



## Developing Measures of L2 Korean Vocabulary Knowledge

**Daniel R. Isbell**

*University of Hawai'i at Mānoa*  
disbell@hawaii.edu

**Leeseul Park**

*University of Hawai'i at Mānoa*  
leeseul@hawaii.edu

**Meghan Delaney**

*University of Hawai'i at Mānoa*  
delaneym@hawaii.edu

### Abstract

Vocabulary knowledge plays an important role in language learning and language proficiency, but it is difficult to know how much of a role it plays without appropriate vocabulary measures. In this article, we first discuss learner vocabulary in our research on Korean as a second/foreign language. Unable to find appropriate measures for receptive vocabulary knowledge, we undertook the task of developing new tests: Park's *Korean Vocabulary Levels Test* and Delaney's *Korean Vocabulary Knowledge Test*. Although we needed to develop new tests from the ground-up, we were able to draw practical insights from (a) the large body of research on L2 English vocabulary and (b) existing tests of vocabulary knowledge tests for L2 English learners. We summarize the design of each test and initial/ongoing validation research. In sum, we have found these tests to have high reliability, generally desirable item statistics, and positive correlations with language skills that largely align with expectations.

**Keywords:** assessment, Korean, receptive vocabulary knowledge, meaning recall, meaning recognition.

## Introduction

Vocabulary is an important component of second language (L2) ability and communicative skills (Nation, 2013). While that statement is true for any L2, in reality, we know a lot more about the specifics of L2 vocabulary learning and its role in communicative skills for L2 English than other L2s. And before many questions about L2 vocabulary in other languages can begin to be investigated, it is first necessary to develop appropriate measures of vocabulary, which in turn depend on detailed study of a language's lexicon. Lexical resources and vocabulary measures can also be useful in applied settings, like language education and assessment.

As researchers of L2 Korean (and Korean teachers, in Park's and Delaney's cases), we have all become interested in L2 Korean vocabulary. Isbell's work on L2 Korean has focused on speaking and oral proficiency, and vocabulary has consistently popped up in interesting ways. In investigating the properties of a standardized measure of oral proficiency used for research, the Korean Elicited Imitation Test, Isbell and Son (2022) found that lexical sophistication of test stimuli (based on word frequency and presentation at different curricular levels; National Institute of the Korean Language, 2003) were a significant predictor of difficulty. Turning toward production, lexical diversity (operationalized through Guiraud's index, which accounts for variation in text length) was a significant, positive predictor of listener comprehensibility ratings, but not accentedness ratings (Isbell et al., 2023). In that study, speech rate – the number of words produced per unit of time; not traditionally considered a vocabulary/lexical measure, but indicative at some level of lexical processing efficiency – was a positive predictor of both accentedness and comprehensibility; in another study based on a subset of the same data, speech rate also emerged as the strongest predictor of listener fluency ratings (Isbell & Lee, 2023).

The importance of vocabulary in L2 Korean speaking ability has manifested clearly in Isbell's research, and fortunately many interesting aspects of vocabulary are able to be measured in speech samples without needing to develop novel measures. Counting tokens allows fluency-related measures to be calculated, and counting types allows the calculation of a range of lexical diversity measures. Basic examination of things like lexical sophistication only requires some kind of principled vocabulary list or lexical database, based on frequency or other criteria (like the list from National Institute of the Korean Language, 2003). Park and Delaney have not had it so easy. Both Park and Delaney interests involve L2 Korean learner's vocabulary knowledge – a mostly stable, trait-like construct which cannot be directly observed. Measuring this vocabulary knowledge would require some kind of test.

Initially, we searched academic literature and the internet more broadly. We found nothing that seemed appropriate for (adult) L2 Korean learners. There are some Korean vocabulary tests used in early childhood education (e.g., short picture naming tests) and clinical applications (e.g., to screen for speech and developmental disorders), but these were not created to assess breadth of vocabulary knowledge. There are also tests of *hanja*, or Chinese characters, taken by Korean adolescents and adults, but these tests are not well-suited to assessing Korean vocabulary for L2 learners, who are typically not taught *hanja* in Korean courses and have little need for it outside of specialized goals (It would also be the case that L1 Japanese and L1 Chinese learners,

especially those from Taiwan, would have their vocabulary seriously overestimated on such tests!). In short, we could find nothing suitable for L2 Korean learners like you can find for L2 English on Paul Nation's website (<https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests>) or the IRIS database ([www.iris-database.org](http://www.iris-database.org)), among other places.

So, we set about making tests – initially to facilitate investigations into the role vocabulary plays in L2 Korean learning and proficiency, but also with potential pedagogical applications (e.g., placement, diagnosis) in mind. In what follows, we discuss some of our general design considerations for tests of L2 Korean vocabulary knowledge and introduce two recently developed tests of written receptive Korean vocabulary knowledge. For the Korean vocabulary tests, we share some select initial findings related to their psychometric qualities and relationships with other variables of interest. While our focus is on Korean, the issues we touch on are likely to be relevant to many non-English languages, and especially so-called less commonly taught languages, such as Japanese.

## Our Considerations for Designing Korean Vocabulary Tests

Several factors influenced our approach to designing new Korean vocabulary tests. First, and to be candid perhaps foremost, was practicality. To begin with assessing L2 Korean vocabulary knowledge, it made sense to create tests that were practical to administer and score and could be developed using available resources. This pointed us toward tests of written receptive knowledge that could be delivered easily, requiring only paper or easy-to-use online survey tools. Response formats that could be used with linguistically-diverse learners was also desirable. We also identified several Korean lexical resources and found Lee et al.'s (2017) *A Frequency Dictionary of Korean* to be both relevant to our goals and transparent in its methodology – it would inform the content selection and item construction for our tests.

We also wished to avoid reinventing the wheel. There are many existing vocabulary tests for other L2s, mostly for English but also some for languages like French and Spanish, and many of these have been extensively investigated. They have known properties, and following a similar design would allow us to (a) take advantage of what is known about the quality of those tests and variations in item format and administration, and (b) in theory support similar interpretations of test scores, if not allow for direct comparisons of scores themselves. The research on L2 English vocabulary tests made it clear that frequency, though not without its own limitations, should form the backbone of test design, and that tests which sampled few words per frequency band, included highly infrequent vocabulary items, and featured yes/no response formats especially susceptible to guessing would be inappropriate for our target population, which included beginner-level learners. Just as for other research instruments (e.g., EITs, see Isbell & Son, 2022), we saw building towards some degree of standardization in instruments across languages as valuable, so adopting the design/design elements of existing tests was a conscious choice.

Although frequency has been foundational in the design of many vocabulary tests, it is not without limitations. While Lee et al.'s (2017) frequency dictionary is a useful resource, the underlying corpus reflects the average word use of native Koreans,

which is not always going to line up with the language adult learners are exposed to. For example, the word 숙제 (sukje, *homework*) ranks 3,687 in frequency, making it a Band 4 word. However, for students learning Korean as a foreign language, the word 숙제 is, more likely than not, going to be a word they hear and use very frequently. Therefore, if in the creation of these tests 숙제 were to be randomly selected as a candidate for a Band 4 word, it would have been passed over as it would not be considered a good representative of a lower-frequency word. Vice versa, some words that are considered frequent based on a corpus meant to broadly represent text types adult native speakers are exposed to (e.g., the largest text type in Lee et al.'s 2017 is novels and children's books) may not match the word learning and exposure of learners. As Hashimoto (2021) pointed out, frequency does not explain everything about word difficulty in vocabulary tests, and there are clearly other lexical aspects at play when modeling learner vocabulary knowledge (e.g., Vitta et al., 2023). From a pedagogical perspective, He and Godfroid (2019) showed how the combination of human intuition and objective frequency information can result in a more meaningful list for teaching, and we believe the same can hold true for testing – and expert intuition might be especially valuable when, for example, psycholinguistic word norms are not readily available for a given language. Therefore, combining frequency information with teacher/expert intuition, rather than just relying on frequency information, informed vocabulary selections in our tests.

We also had to consider Korean-specific factors. For instance, there is some contention in Korean linguistics as to whether the language truly has adjectives as a distinct word class (as opposed to verbs with various inflections). Many Korean nouns can also be verbed, in particular nouns with Sino-Korean roots can function as verbs by adding -하다 (hada, *to do*), e.g., 일 (il, *work*) and 일하다 (ilhada, *to work*). We decided to treat adjectives as a distinct word class, and nouns and verbs with the same root as distinct, corresponding to different lemmas (though most would be part of the same *flemma*, McLean, 2018). These decisions are in line with, and influenced by, Lee et al. (2017). However, we did not include multiple items from the same *flemma* in the tests (e.g., a test could include 일/il or 일하다/ilhada, but not both). Relatedly, it was easy for us to rule out word families as a counting/sampling unit, drawing on research in English (e.g., McLean, 2018) and our experiences teaching and learning Korean, as it seemed highly unlikely we could assume *any* learner who knows the high-frequency 소리 (sori, *sound*) would know 헛소리 (heotsori, *nonsense* or *bullshit*, where the derivational prefix 헛, heot, *useless*, modifies 소리/sori, *sound*).

In the end, we landed on Webb et al.'s (2017) revised Vocabulary Levels Test (VLT) as the design template for our new Korean vocabulary tests. Webb et al.'s VLT is well documented, easily accessible, and supported by considerable validity evidence. It contains a large number of items per frequency band, which can support inferences about mastery for educational purposes, and focuses on the first 5,000 most frequent vocabulary items, which is well-suited to many L2 Korean learners, especially those at lower proficiency levels, and is conveniently aligned with the scope of a high-quality Korean lexical resource (i.e., Lee et al., 2017). We also note that its design, with some modifications, has been adapted to different stimuli types and response formats (e.g., McLean et al.'s Listening Vocabulary Levels Test), which we found encouraging.

## **Introducing Two Newly Developed Korean Vocabulary Tests**

Collectively, we have recently developed two tests of receptive Korean vocabulary knowledge, focused on form-meaning knowledge, that are based on the design of the (English) Vocabulary Levels Test (Webb et al., 2017). The first of these tests we describe is the Korean Vocabulary Levels Test (KVLТ), which assesses vocabulary knowledge through meaning recognition. Park was the lead developer and researcher of the KVLТ. The second test is the Korean Vocabulary Knowledge Test (KVKT), which assesses vocabulary knowledge through meaning recall. The KVKT was developed by Delaney, who is also carrying out ongoing research on the test as part of her dissertation. Isbell has a supervisory role in both projects, advising on design and validation efforts. In the following sections, we provide a brief overview of each test, including its purpose, design, and some highlights from initial/ongoing research.

### **Park's Korean Vocabulary Levels Test**

The purpose of the KVLТ is to assess learners' written receptive meaning-recognition vocabulary knowledge for research and low-stakes educational purposes. The KVLТ was administered online through Google Forms. An initial validation study focused on the KVLТ's measurement properties, internal structure, and relationship with language skills among 112 learners of Korean from Canada, Japan, Thailand, and the US (Park, 2022). Test directions were presented in the participants' L1, but test items were consistently presented in Korean for all subjects.

#### ***Design of the Test***

The KVLТ's design closely follows Webb et al.'s (2017) VLT and features a total of 150 items selected from the 5,000 most frequently occurring Korean words based on Lee et al. (2017). It is important to note, however, that unlike Webb et al.'s VLT, KVLТ items were sampled based on lemmas rather than word families. There are five subsections of the test, each with 30 items, organized by 1,000-word frequency bands: 1K (the most frequent 1–1000 word families), 2K (the most frequent 1001–2000 word families), 3K (the most frequent 2001–3000 word families), 4K (the most frequent 3001–4000 word families), and 5K (the most frequent 4001–5000 word families). The ratio of word types in each band approximates the overall ratio in Lee et al.'s (2017) frequency list: 19 nouns, 9 adjectives, and 2 verbs. Adverbs were excluded due to difficulty in writing simple, learner-friendly definitions in Korean.

Like the English VLT, KVLТ items are presented in a matching format, with clusters of three target words and two distractors (Figure 1). For each cluster, there are three short definitions written in Korean, and all target words and distractors are of the same type (noun, verb, or adjective). The definitions are written in simplified Korean using words more frequent than the target words whenever possible. For example, definitions in the 3K band were written using words in the 1K and 2K bands; in the 1K band extra care was taken to ensure that highly frequent words likely to be familiar to beginner learners were used.

|          | 부모                    | 역할                    | 손                     | 키                     | 대학                    |                             | parents               | role                  | hands                 | height                | university            |
|----------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| 아버지와 어머니 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | father and mother           | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| 해야 하는 일  | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | things that need to be done | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |
| 물건을 잡는 것 | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | things to grasp objects     | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> | <input type="radio"/> |

Figure 1 Example KVLТ Item.

Psychometric Qualities

The KVLТ demonstrated strong internal consistency reliability with a Cronbach’s  $\alpha$  of .95. As expected, the difficulty of the test items was affected by word frequency (Table 1). Items from the first three bands were generally easy for participants, but difficulty increased in the 4K and 5K bands. (The difficulty increased as the frequency level decreased. Item discrimination was generally good throughout the test.)

Relationships with Korean Language Skills

Participants were asked to self-assess their proficiency in listening, reading, writing, and speaking ( $n = 104$ ) as well as self-report standardized test scores from the Test of Proficiency in Korean (TOPIK). For self-assessments, KVLТ scores correlated at  $r = .56$  ( $p < .001$ ) for listening,  $r = .34$  ( $p < .001$ ) for speaking,  $r = .47$  ( $p < .001$ ) for reading, and  $r = .34$  ( $p < .001$ ) for writing. Among lower-proficiency students who took the lower level of the TOPIK exam ( $n = 22$ ), the vocabulary size showed a small to moderate correlation with TOPIK I listening ( $r = 0.19$ ;  $p = .397$ ) and reading ( $r = 0.54$ ;  $p = .009$ ) scores. Among higher proficiency students who reported TOPIK II scores ( $n = 31$ ), small to medium relationships were found for listening ( $r = 0.34$ ;  $p = .061$ ) and reading skills ( $r = 0.27$ ;  $p = .142$ ).

Delaney’s Korean Vocabulary Knowledge Test

The purpose of the KVKT is to assess the written receptive meaning-recall vocabulary knowledge of L2 Korean learners for research and low-stakes educational purposes, especially as it relates to reading ability. The data discussed in this report was collected as part of a larger, ongoing study that explores the relationship between overall reading comprehension and various subcomponents of reading, including vocabulary knowledge. Data from the KVKT was collected via Google forms and a third-party program was used to impose a 30-minute time limit. The results discussed here are from the first round of data collection (Spring 2023) which consisted of 36 participants (23 female, 9 male, 4 non-binary). Participants were recruited from undergraduate and graduate-level Korean language courses. The lowest course represented was KOR 101 (first semester, introductory course), while the highest course represented was a graduate-level course in a Korean for professional master’s program.

*Design of the Test*

The KVKT is a meaning-recall test designed specifically for investigating relationships with reading (McLean et al., 2020). Its purpose is to measure word knowledge of the 5,000 most frequent Korean words compiled in Lee, Jang, and Seo (2017). The test includes 150 words and is divided into five frequency levels, as in Park’s KVLTL. Among each band’s 30 words were 18 nouns, 7 verbs, 3 adverbs, and 2 adjectives. This ratio closely matches the ratio of word types found in Lee et al.’s (2017) frequency dictionary. In each frequency level nouns are presented first, followed by verbs, then adverbs, and then adjectives. For each item, test takers are presented with a Korean word and a short sentence that provides part-of-speech information. They are instructed to write the English definition of the word or to write “S” (for ‘skip’) if they did not know the word. A sample item is given below (Figure 2).

As in Park’s KVLTL, word frequency information was taken from Lee, et al.’s (2017) dictionary. The 30 words in each level were chosen randomly using the random number generator at random.org, balanced for word class, and final selections were made using researcher discretion.

Although the KVKT was initially based on Webb, Sasao, & Balance’s (2017) VLT, the meaning recall format was chosen over the word-meaning recognition format for a few reasons. One was to reduce incorrect responses due to unfamiliarity with words and/or grammar patterns used in the given definitions. The meaning-recall format isolates knowledge of the target word whereas a meaning-recognition format tests knowledge of the target words *and* the words and grammar of the definition and may therefore lead to an inaccurate assessment of target-word knowledge. Another advantage of the meaning recall format is that it eliminates the random guessing that often happens on multiple choice tests, which may lead to an over estimation of test-takers’ vocabulary knowledge (Schmitt et al., 2020). And finally, as Kremmel and Schmitt (2016) point out, in an actual reading task, the reader must depend on automatic meaning-recall, with little to no outside hints. In this regard, a meaning-recall test such as the KVKT, may be a better representation of real-world reading than a word-meaning recognition test.

*Psychometric Qualities*

Overall, the KVKT showed very high internal consistency ( $\alpha = 0.99$ ). Furthermore, within-band reliability ranged from 0.95–0.96, with an average of 0.95. Item discrimination values, which indicate how well item scores align with total scores minus the item, ranged from 0.10 to 0.81, with an average of 0.56 (excluding items

| Actual Item                | English Translation                        |
|----------------------------|--------------------------------------------|
| 1. 앞   학교 앞에 있어요.<br>_____ | 1. front   (It) is in front of the school. |

**Figure 2** Example KVKT Item.

**Table 1** *Average Item Facility and Discrimination by Frequency Band*

| Test                            | Band | Mean Item Facility | Mean Item Discrimination |
|---------------------------------|------|--------------------|--------------------------|
| KVLT Meaning-recognition (Park) | 1K   | 0.89               | 0.60                     |
|                                 | 2K   | 0.83               | 0.63                     |
|                                 | 3K   | 0.82               | 0.58                     |
|                                 | 4K   | 0.74               | 0.58                     |
|                                 | 5K   | 0.67               | 0.57                     |
| KVKT Meaning-recall (Delaney)   | 1K   | 0.81               | 0.49                     |
|                                 | 2K   | 0.59               | 0.58                     |
|                                 | 3K   | 0.35               | 0.63                     |
|                                 | 4K   | 0.29               | 0.58                     |
|                                 | 5K   | 0.14               | 0.50                     |

which all participants answered correctly or incorrectly). Item facility ranged from 0–1 (including values from all items), with an average of 0.43. The average item discrimination and item facility values for each band are shown in Table 1. In line with expectations, items were easier in the high-frequency bands and more difficult in the low frequency bands.

*Relationships with Korean Reading Ability*

One reason to assess vocabulary knowledge is to help identify underlying issues when it comes to reading comprehension. To that end, the KVKT was analyzed in relation to a reading comprehension test also taken by participants. The reading comprehension test used in this study combined questions and passages from Cho (2018) and the newly developed placement test for beginning and intermediate levels at the University of Hawai‘i at Mānoa. The test contained 13 reading passages and 20 multiple-choice items. Each item had four answer choices; the stem and all options were presented in English. Participants were instructed not to guess if they were unsure of the answer. Overall, the reading comprehension test showed a good level of internal consistency ( $\alpha = 0.85$ ) with an average score of 14 (range: 4 to 19). A strong, positive correlation was found between the KVKT and the reading comprehension test ( $r = 0.81$ ;  $p < 0.001$ ,  $n = 36$ ).

**Conclusion**

As researchers (all) and teachers (Park and Delaney) of Korean as a second/foreign language, we came to the realization that no one was going to build Korean vocabulary tests for us. Fortunately, a large body of research and numerous vocabulary tests developed for L2 English learners provided strong conceptual and practical foundations for developing our own tests, and a corpus-based frequency dictionary

allowed us to efficiently select and write test content. This has led, so far (and it is very important to point out the limited sample sizes for some correlations involving the KVLIT and all analyses reported for the KVKT, which is still a work-in-progress), to quite predictable results – the Korean vocabulary tests introduced here had high reliability, item difficulties that tended to increase as frequency decreased, and sizable correlations with reading skill. Such findings should be encouraging to researchers and practitioners interested in other languages with limited vocabulary assessment resources. It would seem quite feasible to develop standardized, VLT-inspired tests of Japanese vocabulary suitable for a wide range of L2 learners using Tono et al.'s (2013) *A Frequency Dictionary of Japanese*, for example, or another suitable lexical resource.

Of course, the tests introduced here are not without limitations, and each only captures one aspect of vocabulary knowledge relevant to the most common Korean words. There is certainly more important work to be done. For example, measures that target more advanced or specialized Korean vocabulary would be welcome; such tests could better serve research and education targeted at advanced L2 learners, like students in the Korean Flagship program at our university and people pursuing higher education in Korean (as a discipline or medium of instruction). Similarly, tests that target phonological word forms are needed to better understand vocabulary knowledge's role in listening comprehension and speaking ability. We also see tests of word part knowledge in Korean as particularly interesting for future work. Developing tests will be an important step in understanding how much morphological knowledge learners of Korean have and how that knowledge relates to language skills. We see this as an exciting and important research agenda that will lead to better understanding of L2 Korean vocabulary knowledge and potentially yield useful information for practice.

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Daniel R. Isbell, Department of Second Language Studies, University of Hawai'i at Mānoa, 1890 East-West Road, Moore Hall 555, Honolulu, Hawaii, 96822, USA. Email: disbell@hawaii.edu

## References

- Hashimoto, B. J. (2021). Is frequency enough? The frequency model in vocabulary size testing. *Language Assessment Quarterly*, 18(2), 171–187. <https://doi.org/10.1080/15434303.2020.1860058>
- He, X., & Godfroid, A. (2019). Choosing words to teach: A novel method for vocabulary selection and its practical application. *TESOL Quarterly*, 53(2), 348–371. <https://doi.org/10.1002/tesq.483>
- Isbell, D. R., Lee, J., & Jang, J. (2023). *Linguistic influences on comprehensibility and accentedness in second language Korean speech* [Preprint]. Open Science Framework. <https://doi.org/10.31219/osf.io/pj83v>
- Isbell, D., & Lee, K. (2023). L2 Korean speech fluency: Listener perceptions and objective measures. Paper presented at the 28th American Association of Teachers of Korean (AATK) Conference, Philadelphia, PA.

- Isbell, D. R., & Son, Y.-A. (2022). Measurement properties of a standardized elicited imitation test: An integrative data analysis. *Studies in Second Language Acquisition*, 44(3), 859–885. <https://doi.org/10.1017/S0272263121000383>
- Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, 13(4), 377–392. <https://doi.org/10.1080/15434303.2016.1237516>
- Lee, S.-H., Jang, S. B., & Seo, S. K. (2016). *A frequency dictionary of Korean: Core vocabulary for learners* (1st ed.). Routledge. <https://doi.org/10.4324/9781315561417>
- McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, 39(6), 823–845. <https://doi.org/10.1093/applin/amw050>
- 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>
- Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.
- National Institute of the Korean Language. (2003). 한국어 학습용 어휘 목록 [A Korean vocabulary list for learners]. Retrieved from [https://www.korean.go.kr/front/etcData/etcDataView.do?mn\\_id=46&etc\\_seq=71](https://www.korean.go.kr/front/etcData/etcDataView.do?mn_id=46&etc_seq=71)
- Park, L. (2022). *The Korean Vocabulary Levels Test (VLT) as a useful proficiency measure tool: The relationship between receptive vocabulary knowledge and proficiency*. Paper presented at the 27th American Association of Teachers of Korean (AATK) Conference, Ithaca, NY.
- Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, 53(1), 109–120. <https://doi.org/10.1017/S0261444819000326>
- Tono, Y., Yamazaki, M., & Maekawa, K. (2013). *A frequency dictionary of Japanese*. Routledge. <https://doi.org/10.4324/9781315823287>
- Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL – International Journal of Applied Linguistics*, 168(1), 33–69. <https://doi.org/10.1075/itl.168.1.02web>
- Vitta, J. P., Nicklin, C., & Albright, S. W. (2023). Academic word difficulty and multidimensional lexical sophistication: An English-for-academic-purposes-focused conceptual replication of Hashimoto and Egbert (2019). *The Modern Language Journal*, 107(1), 373–397. <https://doi.org/10.1111/modl.12835>