along the Rasch continuum. In addition, the fixed-effects X^2 test shows that the 30 vocabulary items were statistically different (X^2 = 1895.4, df = 29, $p < .001$).

In sum, although most of the words elicited consistent response behavior from the participants and were capable of measuring their abilities, there were some problematic examples that drew unexpected responses. The possible causes of the misfitting words will be further discussed in the discussion section.

Principal Component Analysis of Rasch Residuals

In order to further examine the dimensionality of the four tests combined using Rasch fit statistics, the FACETS measures for Facet 1 (Person) and Facet 2 (Test) were exported and a principal component analysis of the Rasch residuals (PCAR) was conducted in WINSTPEPS 4.4.7 (Linacre, 2019). When conducting a PCAR, the Rasch measure accounts for the first major dimension in the data, and no systematic relationships should be present in the residuals if the data are unidimensional and fit the Rasch model. In other words, when the data fit the Rasch model, the residuals should be random. The results of the PCAR revealed that the Rasch model accounted for 86.6% of the total variance, which is well above the 60% criterion proposed by Linacre (2007). The unexpected variance in the first contrast was 6.7% (eigenvalue = 2.0), below the threshold of 10% and eigenvalue greater than 2.0 suggested by Linacre, further supporting unidimensionality of the construct (Linacre, 2020). Thus, the results of the PCAR suggest that the four vocabulary tests measure a unidimensional construct which we might label “depth of vocabulary knowledge.”

Table 4

Word Measurement Report Ordered from Most to Least Difficult

Word	Observed Average	Difficulty Measure (in logits)	*SE*	Infit MNSQ	Outfit MNSQ	Estim Discm	Point measure *r*
adhere	.11	2.79	0.16	1.03	1.04	0.98	.29
implement	.20	2.01	0.13	1.27	**1.74**	0.56	.17
province	.21	1.92	0.13	1.31	**2.61**	0.33	**.08**
shrink	.26	1.59	0.13	0.94	0.78	1.13	.49
compose	.28	1.40	0.12	1.00	0.89	1.03	.46
absorb	.33	1.11	0.12	1.00	1.21	0.76	.39
appropriate	.42	0.60	0.12	0.92	0.82	1.19	.57
apply	.44	0.48	0.11	1.42	**1.58**	0.16	.24
coach	.49	0.23	0.11	0.68	0.62	1.61	.72
leaf	.50	0.14	0.11	0.97	0.88	1.09	.55
industry	.51	0.09	0.11	0.77	0.68	1.46	.68
price	.52	0.07	0.11	0.72	0.68	1.52	.70
serve	.53	-0.03	0.12	1.29	1.42	0.44	.34
wealth	.55	-0.13	0.12	1.09	1.03	0.87	.48
smooth	.58	-0.30	0.12	0.94	0.95	1.10	.56
operate	.58	-0.31	0.12	1.16	1.31	0.67	.41
charity	.59	-0.34	0.12	0.88	0.83	1.23	.61
interest	.61	-0.44	0.12	0.80	0.81	1.33	.64
mind	.62	-0.52	0.12	1.36	1.40	0.41	.31
period	.62	-0.53	0.12	1.15	1.03	0.81	.45
bank	.63	-0.58	0.12	1.21	1.14	0.70	.41
employ	.63	-0.59	0.12	0.72	0.66	1.46	.69
season	.63	-0.60	0.12	0.68	0.57	1.54	.72
fresh	.65	-0.67	0.12	0.63	0.51	1.61	.75
fashion	.66	-0.75	0.12	0.74	0.66	1.41	.67
respect	.67	-0.83	0.12	0.78	0.77	1.33	.64
bright	.70	-1.04	0.12	0.76	0.64	1.37	.66
hard	.71	-1.08	0.12	**1.76**	**1.90**	-0.07	.05
character	.76	-1.43	0.15	0.71	0.50	1.40	.66
poor	.86	-2.25	0.15	1.28	1.29	0.74	.23

Note. Because items were scored dichotomously (0 or 1), the observed average of a test is equivalent to the percentage of test-takers who answered the item correctly.

Multiple Components of Vocabulary Knowledge

To evaluate how the four tests function as separate measures, FACETS analyses were conducted on each test. These analyses could potentially be useful for researchers and teachers who are interested in measuring an individual component of vocabulary knowledge such as collocations or polysemy, independently.

Separation and Reliability

Table 5 shows person and item separation and reliability for the four tests. The person separation indices ranged from 1.49 to 1.89 and person reliability from .69 to .78. These results suggest that the 30 items used on each test may not have sufficient precision to separate the test-takers into two statistically significant groups, which may be due to the participants' homogeneous vocabulary knowledge. Item separation estimates ranged from 3.80 to 5.50 and the item reliability ranged from .94 to .97, which means that the 30 items in each test have three to five statistically distinct levels of item difficulty which can be distinguished by the persons in this sample.

Table 5

Separation and Reliability of the Vocabulary Tests

Vocabulary test	Person Separation	Person Reliability	Item Separation	Item reliability
Form-meaning link (*n* = 99)	1.65	.73	3.80	.94
Collocation (*n* = 102)	1.89	.78	5.01	.96
Word parts (*n* = 107)	1.78	.76	5.50	.97
Polysemy (*n* = 115)	1.49	.69	4.55	.95

Person and Item Measurement

Table 6 shows statistics for person measurement for the four individual tests. As for person fit statistics, eight out of 99 participants (8.08%), eight out of 102 participants (7.84%), four out of 107 participants (3.74%), and eight out of 115 participants (6.96%) were considered underfitting (Infit MNSQ > 1.50) for the form-meaning link, the collocation, the word parts, and the polysemy tests, respectively, all within 10%.

Table 6

Summary Statistics of Person Measurement

Vocabulary test	*M* (in logits)	*SD*	Range of Infit MNSQ	Range of Outfit MNSQ
Form-meaning link (*n* = 99)	1.95	1.24	0.41-2.19	0.09-9.00
Collocation (*n* = 102)	1.51	1.21	0.51-1.98	0.33-9.00
Word parts (*n* = 107)	-0.33	1.02	0.54-1.86	0.38-7.11
Polysemy (*n* = 115)	-2.21	1.22	0.51-2.17	0.03-4.47

In terms of item fit statistics, item Infit and Outfit MNSQ statistics were examined for each individual test. Infit MNSQs were within the acceptable range of 0.50 to 1.50 for all four tests. Many instances of out-of-range Outfit MNSQs were identified in the individual analyses, but they generally occurred with very easy items and were likely the result of test-takers' carelessness, or possibly the scoring criteria. Observed average, difficulty measure, Infit and Outfit MNSQs for the four tests are reported in Appendix B.

Differences in Item Performance among the Subtests

As mentioned earlier, generally, the order of difficulty of the 30 items shows a more or less similar pattern across the four tests. For example, *adhere* (Item 29) and *province* (Item 23) appear to be relatively difficult across all four tests, although not exactly the same. Some small changes in the order of difficulty might be expected. However, one thing worth mentioning is that *coach* (Item 16) changes its rank of difficulty drastically across the four tests. Although it was relatively easy both in the form-meaning link and the collocation tests, it was the most difficult item in the word parts test. The possible reason for this result will be discussed in the discussion section.

In sum, the high item reliability and separation show that the persons in this sample varied sufficiently in their ability to produce stable and accurate measures of item difficulty. Furthermore, the good item fit statistics indicate that the 30 items in each test have a reasonable fit to a Rasch hierarchical model of item difficulty. Thus, each of the four vocabulary tests could be used independently to measure a specific vocabulary component if a researcher had such a need.

Correlation and Factor Analysis

A correlation analysis followed by exploratory factor analysis using JASP version 0.18.1 (JASP Team, 2023) was conducted to further examine unidimensionality from a different, non-Rasch perspective. Although both Rasch analysis and factor analysis are employed to examine dimensionality, they are different in that the former is a hierarchical implicational model, whereas the latter is a correlational model (Sick, 2011). Checking the box plots revealed two outliers, and a temporary removal of the outliers led to histograms approximately normally distributed with the exception of the collocation measures. Thus, a non-parametric Spearman's *rho* was run. As shown in Table 7, the correlations between the four vocabulary knowledge components lay between .594 and .727, which means that the vocabulary knowledge components were moderately yet significantly related.

Table 7

Correlations between vocabulary knowledge components

	Polysemy	Word Parts	Collocation	FML
Polysemy	—			
Word Parts	.652*	—		
Collocation	.594*	.685*	—	
FML	.628*	.673*	.727*	—

*Note. Spearman's rho: *p < .001, N = 83.FML = Form-meaning link.*

Before performing factor analysis, two assumptions were examined. First, the overall MSA value in the Kaiser-Meyer-Olkin test was 0.821, exceeding the recommended value of 0.80. Second, the Bartlett's Test reached statistical significance (X^2 = 208.02, *df* = 6, $p < .001$), supporting the factorability of the correlation matrix. These two analyses support the suitability of the data for an exploratory factor analysis. A factor analysis using the minimum residual factoring method with Promax rotation found that all four tests loaded highly on a single factor, with only the polysemy measure showing a substantial degree of unique variance, as shown in Table 8. Table 8 indicates that 70.1% of the variance can be explained by a single factor, which is the combination of form-meaning link, collocation, word parts, and polysemy (eigenvalues = 3.098).

Table 8

Factor Loadings

	Factor 1	Uniqueness
Collocation	0.881	0.224
Word Parts	0.859	0.262
FML	0.833	0.306
Polysemy	0.772	0.405

Note. Applied rotation method is promax.

The chi-squared test for a single factor solution ($X^2 = 6.56$, $df = 2$, $p = .038$) was non-significant at $p < .01$, indicating that the data did not significantly differ from a hypothesized single factor model. In addition, a forced 2-factor solution found that a second factor had an eigenvalue of only 0.472, well below what might be expected from random data, as simulated in a parallel analysis. Thus, the factor analysis is also supportive of unidimensionality with some indication that there may be unique variance introduced by some items, which might be improved with item revision. This finding is in line with the Rasch analyses, which found high Outfit MNSQs for some items. In short, unidimensionality of the four vocabulary components was supported both by Rasch analysis and factor analysis.

Discussion

The results of PCAR showed that the Rasch model accounted for 86.6% of the total variance, demonstrating the unidimensionality of the four components of vocabulary knowledge. RQ1 is thus answered affirmatively. Regarding RQ2, the person reliability was .85, the test reliability was 1.00, and the word reliability was .99. Person and word fit measures showed that all persons' Infit MNSQ values were within the accepted range of 0.50 to 1.50. Despite the presence of some misfitting words when Outfit MNSQ values were considered, the person and word measures in this sample had good fit to the Rasch model in general. As for the reliability and validity of the four subtests analyzed individually as posed in RQ3, the person reliability ranged from .69 to .78 and the item reliability ranged from .94 to .97. The percentage of misfitting persons was below 10% for all subtests. As for item measures, Infit MNSQs were all within the acceptable range of 0.50 to 1.50 for all four tests. Although there were some misfitting items when Outfit MNSQs were considered, the person and item measures in this sample had good fit to the Rasch model in general.

Lastly, the correlations among the four vocabulary components ranged from .594 to .727 ($p < .001$), and all components were significantly related to each other. In addition, an exploratory factor analysis supported the unidimensionality of the four components of vocabulary knowledge. Therefore, RQ 4 is answered affirmatively. These results are in line with González-Fernández and Schmitt's (2019) findings that different vocabulary components are strongly related and make a substantial contribution to the explanation of the vocabulary knowledge construct.

One disadvantage of the MFRA employed in this study was that it assumes a consistent distance in difficulty among the four measures of vocabulary depth for each of the 30 words. Although differences in difficulty among the vocabulary depth tests may generally be consistent, the relative difficulty or even the order of difficulty could vary for individual words. These inconsistencies could stem from various sources, including loanword influence, where and how students learned a word, or faulty item design, and would likely manifest as misfit in the Word facet.

To take one example, *coach* (Item 16) showed inconsistent results. As noted in the Results section, the difficulty of *coach* changed drastically across the four tests. It was the most difficult item in the word parts test, whereas it was relatively easy in the form-meaning link and the collocation tests. A possible reason for this result is the loanword effect. *Coach* is a loanword, so the meaning of *coach* tested on the form-meaning link and the collocation tests seemed to have been especially easy for the participants. On the word parts test, in contrast, *coachable* as an adjective was a target word in the sentence 'He is *coachable*.' Because *coachable* is not a loanword, the difficulty level rose sharply. Many participants thought that *he* in 'He is *coachable*' is a person who coaches someone, not a person who is coached. As this example shows, a close examination of items can be insightful. In the next section, I describe possible reasons for misfitting items and how they can be revised to improve the tests.

Misfitting Items and Possible Revisions

As shown in Table 4, *hard* (Item 5) was underfitting (Infit MNSQ = 1.76, Outfit MNSQ = 1.90). *Hard* as 'solid' in the form-meaning link test was less known to the participants than *hard* as 'difficult' in the polysemy test. A Bias/Interaction analysis between test and item difficulty showed that the average observed score of the item on the polysemy test was 1, where all participants answered the item correctly, which was higher than that of the form-meaning link test (0.83). The reason why *hard* meaning 'difficult' was particularly easy for the participants could be because that meaning is a loanword and is much more familiar to Japanese students. Thus, switching tested meanings between the form-meaning link test and the polysemy test would be appropriate if the tests are to be used together to measure vocabulary depth of Japanese L2 learners.

An examination of Outfit MNSQs revealed that *province* (Item 23) was underfitting (Outfit MNSQ = 2.61) with a point measure correlation of .08 and *implement* (Item 26) was underfitting (Outfit MNSQ = 1.74) with a point measure correlation of .17. A Bias/Interaction analysis between test and item difficulty showed that the average observed score of *province* on the polysemy test (0.36) was higher than the score on the form-meaning link test (0.14). As for *implement*, the average observed score on the word parts test (0.31) was higher than the score on the form-meaning link test (0.17). The reason why *province* as 'a person's particular area of knowledge' was easier for the participants than *province* as 'one of the areas that some countries are divided into with its own local government' might be because the latter meaning is especially unfamiliar to Japanese students. In Japan, the word prefecture is used to express an area governed by a local government much more frequently. Thus, in future revisions to the test, the inclusion of the word *province* needs to be reconsidered. As for *implement*, the reason why *implementation* tested in the word parts test was easier for the participants than *implement* as a verb tested in the form-meaning link test might be because the context helped them understand the meaning of the tested word. *Implementation* was tested in the sentence, 'They agreed on the *implementation* of the new system,' and *implement* was tested in the sentence, 'Schools *implement* plans every year.' The phrase of 'the new system' might have become a hint for the participants. Therefore, in future revisions to the test, the word 'system' could be changed to 'plan.'

Lastly, for Item 18, *apply* (Outfit MNSQ = 1.58), the collocation test was the most difficult. This might be because the tested meaning was different on the form-meaning link test and on the collocation test, whereas for the other items, the tested meaning was the same on the two tests. 'You can *apply* online' was used on the form-meaning link test, and 'New rules will *apply'* was used on the collocation test. Thus, to make all the items consistent across the four tests, this item needs to be revised so that the same meaning would be tested in the two tests.

Conclusion

This study provided preliminary validity evidence for four vocabulary tests designed to measure four components of vocabulary knowledge: form-meaning link, collocation, word parts, and polysemy. The results of a many-facet Rasch analysis (MFRA) showed that the four tests can reasonably be combined to form a single measure with adequate reliability. Thus, the tests can be used either in combination or independently as measures of depth of vocabulary knowledge. Nevertheless, analysis revealed that some items were problematic, most likely due to the influence of English loanwords in Japanese. Future revisions of the test may need to reconsider the inclusion of these words or perhaps reconsider which meaning should be regarded as the most frequent. The findings were based on Japanese learners of English, so future studies should employ MFRA with learners from different first-language backgrounds to examine generalizability.

Acknowledgements:
The authors sincerely appreciate the reviewers' insightful feedback. They are also grateful to the two volunteers who marked the vocabulary tests.

Declaration of competing interests:
The authors have declared no competing interests.

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129. https://doi.org/10.1093/applin/14.2.115

Allwright, D. (2005). Developing principles for practitioner research: The case of exploratory practice.

Kilgour, J. M., & Tayyaba, S. (2016). An investigation into the optimal number of distractors in single-best answer exams. *Advances in Health Science Education, 21*, 571–585. https://doi.org/10.1007/s10459-015-9652-7

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). International Reading Association.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International journal of Lexicography*, *6*(4), 253–279. https://doi.org/10.1093/ijl/6.4.253

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum Associates Publishers.

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge. https://doi.org/1004324/9780429030499

Chui, A. S. Y. (2006). A study of the English vocabulary knowledge of university students in Hong Kong. *Asian Journal of English Language Teaching*, *16*, 1−13.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213−238. https://doi.org/10.2307/3587951

Davies, M. (2008). The corpus of contemporary American English: 450 million words, 1990-present. Retrieved from http://corpus.byu.edu/coca/

Fisher Jr., W. P. (2007). Rating scale instrument quality criteria. *Rasch measurement transaction, 21*(1), 1095. http://www.rasch.org/rmt/rmt211a.htm

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

González-Fernández, B., & Schmitt, N. (2019). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components, *Applied Linguistics*, *41*(4), 481–505. https://doi.org/10.1093/applin/amy057

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524773

JASP Team (2023). JASP (version 0.18.1) [Computer software].

Komiya, C., & McLean, S. (2024). Which type of vocabulary-knowledge better correlates with reading comprehension, meaning-recognition or meaning-recall? *RELC Journal*, Advance online publication. https://doi.org/10.1177/00336882241246619

Kremmel, B., & Pellicer-Sánchez, A. (2020). Measuring vocabulary development. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 211–222). Routledge.

Li, M., & Kirby, J. (2015). The effects of vocabulary breadth and depth on English reading *Applied Linguistics*, *36*(5), 611–634. https://doi.org/10.1093/applin/amu007

Linacre, J. M. (1989). Many-facet Rasch Measurement. Chicago, IL: MESA Press.

Linacre, J. M. (2007). Dimensionality: Contrasts and variances [online]. Available: https://www.winsteps.com/winman/principalcomponents.htm.

Linacre, J. M. (2010). Predicting responses from Rasch measures. *Journal of Applied Measurement, 11*(1), 1–10.

Linacre, J. M. (2017). *Facets computer program for many-facet Rasch measurement, version 3.81.2.* Beaverton, OR: Winsteps.com.

Linacre, J. M. (2019). *Winsteps® Rasch measurement computer program* (Version 4.4.7). Winsteps.com.

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

Nation, I. S. P. (1983). Testing and Teaching Vocabulary. *Guidelines*, *5*, 12–25.

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

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13. https://openaccess.wgtn.ac.nz/articles/journal_contribution/A_vocabulary_size_test/12552197

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

Read, J., & Dang, T. N. Y. (2022). Measuring depth of academic vocabulary knowledge. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688221105913

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913–951. https://doi.org/10.1111/lang.12077

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88. https://doi.org/10.1177/026553220101800103

Schmitt, N., & Schmitt, D. (2020). *Vocabulary in language teaching.* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108569057

Sick, J. (2011). Rasch measurement in language education: Part 6. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *15*(1), 15–17. Retrieved from https://hosted.jalt.org/test/sic_6.htm.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185–201. https://doi.org/10.1111/j.17453984.1987.tb00274.x

Webb, S., Sasao. Y., & Balance, O. (2017). The updated vocabulary levels test: Developing and validating two new forms of the VLT. *ITL—International Journal of Applied Linguistics*, *168*(1), 34–70. https://doi.org/10.1075/itl.168.1.02web

Wright, B. D., & Masters, G. (1981). The measurement of knowledge and attitude (Research Memorandum No. 30). Chicago: University of Chicago, Department of Education, Statistical Laboratory.

Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, *26*(4), 696–725. https://doi.org/10.1177/1362168820913998

Appendix A

Four Vocabulary Tests

1. Form-meaning Link

英語は、1 つの単語に色々な意味がある場合があり、文によって意味が異なります。
例えば
She is **kind.**は「**彼女は親切**です。」という意味です。
一方
What **kind** of house do you live in?は「どんな**種類**の家に住んでいますか?」という意味です。

各文を注意深く読んで、日本語にしましょう。
＊もし文の全部が分からなかったら、分かるところまで書きましょう。
(例) It can accommodate 100 people.　　それは 100 人を　　できる。

1. I don’t know the price.
2. I have an interest in music.
3. He has a healthy mind.
4. I don’t like that bank.
5. This desk is hard.
6. They serve drinks from 10 a.m.
7. How many people do they employ?
8. This room is very bright.
9. The food is fresh.
10. He is poor.
11. I like this season.
12. I like her character.
13. He waited for a long period.
14. I like to follow fashion.
15. I have respect for you.
16. He is my coach.
17. What is the main industry in Japan?
18. You can apply online.
19. They can operate in sunlight.
20. This is very smooth.
21. She has a lot of wealth.
22. It is a charity for children.
23. It is a province in Canada.
24. She composed a letter.
25. This absorbs water.
26. Schools implement plans every year.
27. It is an appropriate choice.
28. This leaf is beautiful.
29. They adhere to what I say.
30. These clothes often shrink.

2. Collocation

英語は、1 つの単語に色々な意味がある場合があり、文によって意味が異なります。
例えば
She is **kind**.は「彼女は**親切**です。」という意味です。
一方
What **kind** of house do you live in?は「どんな**種類**の家に住んでいますか?」という意味です。

各文を注意深く読んで、日本語にしましょう。
＊もし文の全部が分からなかったら、分かるところまで書きましょう。
(例) It can accommodate 100 people.　　それは 100 人を　　できる。

1. I don't want to pay a high price for this.
2. She showed an interest in tennis.
3. Nature clears my mind.
4. This is my bank account.
5. This animal has a hard shell.
6. They serve meals all day.
7. She is employed part-time by the company.
8. I like bright colors.
9. I need fresh air.
10. They come from poor families.
11. Baseball season will start soon.
12. She is the main character in this book.
13. You have to pay after a trial period.
14. Look at the latest fashion.
15. Try to gain the respect of people.
16. The team fired their coach.
17. I want to work in the music industry.
18. New rules will apply.
19. Please tell me how to operate the machine.
20. It was smooth sailing.
21. She is trying to build wealth.
22. There is a local charity here.
23. This is a central province.
24. I want to compose songs.
25. This can absorb heat.
26. They implement the program.
27. This is an appropriate level.
28. Please give me a lettuce leaf.
29. I adhere to the rules.
30. The economy is shrinking.

3. Word Parts

英語は、1 つの単語に色々な意味がある場合があり、文によって意味が異なります。
例えば
She is **kind**.は「彼女は**親切**です。」という意味です。
一方
What **kind** of house do you live in?は「どんな**種類**の家に住んでいますか?」という意味です。

各文を注意深く読んで日本語にしましょう。
＊もし文の全部が分からなかったら、分かるところまで書きましょう。
(例) It can accommodate 100 people.　　それは 100 人を　　できる。

1. This is a pricey restaurant.
2. Interestingly, there were a lot of people.
3. Be mindful of others.
4. She is a bankable star.
5. What was the water hardness?
6. He is a server.
7. He is my employer.
8. The sun brightened up the room.
9. It is a freshly printed book.
10. They live in poverty.
11. They are seasonal workers.
12. The banana has a characteristic smell.
13. Periodic checks are important.
14. They are fashionable clothes.
15. Be respectful of him.
16. He is coachable.
17. He is an industrial worker.
18. Please write an application.
19. The system is in operation now.
20. The interview went smoothly.
21. He is wealthy.
22. It is a charitable group.
23. It is a provincial food.
24. There were changes in the composition of the members.
25. This helps the absorption of water.
26. They agreed on the implementation of the new system.
27. He dressed appropriately for a party.
28. I like leafy vegetables.
29. His adherence is a surprise.
30. This T-shirt is shrinkable.

4. Polysemy

英語は、1 つの単語に色々な意味がある場合があり、文によって意味が異なります。
例えば
She is **kind**.は「彼女は**親切**です。」という意味です。
一方
What **kind** of house do you live in?は「どんな**種類**の家に住んでいますか?」という意味です。

各文を注意深く読んで日本語にしましょう。
＊もし文の全部が分からなかったら、分かるところまで書きましょう。
(例) It can accommodate 100 people.　　それは 100 人を　　できる。

1. Working late is often the price of success.
2. You have to pay the interest.
3. I don't mind the cold.
4. I walk along the river bank every morning.
5. This test is hard.
6. He served five years here.
7. They employed force to enter the building.
8. She is very bright.
9. Don't get fresh with me.
10. He is poor at sports.
11. I need to season this dish.
12. I cannot read Chinese characters.
13. We have history class third period.
14. She fashioned a cup by herself.
15. We are lucky in this respect.
16. I travel by coach.
17. He works with great industry.
18. Apply the cream to your leg.
19. We will operate on his eyes.
20. He is a smooth talker.
21. I have a wealth of choices.
22. She showed no charity towards her friends.
23. That question is outside my province.
24. He tried to compose himself.
25. He was absorbed in the book.
26. This is an implement.
27. They appropriated a lot of research money.
28. I will leaf through the book.
29. It did not adhere well.
30. He is seeing a shrink.

Appendix B

Measurement Report for the Four Tests

Item measurement report for the form-meaning link test (item facet)

Item	Observed Avg	Fair Avg	Difficulty Measure(in logits)	*SE*	Infit MNSQ	Outfit MNSQ	Estim Discm	Point measure *r*
adhere	.11	.07	4.55	.36	1.04	1.39	.95	.36
implement	.17	.12	3.91	.30	.98	.74	1.06	.47
province	.25	.20	3.29	.26	.89	.76	1.18	.53
compose	.40	.38	2.40	.23	.98	.96	1.05	.47
absorb	.46	.45	2.09	.23	.82	.75	1.51	.58
shrink	.46	.45	2.09	.23	.87	.85	1.35	.54
operate	.57	.58	1.58	.23	.90	.94	1.25	.50
appropriate	.68	.71	.99	.24	.86	.73	1.33	.52
apply	.69	.72	.93	.24	1.07	1.15	.82	.34
serve	.70	.74	.87	.24	1.13	1.19	.72	.30
wealth	.79	.83	.30	.27	.99	1.44	.94	.33
mind	.81	.85	.15	.28	1.29	1.53	.63	.13
hard	.83	.87	-.01	.29	1.01	1.01	.99	.33
smooth	.84	.88	-.10	.30	.96	.94	1.03	.35
industry	.86	.90	-.28	.31	1.00	.88	1.02	.33
price	.88	.92	-.49	.33	1.22	1.69	.76	.10
fashion	.89	.92	-.60	.34	1.11	.90	.96	.25
interest	.91	.94	-.85	.37	1.07	1.68	.90	.18
leaf	.91	.94	-.85	.37	1.00	.86	1.00	.28
charity	.92	.95	-1.00	.39	1.14	2.07	.85	.11
employ	.93	.96	-1.16	.41	1.06	.89	.98	.23
bank	.95	.97	-1.55	.48	1.00	1.05	.99	.21
poor	.96	.98	-1.80	.53	.93	.38	1.08	.30
coach	.96	.98	-1.80	.53	.99	.50	1.04	.25
bright	.97	.98	-2.12	.60	.85	1.81	1.03	.22
respect	.97	.98	-2.12	.60	.87	1.82	1.02	.21
fresh	.98	.99	-2.56	.73	.86	.26	1.09	.27
period	.98	.99	-2.56	.73	.81	.16	1.12	.31
season	.99	.99	-3.28	1.02	1.03	.70	.98	.08
character	1.00	1.00	-4.51	1.83	Minimum			.00

Item measurement report for the collocation test (item facet)

Item	Observed Avg	Fair Avg	Difficulty Measure (in logits)	*SE*	Infit MNSQ	Outfit MNSQ	Estim Discm	Point measure *r*
province	.14	.09	3.74	.31	1.00	.83	1.01	.34
implement	.20	.14	3.24	.27	1.10	1.05	.88	.28
adhere	.28	.23	2.67	.24	1.14	1.11	.74	.28
shrink	.30	.25	2.55	.24	.93	.99	1.09	.42
apply	.31	.26	2.50	.23	1.03	.97	.94	.37
absorb	.37	.33	2.18	.23	.83	.71	1.52	.54
leaf	.40	.36	2.03	.22	1.01	.94	1.00	.41
compose	.45	.42	1.79	.22	.89	.83	1.37	.51
wealth	.51	.49	1.50	.22	.88	.78	1.44	.53
appropriate	.53	.52	1.40	.22	.87	.79	1.43	.54
smooth	.63	.64	.89	.23	.95	.91	1.13	.48
serve	.66	.68	.73	.23	1.05	1.07	.91	.41
mind	.73	.76	.32	.25	1.05	1.23	.90	.39
interest	.75	.78	.20	.26	.97	1.03	1.02	.45
respect	.76	.80	.06	.26	1.04	1.08	.95	.40
period	.78	.82	-.08	.27	.84	.73	1.19	.55
charity	.78	.82	-.08	.27	1.18	1.18	.81	.31
hard	.79	.84	-.16	.28	1.07	1.21	.90	.36
industry	.79	.84	-.16	.28	1.17	1.17	.83	.31
operate	.90	.94	-1.24	.37	.83	.73	1.10	.46
bright	.92	.95	-1.53	.40	1.01	1.33	.99	.32
fashion	.92	.95	-1.53	.40	.93	.93	1.04	.38
poor	.93	.96	-1.70	.42	1.06	1.71	.91	.23
coach	.94	.97	-1.90	.45	.91	.47	1.09	.40
price	.95	.97	-2.12	.49	1.01	**2.91**	.89	.16
bank	.96	.98	-2.38	.54	1.05	**2.56**	.95	.21
employ	.97	.98	-2.71	.61	1.19	1.20	.88	.08
season	.98	.99	-3.16	.74	1.16	**3.97**	.79	**-.07**
character	.98	.99	-3.16	.74	1.09	.50	.99	.17
fresh	.99	1.00	-3.90	1.02	.86	.09	1.11	.27

Item measurement report for the word parts test (item facet)

Item	Observed Avg	Fair Avg	Difficulty Measure (in logits)	*SE*	Infit MNSQ	Outfit MNSQ	Estim Discm	Point measure *r*
coach	.01	.01	4.72	1.01	1.01	.79	.99	.06
bank	.03	.02	3.58	.59	.95	2.01	.99	.10
adhere	.04	.03	3.28	.52	.95	.43	1.07	.25
province	.07	.06	2.51	.38	.89	.77	1.07	.29
period	.08	.15	1.45	.27	.96	.88	1.05	.30
hard	.21	.18	1.25	.25	1.04	1.10	.93	.23
absorb	.25	.22	.96	.24	.84	.72	1.30	.45
shrink	.27	.24	.85	.23	.86	.80	1.27	.43
price	.30	.27	.69	.22	1.05	.98	.93	.28
compose	.30	.27	.69	.22	.95	.88	1.13	.37
implement	.31	.28	.64	.22	.97	.93	1.08	.34
mind	.34	.31	.50	.22	1.05	1.05	.86	.28
serve	.37	.36	.31	.21	1.14	1.40	.46	.18
apply	.39	.38	.22	.21	1.24	1.33	.11	.11
industry	.46	.45	-.09	.21	1.02	1.02	.91	.33
charity	.50	.50	-.27	.21	.93	.89	1.32	.43
appropriate	.51	.52	-.35	.21	.98	1.06	1.01	.37
season	.54	.55	-.49	.21	.97	.96	1.12	.40
poor	.60	.62	-.75	.21	.88	.83	1.40	.49
operate	.62.	.64	-.85	.22	1.07	1.19	.73	.29
wealth	.65	.68	-1.04	.22	.90	.91	1.22	.46
employ	.66	.69	-1.09	.22	1.19	1.28	.54	.20
fresh	.68	.71	-1.19	.23	1.05	1.07	.90	.34
leaf	.71	.74	-1.34	.23	.95	.90	1.10	.43
interest	.76	.79	-1.63	.25	.94	1.23	1.03	.40
character	.82	.86	-2.10	.28	.85	1.18	1.13	.49
bright	.84	.88	-2.25	.29	1.14	1.06	.90	.28
fashion	.85	.89	-2.34	.30	1.00	1.10	.98	.35
smooth	.88	.91	-2.62	.32	1.00	1.47	.94	.31
respect	.93	.95	-3.27	.41	.92	.65	1.07	.46

Item measurement report for the polysemy test (item facet)

Item	Observed Avg	Fair Avg	Difficulty Measure (in logits)	*SE*	Infit MNSQ	Outfit MNSQ	Estim Discm	Point measure *r*
price	.02	.01	2.36	.72	0.95	0.41	1.05	.18
fresh	.02	.01	2.36	.72	0.99	0.44	1.03	.15
industry	.02	.01	2.36	.72	0.97	0.40	1.04	.17
compose	.02	.01	2.36	.72	1.03	0.97	0.97	.06
appropriate	.02	.01	2.36	.72	1.03	1.39	0.96	.04
shrink	.02	.01	2.36	.72	0.97	0.40	1.04	.17
adhere	.03	.02	1.94	.59	1.02	0.74	1.00	.12
employ	.04	.03	1.39	.47	0.93	0.55	1.06	.25
fashion	.04	.03	1.39	.47	1.04	0.88	0.98	.14
smooth	.04	.03	1.39	.47	1.07	**2.30**	0.88	**.02**
leaf	.05	.03	1.19	.43	0.93	0.52	1.00	.27
interest	.08	.05	.73	.36	1.02	0.94	0.98	.20
respect	.10	.07	.49	.33	1.05	0.93	0.97	.21
season	.10	.07	.39	.32	0.87	0.55	1.13	.37
implement	.11	.08	.29	.31	1.02	0.88	0.99	.25
coach	.12	.09	.20	.30	1.10	1.08	0.91	.19
bright	.16	.11	-.13	.27	1.14	1.21	0.85	.19
charity	.22	.17	-.59	.24	1.01	0.91	1.01	.34
absorb	.25	.21	-.82	.23	0.94	0.77	1.14	.42
operate	.29	.24	-1.03	.23	1.21	1.46	0.53	.20
wealth	.29	.24	-1.03	.23	0.85	0.71	1.32	.50
character	.30	.26	-1.13	.22	0.97	0.84	1.10	.42
province	.36	.32	-1.41	.22	0.86	0.78	1.35	.51
apply	.39	.36	-1.60	.21	0.98	0.99	1.04	.43
serve	.43	.41	-1.82	.21	1.07	1.14	0.77	.37
period	.58	.60	-2.58	.21	1.07	1.02	0.87	.42
bank	.62	.64	-2.76	.22	1.11	1.21	0.74	.37
mind	.63	.66	-2.81	.22	0.95	0.99	1.08	.49
poor	.95	.98	-5.87	.48	0.82	0.35	1.12	.50
hard	1.00	1.00	-8.85	1.84	Minimum			.00