

The Internal Consistency and Accuracy of Automatically Scored Written Receptive Meaning-Recall Data: A Preliminary Study

Stuart McLean^a, Paul Raine^b, Geoffrey Pinchbeck^c, Laura Huston^d,
Young Ae Kim^e, Suzuka Nishiyama^a, and Shotaro Ueno^f
^aMomoyama Gakuin University; ^bKeio University; ^cCarleton University; ^dJosai
International University; ^eKyoto Seika University; ^fHirakata Junior High School

Abstract

Vocableveltest.org is a testing platform on which users can create on-line self-marking meaning-recall (reading or listening) and form-recall (typing) tests that address a number of limitations of the existing vocabulary level tests and vocabulary size tests. A major limitation of many existing vocabulary tests is the written receptive meaning-recognition (multiple-choice or matching) format which is associated with increased error due to guessing and decreased power to measure the type of vocabulary knowledge suitable for reading practice (McLean et al., 2020; Stewart et al., 2021a; Stoeckel et al., 2021), despite being designed for this purpose (Nation, 2012; Schmitt et al., 2020; Webb et al., 2017). Conversely, scoring meaning-recall tests by hand is labour-intensive, and the internal consistency and accuracy of automatically marked data are unknown. Thus, this study investigated the internal consistency and accuracy of automatically marked responses of 98 words from the fifth 100 most frequent words of English. This study tested for knowledge of high-frequency words as a more robust test of the marking system, as these words possess multiple-meaning senses, making their automatic marking problematic. Furthermore, the predicted limited range of learners' knowledge of these 98 words was expected to result in data of a low internal consistency. However, the automatically marked data had a high internal consistency (Cronbach's $\alpha = 0.868$) and was 98% similar to human marked meaning-recall responses.

Keywords: meaning-recall, automatic marking, accuracy

1 Background: Addressing the Limitations of Existing Vocabulary Levels and Size Tests

Many of the most commonly used vocabulary levels and size tests (hereafter *levels tests*) are based on word families, and sample between 5 and 30 meaning-recognition (multiple-choice or matching format) items to represent 1,000 words. Levels test design has seen few innovations and has been the focus of recent critical scrutiny (Kremmel, 2016; Schmitt et al., 2020; Stewart et al., 2021a; Stewart et al., 2021b; Stoeckel et al., 2021). Vocableveltest.org (McLean & Raine, 2018)

addresses some of the known limitations of levels tests and allows teachers and researchers (hereafter *teachers*) to create online self-marking levels tests to their required parameters. Once tests are created, teachers are provided with web addresses and QR codes that are shared with learners so that they can complete levels tests. Learners can be provided with feedback, and teachers can download actually typed responses, dichotomously scored responses, and the time taken to complete each response.

1.1 Item Format

When a receptive levels test is administered through the written receptive modality (as opposed to spoken), the conventional assumption has been that the test does in fact measure the examinee's knowledge of vocabulary required for reading, rather than listening (Beglar, 2010; McLean & Stoeckel, 2021; Nation, 2012). Figure 1 illustrates the four main vocabulary test item types. Both theory and research support the use of written receptive meaning-recall items to measure the type of lexical knowledge that can be employed when reading (Aviad-Levitzky et al., 2019; McLean et al., 2015a, 2020; Stoeckel et al., 2019; Zhang & Zhang, 2020).

Retrieval forms

	recognition	recall
Meaning-	It is a <u>huge</u> dog. A 巨大な A big B 幸運な B Lucky C 面白い C funny D 怠惰な D lazy	Huge =
Form-	It is a <u>巨大な</u> dog. A huge B lucky C funny D lazy	<u>巨大な</u> = ?

Figure 1. Form-Meaning Link Vocabulary Item Types.

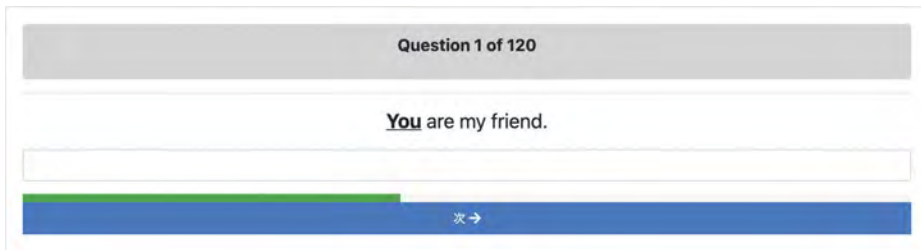


Figure 2. An Image of a Written Receptive Meaning-Recall Item (A Japanese translation of target word - underlined - is supplied by the Test-Taker).

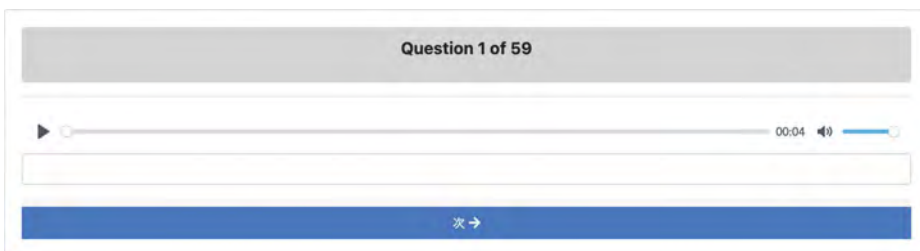
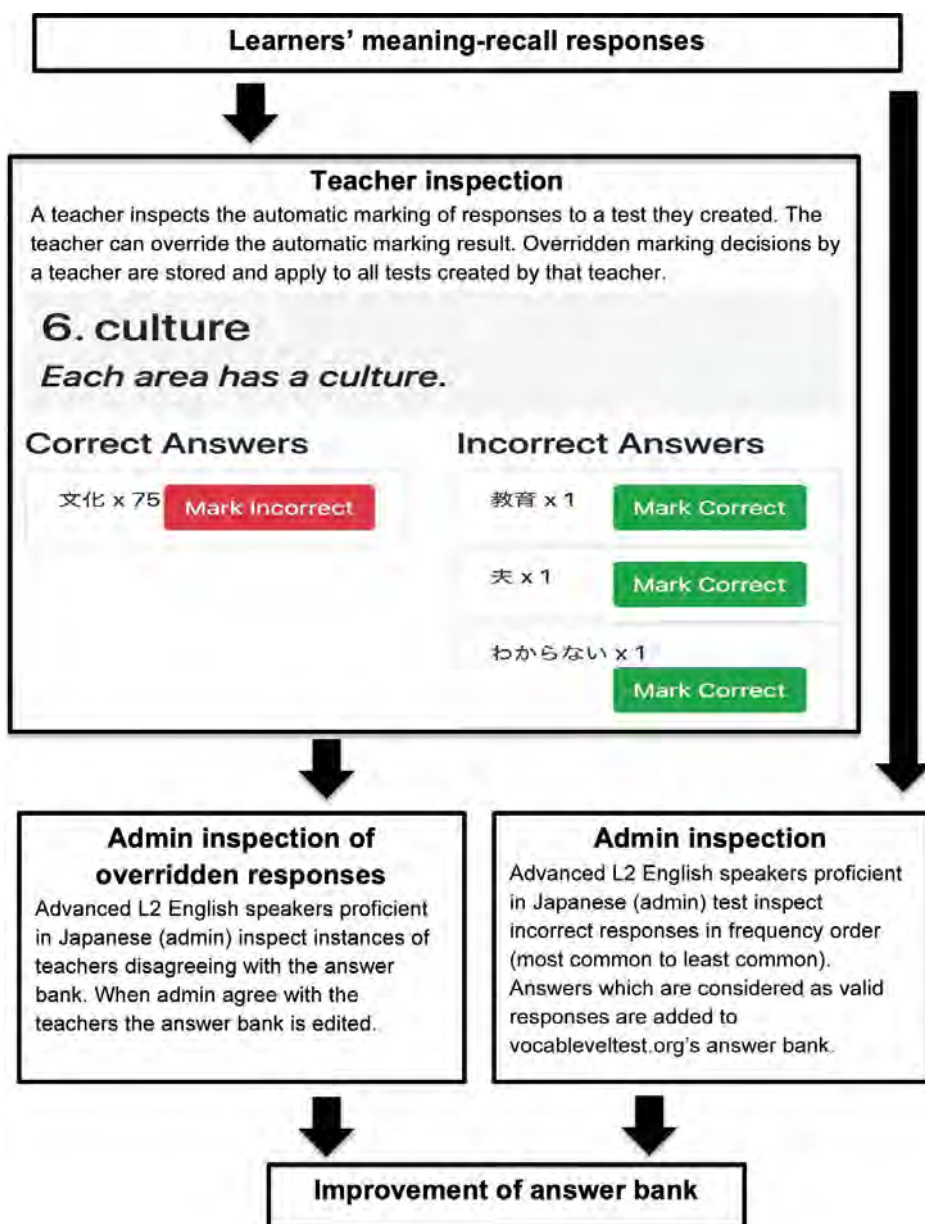


Figure 3. An Image of a Spoken Receptive Meaning-Recall Item. No target word is visible as learners hear the target word first in isolation and then in a sentence. For example, *school: It is a school.* The learner then provides an L1 translation of the target word.



Figure 4. An Image of a Written Productive Form-Recall Item. The target meaning 自転車 (bike) is translated by learners into L2 English. The correct answer is bike. If a learner gives the answer *bicycle*, they are told that it is a valid answer, but not the correct answer for this question. The learner is then given another 20 seconds in which to answer.

Vocableveltest.org allows teachers to create written (Figure 2) and spoken (Figure 3) receptive meaning-recall tests, as well as written productive form-recall tests (Figure 4). Vocableveltest.org automatically marks learners' responses using an extensive bank of possible valid responses collected from dictionaries and the inspection of learners' responses through two methods (Figure 5). In the first method, incorrect responses are inspected and valid responses are added to the answer bank. In the second method, teachers give feedback on completed tests with automatically scored responses. Teachers' suggestions for test bank changes are stored and presented to site administrators, who can supplement and edit the answer bank. The written and spoken versions of the receptive meaning-



Vocableveltest.org's Answer Banks Are Improved.

recall tests can now be completed in Japanese, Vietnamese, French, Chinese, Dutch, and Arabic.

1.2 Word Lists

Existing levels tests have been based on a limited number of word lists. When matching learners with vocabulary-level-appropriate materials (McLean,

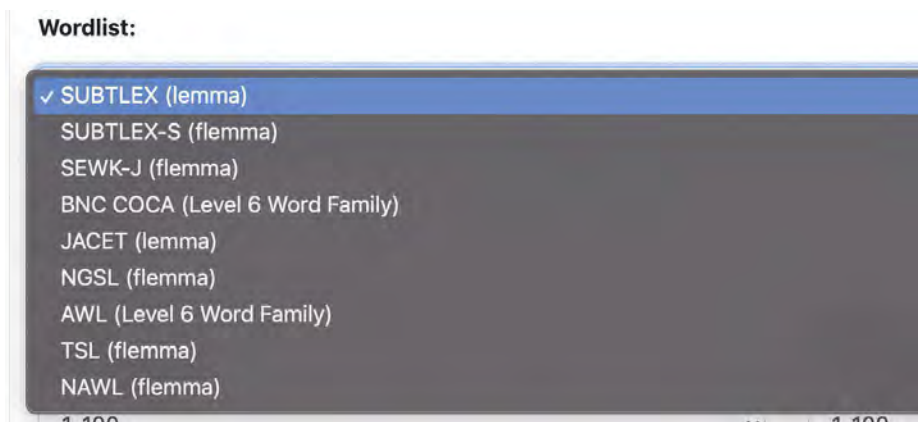


Figure 6. A Screenshot Showing Word List and Lexical Unit Selection.

2014), the use of knowledge-based word lists is preferable to a frequency list (Paul Nation, personal communication, August 8, 2021; Schmitt et al., 2021). Knowledge-based lists rank words according to how well words are known within a given population. *Vocableveltest.org* facilitates the creation of levels tests from nine different word lists (Figure 6). The Scale of English Word Knowledge—Japan (SEWK-J) is a list derived from a predictive model of English word knowledge for native Japanese speakers. Please see Mizumoto et al. (2021) for a description of the parallel text profiler that uses the same word list to estimate the difficulty of candidate-text vocabulary.

1.3 Word Counting Unit

In L2 English research, the lexical units most often discussed are as follows: (a) the “type,” any specific orthographic form (e.g., *use*); (b) the “lemma,” comprised of a base word of a particular part of speech (POS) and its inflectional forms (*use_{verb}*, *used_{verb}*, *uses_{verb}*, and *using_{verb}*); (c) the “flemma,” a base word form and inflectional forms, regardless of POS (*use_{verb}*, *used_{verb}*, *used_{adjective}*, *uses_{verb}*, *using_{verb}*, *use_{noun}*, and *uses_{noun}*); (d) and the “Word Family (WF6),” a base word form, inflectional forms, and its derivational forms regardless of POS to level 6 of Bauer and Nation’s (1993) affix criteria (*use_{verb}*, *use_{noun}*, *uses_{noun}*, *misuse_{verb}*, *misused_{verb}*, *misused_{adjective}*, *misuser_{noun}*, *misusers_{noun}*, *misuses_{verb}*, *misusing_{verb}*, *reusable_{adjective}*, *reuse_{verb}*, *reused_{adjective}*, *reused_{verb}*, *reuses_{verb}*, *reusing_{verb}*, *unusable_{adjective}*, *unused_{adjective}*, *usability_{noun}*, *usable_{adjective}*, *used_{verb}*, *used_{adjective}*, *useful_{adjective}*, *usefully_{adverb}*, *uselessness_{noun}*, *useless_{adjective}*, *uselessly_{adverb}*, *user_{noun}*, *users_{noun}*, *uses_{verb}*, and *using_{verb}*). “Flemma” and “lemma” are terms that have sometimes been used to refer to the same thing.

Views differ on the appropriateness of different lexical units with different learners for different purposes, with some supporting the use of WF6 (Laufer & Cobb, 2020; Laufer et al., 2021). Others the flemma or lemma (Brown et al., 2020, 2021; Kremmel & Schmitt, 2016; McLean, 2018, 2021; McLean & Stoeckel, 2021; Mochizuki & Aizawa, 2000; Stewart et al., 2021a; Stoeckel

et al., 2020, 2021; Ward & Chuenjundaeng, 2009). While the majority of the evidence supports the use of the flemma or lemma with some EFL and ESL learners, the WF6 is appropriate with native English speakers and in some EFL and ESL settings (e.g., Northern Europe). Thus, Vocableveltest.org allows teachers to select lists based on various lexical units (Figure 6).

1.4 Band Sizes and Number of Bands

Levels tests have traditionally been based on 1,000-word bands, a practice for which the rationale has not been explained. Kremmel (2016) and McLean (2021) argue for the adoption of 500-word bands for high-frequency words as these words provide such a great deal of coverage. Furthermore, for beginning learners with gaps in their knowledge of high-frequency words, the most frequent 1,000 words might never be mastered. In a survey of 3,427 Japanese learners, McLean et al. (2014) found that even learners from university departments with a *hensachi* of 61 and over (a high rank in Japan, based on average scores for standardised academic tests) did not demonstrate mastery of the first 1,000 words of English (Figure 7). Vocableveltest.org facilitates the creation of levels tests at 100-, 250-, 500-, and 1,000-word band sizes (Figure 8). Teachers can specify which bands they want their levels test to cover. For example, a teacher who wants to check if speed-reading materials written at the 1,000-word level are lexically appropriate only needs to test learners' knowledge up to the 1,000-word level. If the teacher believes that their learners have already mastered the first 500 words of English, it would not be necessary to test learners' knowledge of the first 500 words of English. If learners' knowledge of only a few bands is tested, more items can be deployed in the test to better represent those target bands.

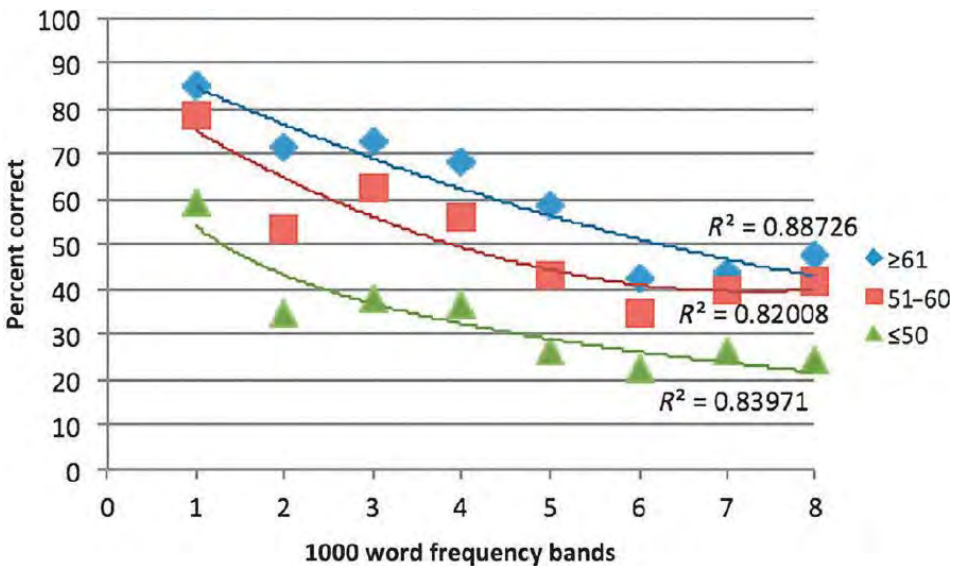


Figure 7. A Scatter Plot of VST Item Accuracy at Each 1,000-Word Frequency Level for the Three *hensachi* Groups with Best-Fit Lines.

Band Size:
500

Starting Band:
501-1000

Items Per Band:
5
[Select my own items..](#)

Question Type:
Read the English prompt and write the answer in L1
Receptive Reading (Meaning Recall)

Ending Band:
1-500
✓ 501-1000
1001-1500
1501-2000
2001-2500
2501-3000
3001-3500
3501-4000
4001-4500
4501-5000

Figure 8. A Screenshot Showing Band Size, Starting Band, and Ending Band Selection, taken from www.vocableveltest.org (McLean & Raine, 2018).

1.5 Sample Size

In practice, teachers and researchers cannot require learners to complete thousands or tens of thousands of items. Thus, levels tests present learners with samples of between 5 and 30 items which represent target word bands of 1,000 or 560 words. While it is clear that the number of items used to represent a word band affects how representative a sample is, in the case of levels tests that sample from word bands, sampling also influences the reliability, accuracy, and construct validity of the test.

Reliability is defined as the “consistency of measurement” (Bachman & Palmer, 1996, p. 19). Thus, a reliable sample from 1,000 words is a number of items that, if reselected, does not result in a significantly different estimate of a learner’s vocabulary knowledge. For example, if 30 items can represent a reliable sample, no two sets of 30 randomly sampled items (or sampled in a stratified way) will result in significantly different test scores from the same learner. Figures 9 and 10 show data from a learner who correctly answered 750 of 1,000 items representing the third 1,000-word band of English. From this data, samples of 5, 10, 20, 50, and 100 items from the third 1,000-word band were selected. The frequency (Y-axis) of the different resulting knowledge estimates (X-axis) are shown in Figures 9 and 10. The accuracy of the test is therefore a function of the sample size of words that represent the target band, and the limited representativeness, accuracy, and reliability of a sample size reduce the construct validity of both a sample size and a test. Figure 11 suggests that even samples of 100 or 200 items can occasionally result in inaccurate estimates. However, the degree of inaccuracy or value of adding more items declines significantly from 40-item tests. Vocableveltest.org allows users to select the number of items that is most appropriate for their setting (Figure 12).

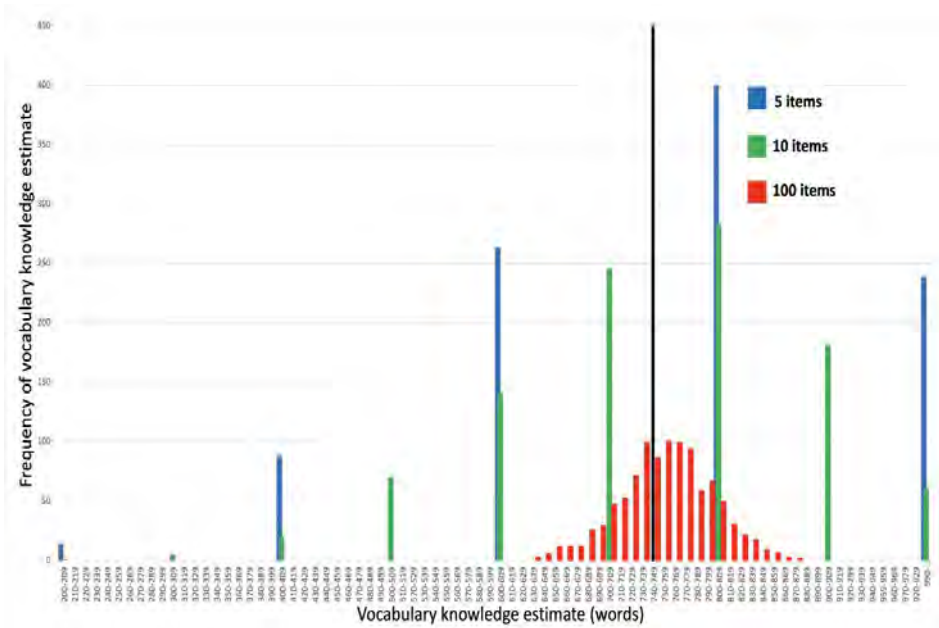


Figure 9. Monte Carlo Study of Vocabulary Size Estimates Using Tests of 5, 10, and 100 Items (Adapted from Gyllstad, McLean, & Stewart, 2021).

Note: The true number of words known by this learner is 750 (black line).

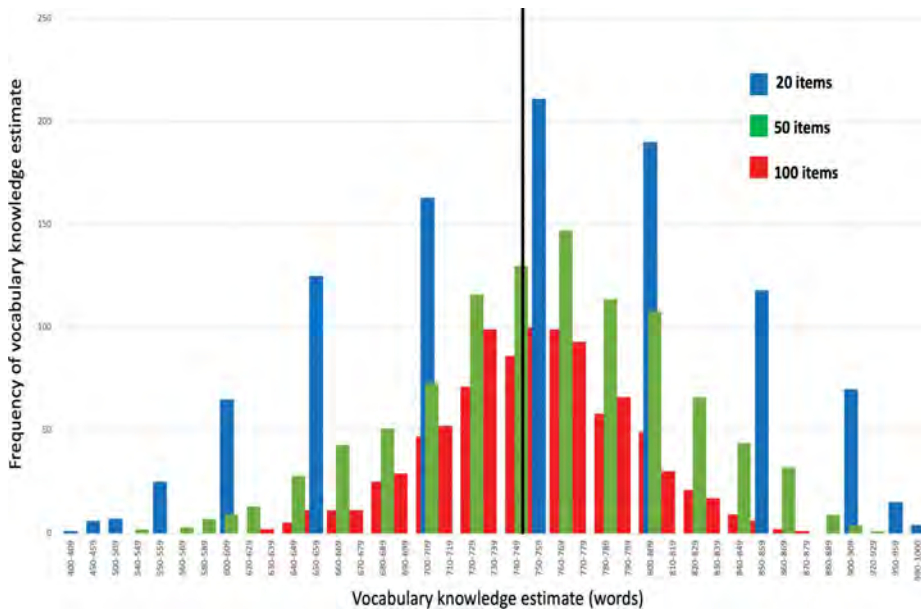


Figure 10. Monte Carlo Study of Vocabulary Size Estimates Using Tests of 20, 50, and 100 Items (Adapted from Gyllstad et al., 2021).

Note: The true number of words known by this learner is 750 (black line).

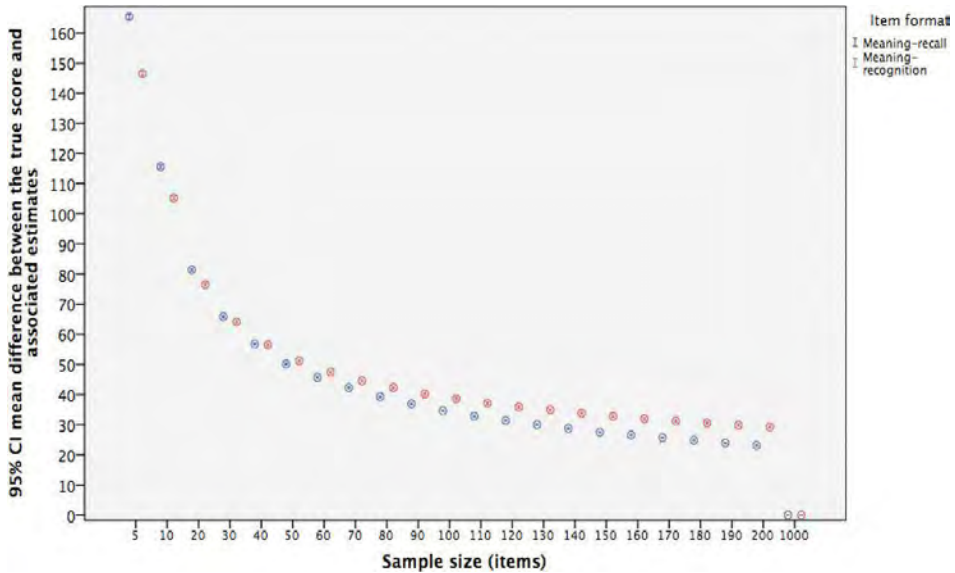


Figure 11. Mean Difference in Scores Between Learner's True Score on 1,000 Item Tests and Estimates from Bootstrapped Samples (adapted from Gyllstad et al., 2021).

Note: The data presented in this figure was from 103 participants. Thus, each data point represents the mean inaccuracy of 103,000 vocabulary knowledge estimates relative to a learner's true score.

Band Size:

Starting Band: **Ending Band:**

Items Per Band:

- 5
- 10
- 30
- 40
- 60
- 100
- 500
- 1000

Figure 12. A Screenshot Showing Sample Size Selection.

1.6 Customised Level Tests, Pretests, and Posttests

Usually, Vocableveltest.org will randomly select items within the parameters selected by the teachers. However, teachers can opt to select which items will be present in tests they create from over 7,000 items. Thus, teachers can customise level tests, pretests, and posttests.

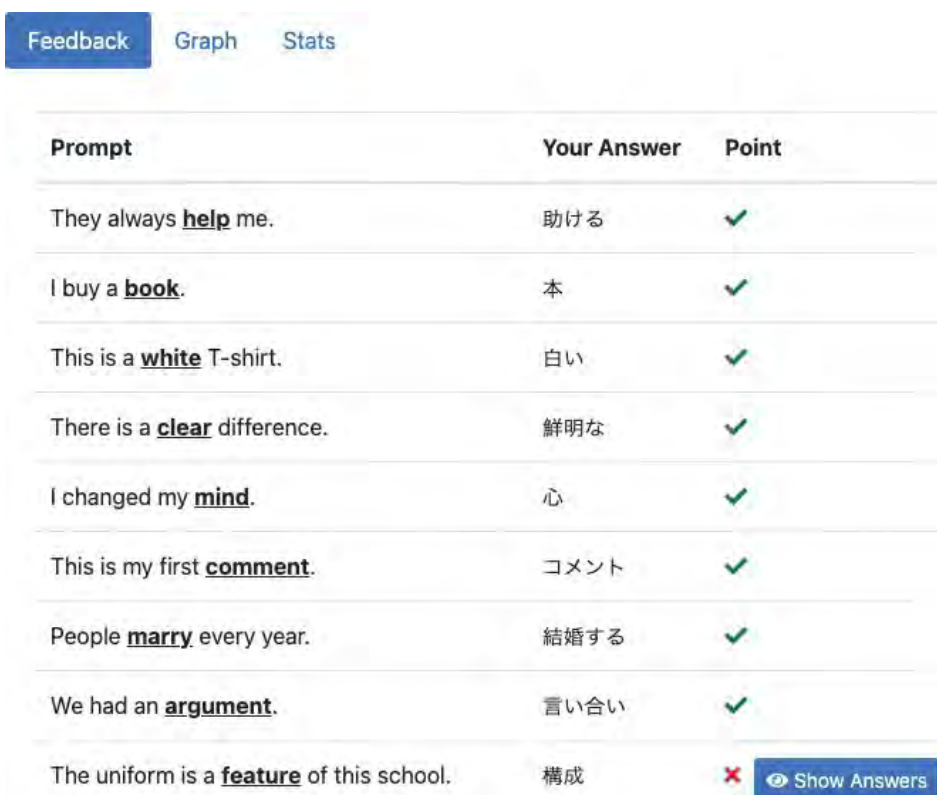
1.7 Feedback

1.7.1 Feedback for students

The existing levels tests do not automatically provide feedback to learners. Upon completion of levels tests on *Vocableveltest.org*, if teachers select the *feedback* option, students are provided with item-level feedback (Figure 13), and the percentage of correct responses at each band (Figure 14).

1.7.2 Feedback for teachers

Teachers can view the same feedback that students view, for all of their learners (Figure 13). Teachers can view mean scores for all learners who have completed a single test (Figures 15), thereby helping teachers quickly and simply estimate a class's level of lexical mastery. Teachers can download an Excel sheet of the learners' (a) typed responses, (b) dichotomously-marked responses, (c) the time taken to complete each response, and (d) class name and standardised test scores.



Prompt	Your Answer	Point
They always <u>help</u> me.	助ける	✓
I buy a <u>book</u> .	本	✓
This is a <u>white</u> T-shirt.	白い	✓
There is a <u>clear</u> difference.	鮮明な	✓
I changed my <u>mind</u> .	心	✓
This is my first <u>comment</u> .	コメント	✓
People <u>marry</u> every year.	結婚する	✓
We had an <u>argument</u> .	言い合い	✓
The uniform is a <u>feature</u> of this school.	構成	✗

[Show Answers](#)

Figure 13. An Image of Feedback for Learners.

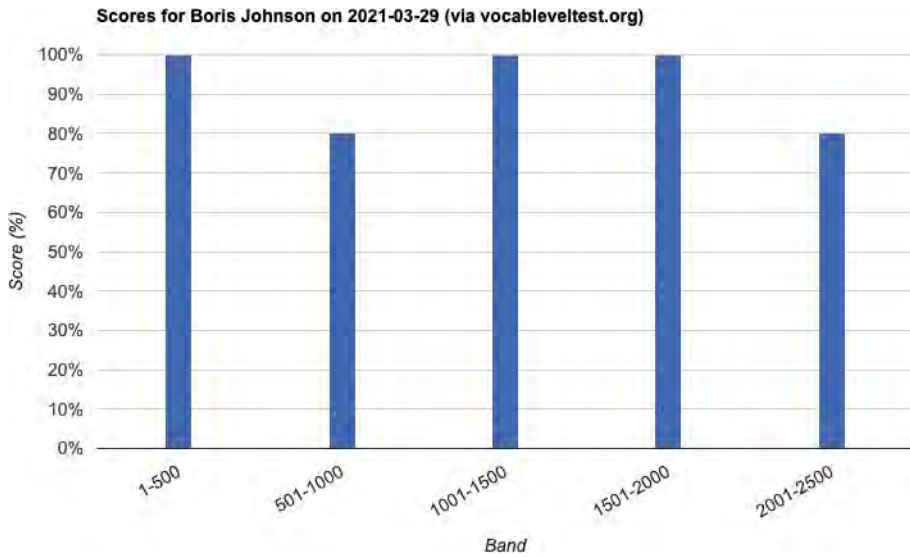


Figure 14. Student Feedback from Vocableveltest.org in the Form of a Lexical Profile.

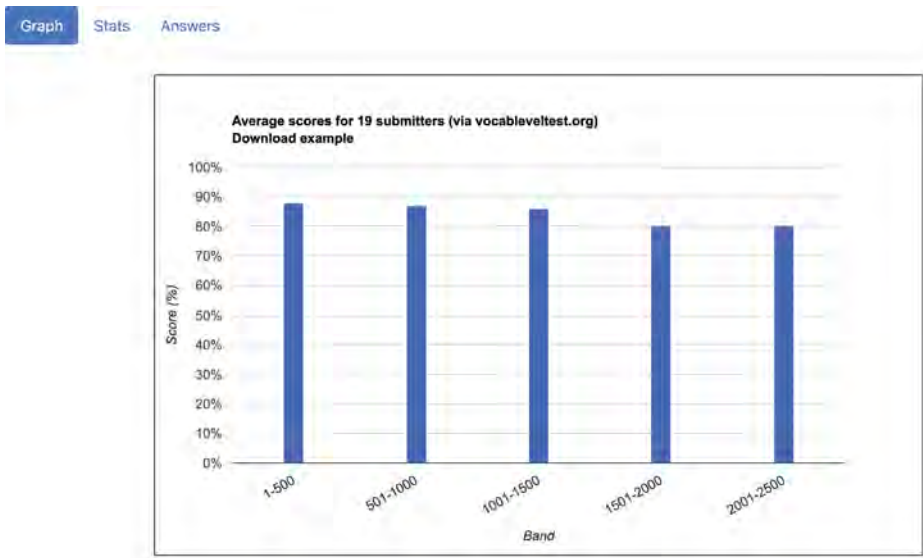


Figure 15. An Image of a Lexical Profile for a Class.

To reduce possible cheating, Vocableveltest.org has the four following features. Firstly, the text within the site cannot be automatically translated by computer or smartphone browsers. Secondly, the text within the test item stems cannot be copied. Thirdly, text cannot be pasted into answer boxes. Fourthly, the learners must complete the items within 20 seconds.

2 Methods

The present study asks the following research questions.

1. What is the internal consistency of responses that are automatically marked by Vocableveltest.org relative to human markers?
2. Can Vocableveltest.org automatically mark written receptive meaning-recall data accurately?

2.1 Participants

The participants were 78 female Japanese university students at a private university in Western Japan. The participants came from two English classes and elected this English class from several available English classes. The participants were within a range of English proficiencies. Test of English for International Communication (TOEIC®) test scores ranged between 300 and 700. All participants agreed to the use of their data in this study.

2.2 Instruments

2.2.1 Target word selection

The participants completed 98 items from the fifth 100 words of the New JACET8000 list.¹ At the time of data collection, 98 of the fifth 100 words (401–500) of New JACET8000 were words that could be tested through Vocableveltest.org (excluding *meeting* and *following*, as at the time of data collection these words did not have item stems).

The internal consistency of vocabulary test data is often the product of the number of items tested and the range of knowledge of items among the participants. Thus, a limited and ecologically-valid number of items was tested: 98 items is 60% the length of commonly used tests (McLean et al., 2015b; McLean & Kramer, 2015, 2016). Secondly, as all 98 items were from the fifth 100-word band, it was expected that the learners would be familiar with these items and would therefore be homogenous in their knowledge of items, reducing the variance in the data and its internal consistency. The high mean scores and limited standard deviations (Table 1) support these two assumptions.

We wanted to rigorously test the automatic scoring accuracy against human raters. Skipped items are unambiguously incorrect. Thus, if this item sample included a large number of low-frequency words which would likely be unknown to students and therefore skipped, it would result in an artificially high similarity of marking between the Vocableveltest.org automatic scoring and the human raters. Instead, high-frequency words, which were less likely to be skipped, were selected. Furthermore, a second reason for using high-frequency words in this study was that they are often associated with multiple possible meanings, leading to multiple valid typed L1 responses for each test item. This serves as a robust test of Vocableveltest.org's ability to accurately mark meaning-recall levels test responses.

Table 1. Descriptive Statistics for the Number of Correctly Answered Items ($K = 98$)

	<i>N</i>	Min	Max	<i>M</i>	<i>SD</i>
Marker 1	98	25	78	72.092	7.851
Marker 2	98	25	78	71.969	7.826
Automatic marking	98	25	87	71.939	8.275

2.2.2 Item presentation

The meaning-recall items were completed on Vocableleveltest.org. The website presents learners with a non-defining context sentence with the target word bolded and underlined (Figure 1). Before completing the test, test-takers read instructions and complete questions that encourage learners to consider and express the part of speech and affixes within the target forms.

2.3 Procedures

Each week the participants completed target items within each 100-word band of the NEW JACET 8000 with feedback on answers. The participants submitted a screenshot of the scores, and wrote unknown words in lexical journals which were submitted as homework and used when conducting writing tasks to encourage recycling of previously unknown words. The first week of the semester, the participants completed the target items.

2.4 Marking

The responses from the 78 participants were downloaded from Vocableleveltest.org, and the automatically marked dichotomous data was used. The participant-typed responses were presented to two native Japanese speakers, Marker 1 and Marker 2, teachers of English, who dichotomously scored the responses. The two markers were instructed to score responses that demonstrated knowledge of the target word including any affixes and any meaning-senses for the target word as correct.

3 Results and Discussion

3.1 Internal Consistency

The internal consistency of the hand-marked data by Marker 1, Marker 2, and Vocableleveltest.org was Cronbach's $\alpha = 0.863, 0.858, \text{ and } 0.869$, respectively. Under Nunnally's (1978) guidelines, an $\alpha =$ value of 0.80 is required for tests used in basic research, and a value of at least 0.90 is advisable for applied settings, although a value of 0.95 or higher is ideal. Considering the limited number of participants, the number of items, and the high degree of homogeneity of learners' knowledge of the items, the computer marking yielded reasonably high internal consistency, which is slightly higher than the human markers.

3.2. Marking Accuracy

Table 2 shows that the inter-rater reliabilities among the two markers and automatically marked data were sufficient for research purposes. Tables 2 to 5 show the degree of similarity in marking between the three marking methods. This degree of similarity provides evidence that Vocableveltest.org can mark data similar to human markers. The discrepancies between marking are due to two main causes. Firstly, the participants added particles to nouns. For example, in response to the stem, “*He is in a hospital.*”, some learners added a に after locations (e.g., 病院, hospital), or を after object nouns, which the human markers marked as correct, but Vocableveltest.org marked as incorrect. Secondly, Vocableveltest.org’s answer bank included some responses that the human markers scored as incorrect. For example, in response to the target word **best** (stem: *He is the best*

Table 2. Inter-rater Reliability (Kappa) Figures

	Marker	
	Marker 1	Vocableveltest.org
Marker 1		0.874
Marker 2	0.959	0.853

Table 3. Degree of Agreement between the First Marker and Automatic Marking

		Vocableveltest.org	
		Incorrect	Correct
Marker 1	Incorrect	518 (6.777%)	61 (0.798%)
	Correct	76 (0.994%)	6,989 (91.431%)

Table 4. Degree of Agreement between the Second Marker and Automatic Marking

		Vocableveltest.org	
		Incorrect	Correct
Marker 2	Incorrect	512 (6.698%)	79 (1.033%)
	Correct	82 (1.073%)	6,971 (91.196%)

Table 5. Degree of Agreement between the First and Second Marker

		Marker 1	
		Incorrect	Correct
Marker 2	Incorrect	563 (7.365%)	16 (0.209%)
	Correct	28 (0.366%)	7,037 (92.059%)

guitar player) the responses 良い (good) and 優れた (excellent) were scored as correct by Vocableveltest.org and incorrect by both of the human markers.

As shown in Table 5, the human raters scored more similarly to each other (98.8% agreement) than to the automatic scoring system. The automatic system was nevertheless very similar to the human raters. Of 7,644 responses automatically marked, 7,483 (97.894%) and 7,507 (98.208%) were scored in the same way as Marker 1 and Marker 2, respectively. The discrepancies between automatic marking and human raters are the result of inconsistencies between the answer bank and the human raters' decisions, which can largely be resolved by ongoing updates to the answer bank, and/or by providing marking instructions and/or calibration training to raters. It is also important to note that the advantages of recall tests over recognition tests outweigh the small differences observed between the human and automatic rating, particularly when vocabulary tests are being used as a proxy for reading proficiency. Stewart et al. (2021a) found that the Pearson's correlation between data from 30 written receptive meaning-recall items and TOEIC reading ($r = 0.74$) was significantly stronger ($p \leq 0.001$, $d = -3.622$) than that of 30 written receptive meaning-recognition items and TOEIC reading ($r = 0.65$). Thus, while the limitations of automatically marked data are salient, they seem small relative to the implicit limitations of meaning-recognition items, which offer sub-optimal construct validity. Thus, we would argue that the initial investment required to produce this platform has been worthwhile.

4 Conclusion

This article introduces Vocableveltest.org and explains how it addresses a number of limitations of the existing levels tests. In this study, student participants were tested on their knowledge of high-frequency words, and these responses were then used to evaluate the automatic scoring system, which was compared to human markers. In this preliminary study, it was found that automatically marked data was found to have high internal consistency ($\alpha = 0.869$), which was slightly higher than two human markers ($\alpha = 0.858$ and 0.863). The 7,644 automatically marked responses were found to agree 97.894% (7,483 responses) and 98.208% (responses) of the time with the two human markers. We will continue to work with teachers and other researchers to provide a user-friendly vocabulary testing platform that continues to be improved and optimised for different groups of learners.

Acknowledgments

This research was made possible by a Grant-in-aid for Scientific Research (20K00792 and 20K00898) from the Japan Society for the Promotion of Science.

Note:

1. This is the updated version of JACET8000 list (JACET, 2003), compiled by the Japan Association of College English Teachers (JACET). The New JACET8000 can be downloaded from Dr. Shin Ishikawa's website. JACET reserves the copyright to the list (JACET, 2016).

References

- Aviad-Levitzky, T., Laufer, B., & Goldstein, Z. (2019). The new computer adaptive test of size and strength (CATSS): Development and validation. *Language Assessment Quarterly*, 16(3), 345–368. <https://doi.org/10.1080/15434303.2019.1649409>
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.
- Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, 6(4), 253–279.
- Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, 27(1), 101–118. <https://doi.org/10.1177/026553229901600202>
- Brown, D., Stewart, J., Stoeckel, T., & McLean, S. (2021). The coming paradigm shift in the use of lexical units. *Studies in Second Language Acquisition*, 43(5), 950–953.
- Brown, D., Stoeckel, T., Mclean, S., & Stewart, J. (2020). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*. Advanced online publication. <https://doi.org/10.1093/applin/amaa061>
- Gyllstad, H., McLean, S., & Stewart, J. (2021). Using confidence intervals to determine adequate item sample sizes for vocabulary tests: An essential but overlooked practice. *Language Testing*, 38(4), 558–579. <https://doi.org/10.1177/0265532220979562>
- JACET Kihongo Kaitei Iinkai (JACET, Committee for Revision of the JACET Wordlist). (2003). *JACET list of 8000 basic words*. JACET.
- JACET Kihongo Kaitei Tokubetsu Iinkai (JACET, Special Committee for Revision of the JACET Wordlist). (2016). *The new JACET list of 8000 basic words*. Kirihara Shoten.
- Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, 50(4), 976–987. <https://doi.org/10.1002/tesq.329>
- 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, 377–392. <https://doi.org/10.1080/15434303.2016.1237516>
- Laufer, B., & Cobb, T. (2020). How much knowledge of derived words is needed for reading?. *Applied Linguistics*, 41(6), 971–998. <https://doi.org/10.1093/applin/amz051>
- Laufer, B., Webb, S., Kim, S. K., & Yohan, B. (2021). How well do learners know derived words in a second language? The effect of proficiency, word frequency and type of affix. *ITL-International Journal of Applied Linguistics*, 172(2), 229–258. <https://doi.org/10.1075/itl.20020.lau>
- McLean, S. (2014). Evaluation of the cognitive and affective advantages of the Foundations Reading Library series. *Journal of Extensive Reading*, 1, 1–14.

- 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. (2021). The coverage comprehension model, its importance to pedagogy and research, and threats to the validity with which it is operationalised. *Reading in a Foreign Language*, 33(1), 126–140. <https://nflrc.hawaii.edu/rfl/item/528>
- McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction*, 3(2), 4755. <https://doi.org/10.7820/vli.v03.2.mclean.et.al>
- McLean, S., & Kramer, B. (2016). The development of a Japanese bilingual version of the New Vocabulary Levels Test. *Vocabulary Education and Research Bulletin*, 5(1), 2–5. https://jaltvocab.weebly.com/uploads/3/3/4/0/3340830/verb-vol5.1_1.pdf
- McLean, S., Kramer, B., & Stewart, J. (2015a). An empirical examination of the effect of guessing on vocabulary size test scores. *Vocabulary Learning and Instruction*, 4, 26–35. <https://doi.org/10.7820/vli.v04.1.mclean.et.al>
- McLean, S., Kramer, B., & Beglar, D. (2015b). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, 19(6), 741–760. <https://doi.org/10.1177/1362168814567889>
- McLean, S., & Raine, P. (2018). VocabLeveltest.org. [Online program]. Retrieved from <https://www.vocableveltest.org/>
- 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>
- McLean, S., & Stoeckel, T. (2021). Lexical mastery thresholds and lexical units: A reply to Laufer. *Reading in a Foreign Language*, 33(2), 247–259.
- Mizumoto, M. et al. (2021). Comparisons of word lists on new word level checker. *Vocabulary Learning and Instruction*, 10(1), 1–12. <https://doi.org/10.7820/vli.v10.1.mizumoto>
- Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System*, 28(2), 291–304. [https://doi.org/10.1016/S0346-251X\(00\)00013-0](https://doi.org/10.1016/S0346-251X(00)00013-0)
- Nation, P. (2012). *The vocabulary size test*. <http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf>
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.
- Schmitt, N., Dunn, K., O'Sullivan, B., Anthony, L., & Kremmel, B. (2021). Introducing Knowledge-based Vocabulary Lists (KVL). *TESOL Journal*, n/a(n/a), e622[early view]. <https://doi.org/10.1002/tesj.622>
- 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>

- Stewart, J., McLean, S., & Batty, A. (2021a). Correlations of modalities of written vocabulary knowledge to listening and reading proficiency: A comparison. *Vocabulary Learning and Instruction*.
- Stewart, J., Stoeckel, T., McLean, S., Nation, P., & Pinchbeck, G. (2021b). What the research shows about written receptive vocabulary testing—A reply to Webb. *Studies in Second Language Acquisition*, 43(2), 462–471. <https://doi.org/10.1017/S0272263121000437>
- Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, 43(1), 181–203. <https://doi.org/10.1017/S027226312000025X>
- Stoeckel, T., Stewart, J., McLean, S., Ishii, T., Kramer, B., & Matsumoto, Y. (2019). The relationship of four variants of the Vocabulary Size Test to a criterion measure of meaning-recall vocabulary knowledge. *System*, 87, 102161. <https://doi.org/10.1016/j.system.2019.102161>
- Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System*, 37(3), 461–469. <https://doi.org/10.1016/j.system.2009.01.004>
- 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>
- Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*. Advanced online publication. <https://doi.org/10.1177/1362168820913998>