Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach

Stuart McLean

Osaka Jogakuin University

Jeffrey Stewart

Tokyo University of Science

Aaron Olaf Batty

Keio University

NOTES:

This is the accepted version of the manuscript. It contains the following substantive error:

- p. 3: Kremmel and Schmitt (2016) used four (4) tests, not seven.

The final published version also includes corrected typos and several small additions for clarification. The reference for the final, published version (Online First) is:

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

Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach

Vocabulary has long been seen as a multicomponential or multidimensional construct (e.g., Henriksen, 1999; Hunston, Francis, & Manning, 1997; Laufer & Goldstein, 2004; Read, 2000; Schmitt, 2014). Although diverse conceptualizations of the vocabulary components or dimensions abound in the literature, many researchers agree on a receptive-productive continuum, typically proposing a hierarchy of vocabulary knowledge moving from passive recognition through active recall (Henriksen, 1999; Laufer & Goldstein, 2004; Schmitt, 2014), a cline of vocabulary development that has been repeatedly demonstrated in empirical studies (e.g., Laufer & Paribakht, 1998; Laufer & Aviad-Levitzky, 2017; Laufer & Goldstein, 2004; Stewart, 2012), and for which a wide array of tests have been developed, requiring test takers to either recognize or recall (produce) words or their meanings.

Perhaps the most common justification for the assessment of L2 vocabulary knowledge is its well-established relation to reading proficiency (Gyllstad, Vilkaitė, & Schmitt, 2015; Nation, 2006; Nation & Beglar, 2007). However, some uncertainty remains over whether receptive or productive vocabulary knowledge is more predictive of reading proficiency. Many would argue that since reading itself is essentially a receptive skill, a test focused on receptive recognition of vocabulary is likely to be the most appropriate predictor of reading ability, and the majority of vocabulary tests used to predict reading proficiency are indeed receptive tests of vocabulary size or levels (Jeon & Yamashita, 2014).

However, there have been increasing calls to reexamine the relationship between vocabulary tests and reading ability. Many of the most commonly-used of these tests, such as the Vocabulary Levels Test (VLT) (Schmitt, Schmitt & Clapham, 2001) and the Vocabulary Size Test (VST) (Nation & Beglar, 2007) present the target word and ask learners to choose a definition from a list of options. It has been argued that many of the most common multiple choice tests of receptive vocabulary knowledge may inflate estimates of vocabulary size and levels (Stewart, 2014) and may not indicate an ability to reliably use the tested words in other contexts, even for reading (Kremmel & Schmitt, 2016). It has therefore been proposed that *meaning-recall* tests of vocabulary knowledge, in which meanings are written or spoken by learners rather than selected, are more effective in establishing learner knowledge of the form-meaning link (Kremmel & Schmitt, 2016), and better approximate the form of vocabulary knowledge required of readers (Gyllstad et al., 2015).

The question of what tested modality or modalities of vocabulary knowledge best predict reading proficiency scores remains an open one despite sustained interest in the field. In the present research we attempt to address methodological gaps in that literature by employing bootstrapping, a robust statistical method which simulates many replications of a study from a single dataset, in order to better elucidate the relationship between vocabulary knowledge and reading proficiency.

Background

Tests of L2 Vocabulary Knowledge

Tests of L2 vocabulary knowledge can be separated into two broad categories: recognition tests and recall tests. Each can be broken into two modalities of vocabulary knowledge assessed. Each modality may relate more, less, or differently to reading comprehension.

Recognition tests.

In their simplest form, vocabulary recognition tests consist of self-reported Yes/No tests, on which learners report which words they believe they know. To help identify response untruthfulness or false positives, researchers have proposed a wealth of corrective scoring formulas, including adding pseudowords to the Yes/No tests (e.g., Meara, 1992; Meara & Buxton, 1987). Overall, these tests are easy to create, and that makes them attractive to researchers (Schmitt, Jiang, & Grabe, 2011). A drawback of the modality is that despite the wealth of proposed scoring formulas to detect false positives, it is still unclear how effective these formulas are (Huibregtse, Admiraal, & Meara, 2002; Mochida & Harrington, 2006). It therefore remains difficult to confidently establish if learners know the words they report they know and vice versa.

Another common vocabulary test modality is meaning-recognition tests that require learners to select the meaning of a target L2 form from a list of options. Such tests can confirm learners' knowledge of words while keeping the marking of results relatively simple. A number of meaning-recognition vocabulary tests have been developed that measure learners' receptive knowledge of L2 written word forms, such as the VLT (Nation, 1990, Schmitt, Schmitt & Clapham, 2001), and, most recently and most prominent in current literature, the VST (Nation & Beglar, 2007).

Recall tests.

Two alternatives to recognition tests are meaning-recall[1] tests and form-recall tests. Meaning-recall tests provide test-takers with the L2 word form and ask them to write the L1 meaning, whereas form-recall tests provide learners with the L1 meaning and require learners to produce the L2 form. Both formats are relatively easy to create as long as test takers share an L1 known by the educator or researcher, as translations to L1 words can serve as either cues (for tests of L2 word form) or acceptable answers (for tests of L2 word meaning). However, marking such written response tests can be difficult and time-consuming, particularly if the test-takers do not share a common L1, which may contribute to the popularity of closed-response recognition tests such as the VST.

Correlations Between Vocabulary Knowledge and Reading Proficiency

Many researchers have investigated the correlation(s) between vocabulary tests and reading proficiency. Schmitt, Jiang, and Grabe (2011) reported a relatively weak Spearman correlation of .41 between a Yes/No test and a test of reading comprehension. Beglar and Hunt (1999) reported a Pearson correlation of .66 for a General Service List (GSL) (West, 1953) VLT meaning-recognition test and .72 for one of the University Word List (UWL; Xue & Nation, 1984) with TOEFL® (https://www.ets.org/toefl) Reading scores. Qian (2002) later confirmed a similar finding with a reported Pearson correlation of .74 between a VLT and the

[1] This paper uses the terms form-recall (L1 to L2 translations), meaning-recall (L2 to L1 translations), form-recognition (recognizing and selecting the L2 form from a number of possible forms, albeit from the meaning presented in the test takers' L1 or L2), and meaning-recognition (recognizing and selecting meaning, albeit from the meaning presented in the test takers L1 or L2, from a number of possible meanings), as used by Nation (2013), Nation and Webb (2011), Schmitt (2010). However, it should be noted that Laufer and Goldstein (2004) refer to form-recall as *active recall*, meaning-recall as *passive recall*, form-recognition as active recognition, and meaning-recognition as passive recognition. More recently, Laufer and Aviad-Levitzky (2017) refer to meaning-recognition as *cued-recall*.

TOEFL Reading section. Perhaps in contrast to these findings, an earlier study by Laufer (1992) found a Pearson correlation of .50 between reading ability and VLT meaning-recognition scores, but a correlation of .75 between the same measure of reading ability and a Eurocentres Vocabulary Test (Meara, 1990), which uses a test modality analogous to a Yes/No test seeded with pseudowords. However, it is difficult to draw direct comparisons between these results, as the samples, the participants' levels of proficiency, and the specific measures of reading proficiency differ from study to study, likely affecting the shapes of the associations and the ultimate strength of the various correlations.

Lack of direct comparability is a well-documented problem in the literature. Jeon and Yamashita, 2014, in their meta-analysis of factors correlated with L2 reading comprehension, had to exclude the majority of candidate papers from their analysis due to incomplete reporting or incomparable results. However, their meta-analysis of 31 correlations between L2 vocabulary knowledge and reading proficiency from 29 reading studies indicated a high average Pearson correlation of .79, CI [.69 – .86], p = .00. Although they found a non-significant difference between productive and receptive tests' correlations with reading proficiency, they noted that there were few studies employing productive tests for this purpose. Receptive tests were found to predict reading proficiency with a correlation of .74, CI [.62 – .82], while productive tests were found to be highly predictive of reading proficiency with a correlation of .92, but with a wide CI of .68 - .98. This CI is most likely owing to the small number of productive studies (eight) available to them for analysis, highlighting the need to reduce probabilities of Type II error in order to adequately investigate this topic, as we set out to do in the present research.

Correlations Between Modalities of Vocabulary Knowledge and Reading Proficiency

An important question left open is which vocabulary test modality has the strongest relationship to L2 reading proficiency. Opinions on this topic, however, are somewhat divided.

With regard to their utility as assessments of the form of vocabulary necessary for reading comprehension, detractors of meaning-recognition tests such as the VST (e.g., Gyllstad et al., 2015; Stewart, 2014) have argued that although such tests are technically receptive in nature, they do not accurately represent receptive knowledge of written word forms, because in real-world contexts learners typically are not presented with definitions from which they can choose a correct match (Kremmel & Schmitt, 2016). Indeed, a number of empirical studies have shown that a portion of correct answers on meaning-recognition tests such as the VST can be attributed to blind guessing (e.g., McLean, Kramer & Stewart, 2015). Furthermore, Kremmel and Schmitt (2016) have demonstrated that scores on vocabulary recognition or form-recall tests have relatively little match-up with meaning recall. In their study comparing seven tests of various modalities of vocabulary knowledge, test-takers could only offer a correct definition of roughly 75% of the words that a VLT, VST, or form-recall test had indicated that they knew.

However, in contrast to that view, proponents of fixed option meaning recognition tests view fixed responses as potentially advantageous when measuring vocabulary. Although some answers on meaning-recognition tests can be attributable to random guesses, they can also be an indicator of partial knowledge of these words (Nagy, Herman, & Anderson, 1985). For this reason, Nation (2012) has encouraged learners to attempt all items on the VST to ensure they receive credit for sub-conscious knowledge, and distractors in meaning-recognition options were not written with the goal of misleading learners into choosing the wrong answer.

In an attempt to determine whether vocabulary tests of meaning recognition or meaning recall were more predictive of reading proficiency, Laufer and Aviad-Levitzky (2017) directly compared two such vocabulary test modalities. The researchers administered to their participants the VST, a meaning-recognition test, and a parallel meaning-recall test of a subset of the same words, along with a reading test. Mean scores and distributions were highly similar for both tests, with a mean score of 45.72 (*SD* = 21.99) on the VST meaning-recognition test and a mean of 42.32 (*SD* = 22.04) for the recall test. Pearson correlations of the two tests with the reading measure were also extremely similar at .91 and .92, a difference that is not statistically significant (Steiger's $z = 1.36, p > .05$).

Despite the non-significant difference in results, the authors concluded that the meaning-recognition test was the superior predictor, and that the recall test "underestimated the vocabulary of almost half of the learners" in the low-intermediate ("basic") group of participants (p. 737). They further explicated the hypothesis that the options of meaning-recognition items elicit displays of partial knowledge, arguing that although learners might not be able to recall these words in isolation, their partial knowledge can allow them to infer their meanings when they are read in context due to clues about their meaning located in the surrounding text. Under this hypothesis, they interpreted that meaning-recognition test item distractors "may trigger the memory of the learner" (p. 738), allowing them to demonstrate and receive credit for partial knowledge of words that they would not receive on a meaning-recall test. They labelled this form of vocabulary knowledge *cued recall* and argued that tests that allow its measurement have greater utility for predicting reading ability than sight vocabulary alone, which is recalled without cues.

In contrast to the above findings, however, Cheng and Matthews (2018), when investigating the degree to which three modalities of vocabulary knowledge test scores predicted listening and reading proficiency, found that a form-recall productive VLT (Laufer & Nation, 1999) was more predictive of scores on a self-developed reading comprehension test than either a meaning-recognition VLT or a productive phonological test, with a Pearson correlation of 0.57 to the written productive test and 0.46 to the written receptive test. These results highlight the uncertainty that remains regarding the degree to which various modalities of vocabulary knowledge relate to reading proficiency.

Challenges Related to Comparisons of Correlations

There are numerous challenges in conducting and interpreting correlational research of this nature. The first of these is differences in testing procedures, which can change outcomes. In contrast to Nation's (2012) test specifications, which recommend that test-takers attempt to demonstrate partial knowledge by answering all questions, participants in the Laufer and Aviad-Levitzky (2017) study were instructed to skip VST items testing words that they did not believe they knew. The non-significant difference between the modalities' predictive power, therefore, may simply be a consequence of this change in instructions. It may be the case that because test-takers skipped words of which they were unsure, the two test formats became too similar for a score difference between them to be demonstrated, as there was no added benefit of guessing on the meaning-recognition test. Given the size of the correlations to the reading test and the sample size of 116, a Steiger's *z* value of 1.36 implies that the two tests were highly correlated with one another, likely to a degree that would indicate multi-collinearity ($r > .85$; Tabachnick & Fidell, 2007). That is to say that the two tests may not have been operating independently, but were instead measuring the same construct, and were effectively interchangeable from both empirical and practical standpoints.

Beyond issues of methodology, however, further challenges in conducting and interpreting correlational research remain. Even if a similar study were carried out in which the test takers attempted all items on the meaning-recognition test, differences in correlations may remain too small to be considered significant unless the sample contains several hundred students, which is, unfortunately, uncommon in the L2 acquisition/assessment field, where small sample size problems are well-documented (Larson-Hall & Herrington, 2010; Plonsky, 2013). Beyond this, even when statistical power is sufficient, correlations can vary greatly from study to study due to variability between individual observations (Glass & Hopkins, 1996; Norris, 2015), which could change between specific test forms.

Another issue is measurement error. Since internal reliability statistics such as Cronbach Alpha are essentially measures of the correlations of tests to themselves (Goodwin & Leech, 2006), if the reliability of a given test form is low, it will degrade the power with which it can be correlated to other tests (Lockhart, 1998). Test reliability can serve as an upper bound on how well a test can correlate to another measure (Glass & Hopkins, 1996). Therefore, changes in reliability due to differences in specific items used on a test may alter its precise correlation to other variables. This was demonstrated by the VLT validation studies conducted separately by Beglar and Hunt (1999) and by Schmitt et al. (2001), wherein reliability changed with the addition and subtraction of items from the test forms.

A related consideration is test length, which can exert influence on a test’s measurement error, as internal reliability varies with the number of the items on the test (Nunnally, 1978, p. 244). Most tests of receptive vocabulary knowledge have fairly low numbers of items, with as little as five items per 1,000-word level (e.g., Coxhead, Nation & Sim, 2015), ranging up to forty (e.g., Meara, 2010). In fact, the VLT validated by Beglar and Hunt (1999) was calculated to comprise 2.7% and 6.7% of the GSL and UWL lists, respectively, and despite achieving mostly good reliability, the authors advocated for longer test forms to gain better coverage and increased reliability. Producing enough items with sufficient discrimination for a long-enough test was a challenge faced by Schmitt et al. (2001), as well. An ideal test is one that is long enough to achieve acceptable reliability, but short enough to be practical. This practical consideration is important when comparing test modalities that require different lengths of time for learners to complete, as it is possible that more items of one format can be answered in a fixed period of time than another, thereby affecting test length, and thereby affecting test reliability. In the literature discussed here, test time was not controlled (Beglar & Hunt, 1999; Laufer & Aviad-Levitzky, 2017; Schmitt et al., 2001).

As a result of the factors explained above, a study suggesting that one test format has a stronger correlation than another with some criterion measure may, in fact, yield opposite results if different test forms are used, due to differences in the specific items used. Although it goes without saying, a single correlation is insufficient for drawing conclusions about the natures of the examined variables. Ideally, researchers would be able to examine a multitude of correlations drawn from a large number of tests and samples in order to determine average correlations for the predictor values, and determine to what degree the distributions of these correlations overlap for each test format. Doing so in the present case using a bootstrapping method would give a more accurate picture of which construct of vocabulary knowledge has a stronger relationship to reading proficiency, even if the differences were relatively small.

Bootstrapping Methodology

One way to address the issues explained above is the use of bootstrapping methods, which allow researchers to estimate a mean and distribution for a measure's correlation to another variable. Under bootstrapping methods, data are continually resampled with replacement (i.e., cases, once sampled, are returned to the population before sampling occurs again) from the observed data. By doing this, it is possible to observe thousands of means (or other values of interest, such as correlations) rather than just one, and see how much the values vary between what are effectively multiple replications of the same study. Doing this allows researchers to create an empirical sampling distribution for the test statistic under consideration (Larson-Hall & Herrington, 2010; LaFlair, Egbert, & Plonsky, 2015). Histograms of the resampled values can shed light on the distribution of the correlations between various tests of a given modality, showing how much these values vary between different test forms, and the degree to which correlations overlap or differ.

With regard to the correlations of modalities of vocabulary knowledge with reading proficiency, it may be the case that while one vocabulary test has a higher correlation to reading proficiency than another on average, the distributions of these correlations may be so close to one another that it is justifiable to follow the advice of Laufer and Aviad-Levitzky (2017) and use the meaning-recognition measure for research or diagnostic purposes, if for no other reason than its relative ease of marking. Conversely, if the correlational distributions differ substantially, one modality of vocabulary knowledge may prove to have a clearly stronger relationship to reading proficiency than another, and tests of that modality should be favored when wishing to predict reading proficiency.

In the present study we employ bootstrapping methodology to determine the correlation of various modalities of vocabulary knowledge to reading proficiency. One-hundred and three (103) learners answered 1,000 vocabulary test items spanning the 3rd 1,000 most frequent words in English in the New General Service List (NGSL; Browne et al., 2013). Items were answered under four modalities each: Yes/No, form-recall, meaning-recall, and meaning-recognition. These large pools of test items were then sampled with replacements to create 84,000 simulated vocabulary tests for each participant by which to investigate the distributions of correlations between the modalities and reading proficiency, as operationalized by Test of English for International Communication (TOEIC®) （https://www.ets.org/toeic）Reading Section scores.

Research Questions

The following research questions are posed:

RQ1: What are the mean reliability coefficients for tests by modality and test length?

As noted above, test reliability can affect the degree to which a test of a given construct correlates to a test of another construct. Therefore, the internal reliability of tests of various lengths will be examined and compared.

RQ2: When controlling for test length, which vocabulary test modality (Yes/No, form-recall, meaning-recall or meaning-recognition) provides the strongest correlation with L2 English reading comprehension?

Answering this question could be of benefit to researchers and educators by establishing which of these modalities, all else being equal, is most effective in predicting L2 reading ability. The results could also indicate the degrees of difference in predictive power.

Researchers can weigh the results against the relative difficulty and time required to make and score such tests when selecting an item modality for diagnostic or research purposes.

Test lengths will be examined and controlled for when addressing this question. Doing this may help educators and researchers determine at what point adding more items to a test modality results in diminishing returns to improved predictive power.

RQ3: How does the time required to complete each test version under each test length affect these correlations?

A final consideration is the time required for learners to take such tests. It has been observed by numerous researchers that Yes/No tests take less time to administer than other common test modalities (Culligan, 2015; Meara & Buxton, 1987; Mochida & Harrington, 2006). Learners therefore may be able to complete more Yes/No items or meaning-recognition items in a given interval of time than written response items of the same words. If this is the case, it is possible that Yes/No or selected-response tests given in a short period of time could serve as a better predictor of reading ability than written response tests modalities.

Method

Participants

The participants (N = 103), Japanese university students aged between 18 and 32, were of a wide range of English proficiencies, with a mean TOEIC Listening & Reading Test score of 531.75 (SD = 194.42) points. For the Reading section of the test, the mean score was 226.5 (SD = 101.87). The mean TOEIC Reading section score for test-takers in Japan is 229 (SD = 97) (Educational Testing Service, 2018), indicating that the sample was closely representative of the test-taking population in Japan. Ninety-three (96) participants (all female) were university undergraduates studying International and English Interdisciplinary Studies. Four participants (2 males and 2 females) were Japanese third-year non-English majors who had studied or lived abroad. Three participants (2 females and 1 male) were Japanese applied linguistics master's students. All participants were paid volunteers, were informed of the nature of the research, and signed informed consent forms in accordance with the guidelines of the institutions at which the participants were students.

Instruments

The participants completed four different vocabulary tests measuring their knowledge of the same 1,000 words: Yes/No, form-recall, meaning-recall, and meaning-recognition, but each test utilized a different test modality. All tests were administered via the Survey Monkey web platform (https://www.surveymonkey.com) with no time limit. Survey Monkey logs start and finish times, allowing durations to be calculated.

Target word selection.

The words for the present research were selected from the 31,242-word New General Service List corpus (Browne, Culligan, & Phillips, 2013), and were comprised of the third

1,000 most-frequent flemmas[2] in the corpus as sorted by the words' Standards Frequency Indices (SFIs).

The choice to test the third 1,000 words was a considered one. Prior research has demonstrated that the four modalities of vocabulary knowledge investigated here are of a wide range of difficulty, with tests of the same words commonly producing quite different mean scores depending on the modality (Schmitt, 2010). If more lexically advanced students were to demonstrate mastery of a 1,000-word band on any of the test modalities, or in contrast, lexically poor students demonstrated limited knowledge of the target items, this would reduce the variance within the data and inhibit meaningful analysis. As a result, it was necessary to select a 1,000-word band that would not result in the presence of a floor or ceiling effect for any learners on any item modality. Piloting of the second and third 1,000-word band with a separate sample of advanced English learners indicated that advanced learners might demonstrate mastery of the second 1,000 words of English, but not the third. Thus, the third 1,000 words were selected for use in the main study.

The four test modalities.

Participants completed items measuring knowledge of the third-1,000 words of the NGSL in four different modalities, first Yes/No, form-recall, meaning-recall and finally, meaning-recognition. A detailed description of the four item modalities and how they were marked is provided in the Supplementary File.

TOEIC Reading section subtest.

All participants completed the TOEIC following the most recent changes to TOEIC in Japan in April 2017. The TOEIC Reading section consists of 30 incomplete sentence questions, 16 text completion questions, and 54 reading comprehension questions, 29 of which are over a single passage and 25 of which require the test-taker to compare two related passages (Educational Testing Service, 2017). Descriptive statistics for the participants' TOEIC Reading scores are displayed in Table 1.

Table 1.

Descriptive statistics for participants' TOEIC Reading scores.

N	Mean	*SD*	Min.	Max.	Skew.	*SES*	Kurt.	*SEK*
103	226.51	101.87	50	440	.33	.24	-1.09	.47

Note: SES refers to "Standard Error of Skewness". SEK refers to "Standard Error of Kurtosis".

Procedures

Participants first completed the Yes/No test, then the form-recall test, then the meaning-recall test, and finally the meaning-recognition test. This progression takes advantage of the hierarchy of difficulties of the modalities of vocabulary knowledge in order to reduce the impact of cued recall based on previous tests. Difficulty of the modalities is typically understood to build from meaning-recognition, through meaning-recall, and finally to form-recall (Laufer & Goldstein, 2004). By moving down the

[2] The flemma, like the lemma, consists of a headword and its inflected, irregular, and reduced forms (e.g., -n't). Unlike the lemma, the flemma groups identical forms of different parts of speech. Thus, the verb *developed* and the adjective *developed* are different lemmas, but members of the same flemma (McLean, 2017).

theoretical hierarchy of difficulty, the danger of exposure to the words in earlier tests affecting the scores on later tests was mitigated (Laufer & Goldstein, 2004; Nation, 2013; Nation & Webb, 2011; Schmitt, 2010). Random shuffling of the item orders, in addition to the very large number of items (1,000) also contributed to making it difficult for one version of the test to inform a successive version. For a more detailed description of the process and item formats, please see the Supplementary File. A summary is available in Table 2.

Table 2.

Presentation of test modalities and their effect on participant recall.

Modality	Item Description	Information Revealed	Effect on Participant
Yes/No	• Target presented in English sentence that reveals part of speech only • Pseudowords included • Participant indicates whether known or not	• Form-meaning link unrevealed to the examinee	Unknown words remain unknown
Form-recall	• Target word presented in Japanese translation of English sentence that reveals part of speech only • Participants translate target to English	• Form-meaning link unrevealed to the examinee	Unknown words remain unknown
Meaning-recall	• Target presented in English sentence that reveals part of speech only • Participant translates target to Japanese	• Item count (1000) unlikely to be remembered	Unknown words likely remain unknown
Meaning-recog.	• Target presented in English sentence that reveals part of speech only • Participant selects Japanese meaning from four options	• Form-meaning link may be revealed to the examinee	Partially-known words may be cued

The tests were completed on computers via an online testing system in a computer room in the learners' educational settings. Participants were informed that they had to complete all four tests within two weeks between or after classes, and that they had to complete one, but not more than one, test within a single sitting. Due to the length of the tests, participants were permitted to take short breaks. They were unable to check the meaning of any of the words that they had seen on previous tests. The participants were closely monitored to ensure that they did not use dictionaries, and the Internet browser on which the participants completed the tests did not indicate the incorrect spelling of words, and did not predict the word being typed. The mean length of time necessary to complete the 1,000 Yes/No, form-recall, meaning-recall, and meaning-recognition tests was 75.49 (40.75), 123.50 (57.40), 118.01 (71.72), and 83.62 (22.45) minutes, respectively.

All participants completed the TOEIC test and vocabulary items within a three-month period.

Data Analysis

Using code in Microsoft Excel and following a bootstrapping methodology, responses for the 1,000 items for each test modality were repeatedly sampled with replacement to create thousands of simulated tests. In order to investigate the effect of test length on each modality's predictive power, 1,000 bootstrap samples were made for each test condition and test length, following Wasserman and Bockenholt's (1989) finding that this number of iterations is adequate for obtaining accurate confidence intervals for location estimates under bootstrapping methods. In order to graph trendlines for tests' predictive power by length, tests with lengths of 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200 items were then created (21 × 1,000 = 21,000 tests) for each of the four different test modalities (21,000 × 4 = 84,000 tests) for each participant (84,000 × 103 = 8,652,000). Table 3 lists the data analysis methods used to address the RQs.

Table 3.

Research questions and data analysis methods.

RQ	Method
RQ1 *(What are the mean reliability coefficients for tests by modality and test length?)*	• Comparison of mean reliability coefficients of progressively longer bootstrapped tests of modalities of vocabulary knowledge
RQ2 *(Which vocabulary test modality correlates most strongly with reading comprehension?)*	• Comparison of correlations of progressively longer (items) bootstrapped vocabulary tests to TOEIC Reading section • ANOVA of distributions of correlation coefficients of the 40- and 100-item bootstrapped vocabulary tests
RQ3 *(How does test time affect correlations to reading proficiency?)*	• Comparison of correlations of progressively longer (time) bootstrapped vocabulary tests to TOEIC Reading section

Results

Reliability

In order to answer RQ1, the internal reliability of the test forms were examined by test version and test length. Cronbach Alpha values for each test version and test length are shown in Table 4. In general, Yes/No tests had the highest internal reliability, followed by meaning-recall, form-recall and meaning-recognition tests, in descending order. Under Nunnally's rough guidelines, a Cronbach Alpha value of .80 is the minimum required for tests used in basic research, and a value of at least .90 is advisable for applied settings, although a value of .95 or higher is ideal (Lance, Butts, & Michels, 2006; Nunnally, 1978). As can be seen, Yes/No tests appeared to have the highest internal consistency, reaching an average value of 0.90 with a test of just 30 items. For tests of form and meaning, tests of 40 items had mean values over 0.90. The meaning-recognition test had the lowest internal reliability on average, attaining a mean value over 0.90 with 70 or more items. For tests with lengths of 90 items or more all modalities had reliability equal to or over .95, with the exception of meaning-recognition tests, which required a length of 130 items to reach this average value.

Table 4.

Means and standard deviations of Cronbach Alpha values by test type and test length.

	Yes/No		Form-recall		Meaning-recall		Meaning-recog.	
Items	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
5	.59	*.09*	.54	*.08*	.55	*.09*	.36	*.18*
10	.74	*.04*	.71	*.05*	.73	*.04*	.50	*.11*
20	**.85**	*.02*	**.84**	*.02*	**.84**	*.02*	.68	*.06*
30	**.90**	*.01*	.89	*.01*	.89	*.01*	.78	*.04*
40	.92	*.01*	**.91**	*.01*	**.91**	*.01*	**.83**	*.02*
50	.94	*.01*	.92	*.01*	.93	*.01*	.87	*.02*
60	**.95***	*.00*	.93	*.01*	.94	*.01*	.89	*.01*
70	.96	*.00*	.94	*.01*	**.95***	*.00*	**.91**	*.01*
80	.96	*.00*	**.95***	*.00*	.95	*.00*	.92	*.01*
90	.97	*.00*	.95	*.00*	.96	*.00*	.93	*.01*
100	.97	*.00*	.96	*.00*	.96	*.00*	.94	*.01*
110	.97	*.00*	.96	*.00*	.97	*.00*	.94	*.01*
120	.97	*.00*	.96	*.00*	.97	*.00*	.94	*.01*
130	.98	*.00*	.97	*.00*	.97	*.00*	**.95***	*.01*
140	.98	*.00*	.97	*.00*	.97	*.00*	.95	*.00*
150	.98	*.00*	.97	*.00*	.98	*.00*	.95	*.00*
160	.98	*.00*	.97	*.00*	.98	*.00*	.96	*.00*
170	.98	*.00*	.97	*.00*	.98	*.00*	.96	*.00*
180	.98	*.00*	.97	*.00*	.98	*.00*	.96	*.00*
190	.98	*.00*	.97	*.00*	.98	*.00*	.96	*.00*
200	.98	*.00*	.98	*.00*	.98	*.00*	.97	*.00*

Note: First values > .80 indicated in bold; > .90 bold and underlined; > .95 bold, underlined, and indicated with a *.

Correlations to Reading Proficiency by Test Length

In order to answer RQ2, the various test forms were correlated to learners' TOEIC Reading section scores. Figure 1 and Table 5 depict the average correlations of the four test modalities to reading proficiency (vertical axis) as a function of test length (horizontal axis). The Supplementary File shows scatterplots illustrating mean correlations for various test modalities and tests lengths. For all test lengths, meaning-recall tests had the highest average correlation to reading ability, followed by form-recall. For tests under 30 items, Yes/No tests had a slightly higher mean correlation to reading than meaning-recognition tests. However, for tests of 30 items or more, meaning-recognition tests pulled ahead. After this point Yes/No tests held the weakest correlations to reading proficiency of all modalities examined, despite boasting the highest internal reliability. All observed correlations were significant ($p < .001$). Table 5 displays a limited number of the mean Pearson's correlations of the four vocabulary formats to reading proficiency for various item length calculated, the Supplementary File displays all 84 of them.

Table 5.

Mean Pearson's correlations of the four vocabulary modalities to reading proficiency by numbers of items.

	Items					
Modality	20	30	40	50	100	200
Yes/No	.62 (.05)	.63 (.03)	.64 (.03)	.65 (.02)	.66 (.02)	.67 (.01)
Form-recall	.68 (.04)	.71 (.03)	.72 (.03)	.73 (.02)	.74 (.02)	.75 (.01)
Meaning-recall	.72 (.03)	.74 (.02)	.76 (.02)	.76 (.02)	.78 (.01)	.78 (.01)
Meaning-recognition	.62 (.04)	.65 (.03)	.67 (.03)	.68 (.03)	.70 (.02)	.71 (.01)

Note: Standard deviations for mean values are indicated in parentheses. All *p* values were < 001.

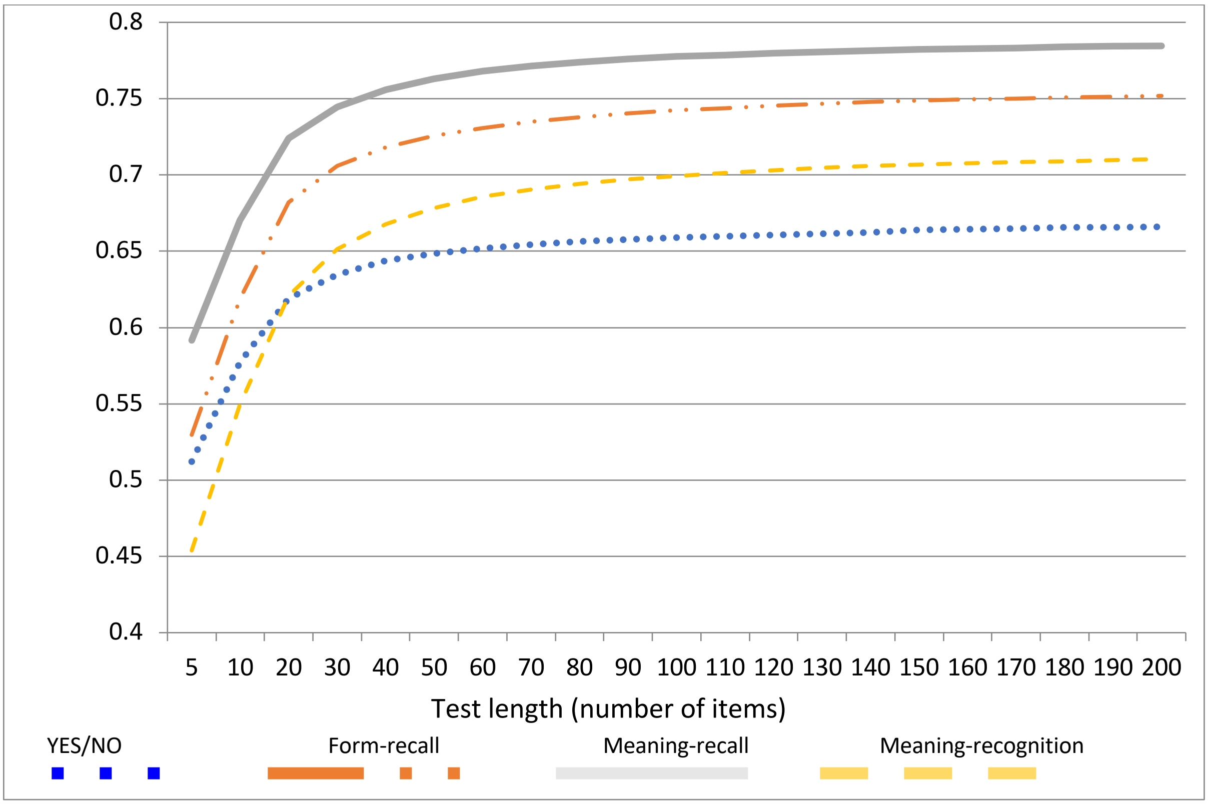

Figure 1. Pearson's correlations of vocabulary tests to reading proficiency by test modality and test lengths ranging from 5 to 200 items.

Differences in distributions of correlations between reading proficiency and modality.

ANOVAs were conducted for 40- and 100-item versions of the tests in order to establish the significance of the difference in correlations. Both were statistically significant [*F*(3, 3996) = 4074, *p* < .001; *F*(3, 3996) = 11164, *p* < .001)] and Tukey post hoc tests indicated differences were significant between all modalities for both test lengths (*p* < .001). Effect sizes and mean differences can be seen below in Tables 6 and 7. Using Plonsky and Oswald's (2014) empirically-based cutoffs for L2 research using within-subject designs, a Cohen's *d* effect size of 1.00 can be considered "moderate" and 1.40 or higher can be considered "large." As such, the listed effect sizes are almost uniformly "large." The sole exception, an effect size of

0.901 for the difference between meaning recognition and Yes/No modalities at a length of 40 items, rises to an effect size of 2.434 when test lengths are increased to 100 items.

Table 6.

Post hoc tests and effect sizes for an ANOVA on bootstrapped modality correlations to reading, 40 items per test.

		Mean Diff.	*SE*	*t*	Cohen's *d*	*p* Tukey
Form Recall	Recognition	.050	.001	45.184	1.877	< .001
	Meaning Recall	-.038	.001	-33.725	-1.639	< .001
	YN	.075	.001	66.813	2.915	< .001
Recognition	Meaning Recall	-.088	.001	-78.91	-3.622	< .001
	YN	.024	.001	21.629	.901	< .001
Meaning Recall	YN	.112	.001	100.538	4.901	< .001

Table 7.

Post hoc tests and effect sizes for an ANOVA on bootstrapped modality correlations to reading, 100 items per test.

		Mean Diff.	*SE*	*t*	Cohen's *d*	*p* Tukey
Form Recall	Recognition	.043	.001	62.269	2.55	< .001
	Meaning Recall	-.035	.001	-51.08	-2.495	< .001
	YN	.083	.001	120.928	5.209	< .001
Recognition	Meaning Recall	-.078	.001	-113.35	-5.279	< .001
	YN	.04	.001	58.659	2.434	< .001
Meaning Recall	YN	.119	.001	172.008	8.562	< .001

An advantage of bootstrapping methods is the creation of sample distributions, which can illustrate ranges of probable values in addition to an average value. These can be seen in the overlapping histograms in Figures 2 and 3, which illustrate the number of tests (vertical axes) with correlations to reading within given ranges (horizontal axes). For tests of 40 items, the distributions of Yes/No test and meaning-recognition test's reading correlations show a considerable degree of overlap, despite a statistically significant difference in their average values. It is likely for this reason that previous studies have shown conflicting results; due to differences in specific test forms used in studies, it is quite possible for one test modality to outperform another in some situations, but not in others. In contrast, for tests of 100 items the distributions of correlations for each modality of vocabulary knowledge are quite pronounced with considerably less overlap between modalities.

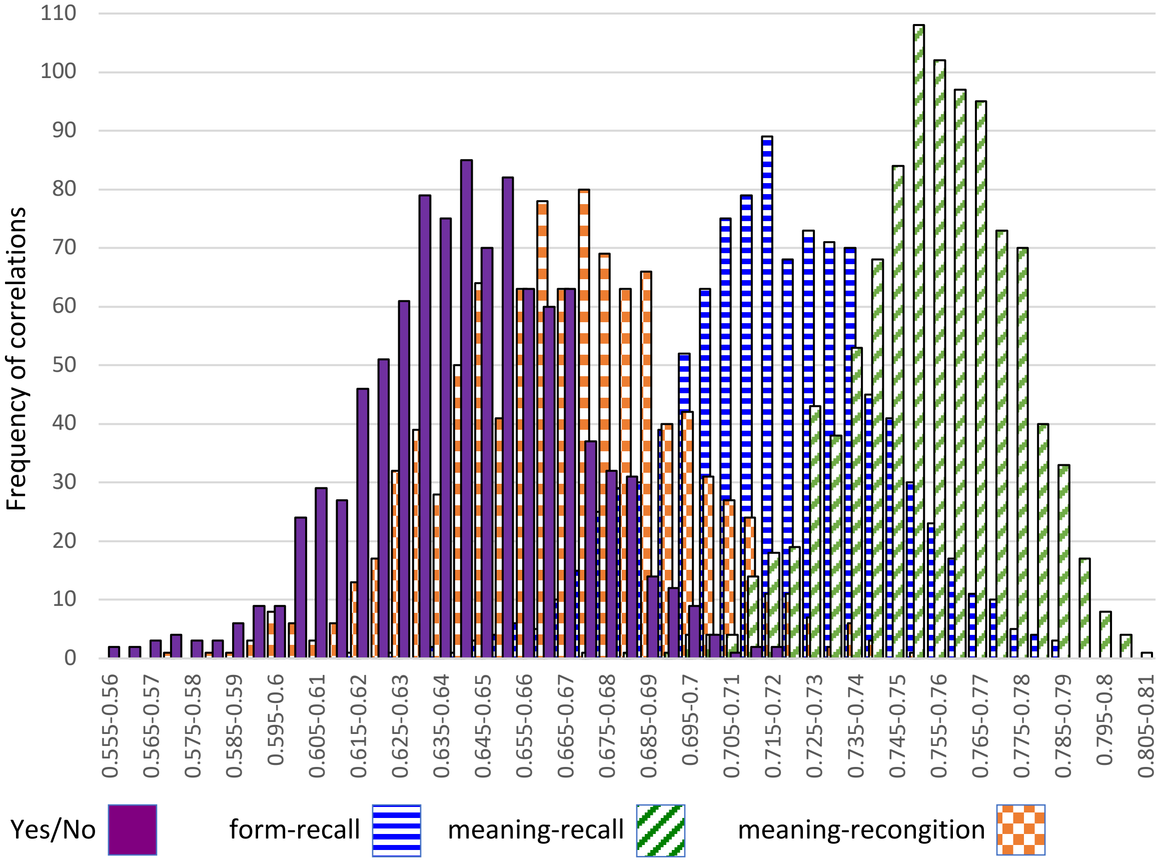

Figure 2. Distributions of Pearson's correlations of vocabulary tests to reading proficiency, grouped by test modality (*K* = 40). The *Y* axis indicates the number of correlations of each range in correlation strength (*X* axis) between a 40-item test and a single participant's TOEIC Reading score.

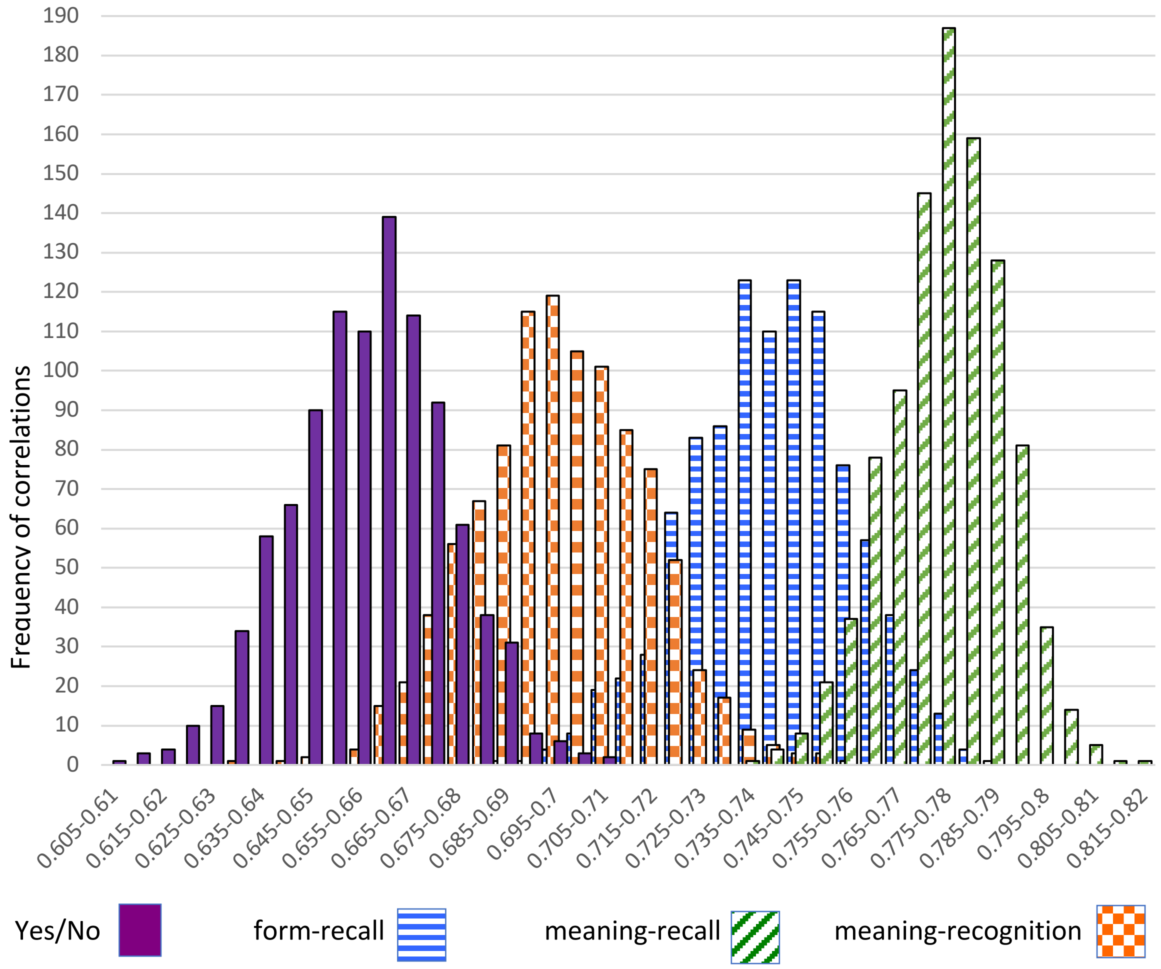

Figure 3. Pearson's correlations of vocabulary test modalities to reading proficiency ($K = 100$). The Y axis indicates the number of correlations of each range in correlation strength (X axis) between a 100-item test and a single participant's TOEIC Reading score.

Relationship between test time and correlation to reading.

Finally, to answer RQ3, the relationship between the time required to take tests in each modality and those tests' correlations to reading proficiency was examined. The time required to take such tests is an important consideration for learners, researchers and educators. Even if tests with Yes/No and meaning-recognition modalities have lower correlations to reading proficiency than meaning-recall tests of the same length, since learners can complete Yes/No or meaning-recognition tests at a faster rate, they may be able to take longer tests within a given time period, which could potentially yield higher correlations to reading than tests using more time-intensive item modalities.

Table 8 shows the mean number of test items that the 103 participants can complete in various time periods under each test modality. Using this information, correlations to reading proficiency were examined again using time as a variable rather than item counts (Figure 4 and Table 9).

Table 8.

The mean number of items completed by participants within various time periods.

	Yes/No	Form-recall	Meaning-recall	Meaning-recog.
Minutes	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)
5	104.68 (41.92)	51.21 (29.00)	49.94 (18.72)	64.17 (17.59)
10	209.36 (83.84)	102.41 (58.00)	99.89 (37.43)	128.34 (35.17)
20	418.72 (167.68)	204.83 (116.00)	199.78 (74.87)	256.68 (70.34)
30	628.08 (241.52)	307.24 (174.00)	299.67 (112.30)	385.02 (105.51)
40	837.44 (335.36)	409.66 (232.00)	399.56 (149.74)	513.36 (140.68)
50	1045.80 (419.20)	512.07 (290.00)	499.45 (187.17)	641.70 (17585)
60	1256.16 (503.04)	614.48 (348.00)	599.34 (224.61)	770.04 (211.02)

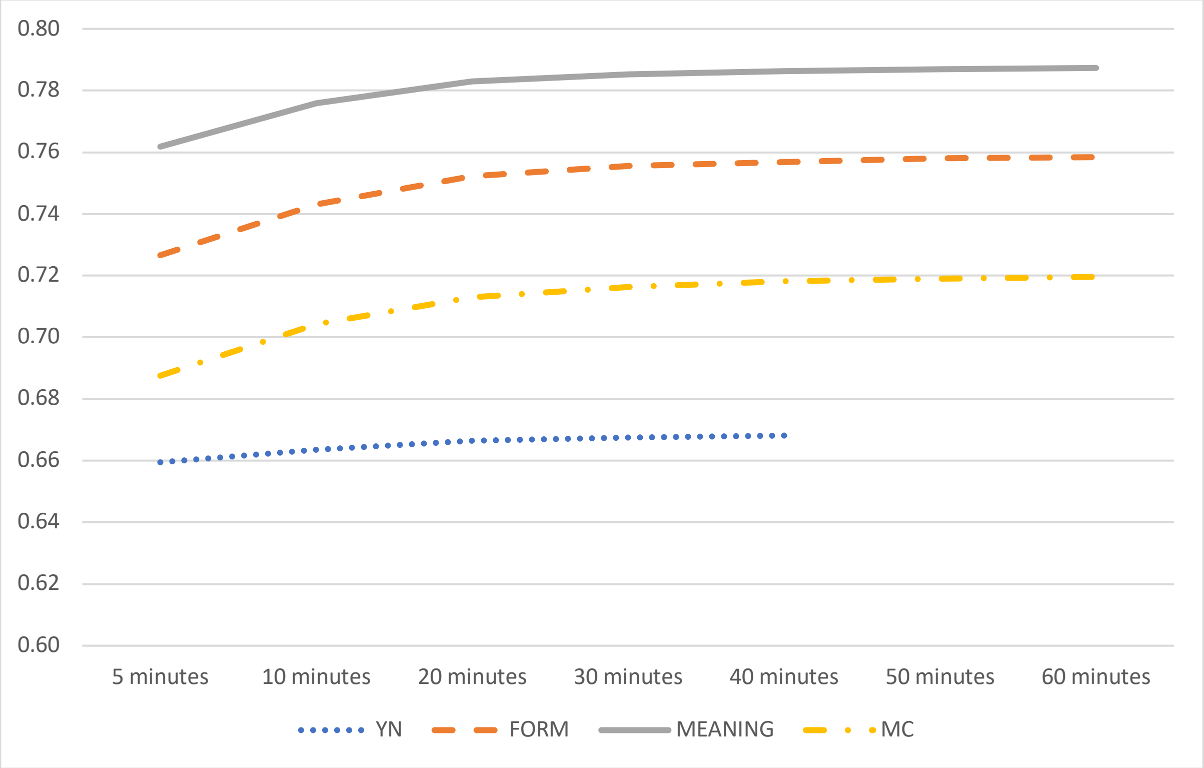

Figure 4. The mean correlation between vocabulary test scores and reading proficiency by test time.

Table 9.

The mean correlation between vocabulary test scores and reading proficiency by test time.

	Yes/No	Form-recall	Meaning-recall	Meaning-recog.
Minutes	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)
5	0.66 (0.02)	0.73 (0.02)	0.76 (0.03)	0.69 (0.02)
10	0.66 (0.01)	0.74 (0.02)	0.78 (0.02)	0.70 (0.02)
20	0.67 (0.01)	0.75 (0.01)	0.78 (0.02)	0.71 (0.01)
30	0.67 (0.004)	0.76 (0.01)	0.79 (0.02)	0.72 (0.01)
40	0.67 (0.002)	0.76 (0.01)	0.79 (0.02)	0.72 (0.01)
50		0.76 (0.01)	0.79 (0.02)	0.72 (0.01)
60		0.76 (0.02)	0.79 (0.02)	0.72 (0.003)

Note: The length (in items) of the vocabulary tests was determined by the mean number of items completed by participants in the time periods indicated. The underlined numbers indicate when correlations between vocabulary test scores and reading proficiency correlations plateau. As the average number of Yes/No items that can be completed in 50 and 60 minutes was calculated to be over 1000, it was not possible to calculate the correlation between TOEIC Reading scores in these instances.

Although Yes/No items required the least time to complete, savings in time did not appear to positively affect correlations to reading; correlations to Yes/No modality tests consistently lagged the other modalities examined and reached a near peak of approximately .67 after 20 minutes, with only marginal increases after this point. In contrast, at the 20-minute mark correlations to meaning-recognition, form-recall and meaning-recall were higher at .71, .75 and .78 respectively. The correlations of these three modalities began to peak at 30 minutes, with correlations of .72, .76 and .79, respectively.

Findings and Discussion

In summary, bootstrapped samples showed that while tests with fewer items exhibited greater overlap with one another in their correlations to reading proficiency, of the modalities examined, tests of meaning recall were on average most strongly associated with reading proficiency, followed by form recall, meaning recognition and Yes/No tests, respectively. On average, meaning recall tests of 40 items had correlations to reading proficiency approximately .09 higher than comparable meaning recognition tests and approximately .12 higher than comparable Yes/No tests, with large effect sizes of 3.62 and 4.90, respectively. Although mean differences remained comparable, for 100 item tests, effect sizes for these differences rose to 5.28 and 8.56 respectively. Given the very large effect sizes, the differences in correlations do not appear to be negligible, as few recognition tests could approach the predictive power or associative strength of otherwise equivalent meaning recall tests. Accounting for the time required to take each test did not alter the rank order of predictive power. For meaning-recognition, form-recall and meaning-recall tests, peaks in correlation occurred after approximately 30 minutes, suggesting that testing learners' vocabulary knowledge for longer lengths of time may provide diminishing returns if the tests are used as an indicator of reading proficiency.

These differences in predictive power or associative strength, which are of statistical and practical significance, appear to be in contradiction to Laufer and Aviad-Levitsky's (2017) hypothesis that tests of meaning recognition are better predictors of reading proficiency than tests of meaning recall, and confirm the findings of Cheng and Matthews (2018) discussed previously. We have two theories as to why this is the case.

The first theory is rather banal and technical in nature. Vocabulary recall measures tend to have higher reliability than receptive measures that use multiple-choice and fixed option formats, especially if test instructions permit learners to guess the answers to items they do not know (Stewart, 2014). Higher reliability implies less error of measurement, which could in turn lead to higher correlations to related psychological constructs, theoretical similarities or dissimilarities between the constructs notwithstanding. We suspect that high reliability of the recall tests relative to that of the meaning recognition test contributed to their comparatively higher correlations to reading proficiency. However, it should be noted that despite having internal reliability equivalent to or higher than the recall tests, Yes/No tests had the lowest correlation with reading proficiency. Therefore, it is likely that differences in the tested modalities of vocabulary knowledge also played a role in the observed differences.

We therefore turn to a more substantive theory regarding the nature of L2 vocabulary knowledge that was originally proposed by Laufer and associates, regarding vocabulary strength (Laufer & Goldstein, 2004). In this view, productive vocabulary knowledge, as measured by meaning-recall and form-recall tests, is a “stronger” form of vocabulary knowledge which takes learners longer to develop than receptive knowledge; scores on productive vocabulary tests typically lag those of receptive vocabulary tests. Furthermore, “stronger” forms of vocabulary knowledge suggest greater mastery of the words in question, which in turn implies possession of “weaker” forms of knowledge of those words as well. As Laufer and Goldstein (2004) argued, “Language learners who can recall the meaning of a given word can [also] typically recognize the meaning among several options” (p. 408). If this were the case, it is possible that even if productive knowledge did not appear to have as direct a significance to reading ability as receptive knowledge, since recall mastery of words also implies recognition mastery, the recall tests may function more effectively and efficiently as predictors of reading proficiency.

Conclusions

From a technical standpoint, meaning-recall test items appear to possess an ideal balance of high internal reliability and correlation to reading proficiency, even at lower item- and time-lengths (see Tables 4 and 5, and Figure 4). Although Yes/No tests reached higher reliability at shorter item-lengths, they correlated rather poorly with reading comprehension and are therefore not recommended for such a purpose. Form-recall tests required slightly more items than meaning-recall tests to reach high reliabilities, and meaning-recognition tests required more still. At all item- and time-lengths, form-recall and meaning-recognition tests demonstrated lower correlations to reading proficiency than meaning recall, with meaning-recognition consistently being the weaker of the two. Finally, all four test formats’ correlations to reading proficiency peaked within 30 minutes, suggesting it is unnecessary to test these modalities for longer than this for the purpose of predicting reading ability.

Although in the present study we demonstrated the facility of meaning-recall tests to indicate reading proficiency, this test format has not been popular in the past due to the time required to mark them relative to Yes/No and meaning-recognition modality tests. However, the recent prevalence of smart phones and free online survey software has greatly simplified the collection and scoring of written responses. In particular, the Vocableveltest.org software (McLean, 2018) is highly useful for scoring written responses to vocabulary tests. In addition to allowing test administrators to download responses for hand marking, answers that have not been encountered before can be flagged for screening by human raters and either whitelisted or blacklisted once marked. With successive administrations of the test, fewer and fewer responses require manual scoring due to this expanding bank of answers.

Limitations and Future Directions

These findings must be interpreted with qualifications in mind. First, it should be noted that 30% of the items on the TOEIC Reading section use an "incomplete sentences" item format, which, in addition to grammar, collocations, and use of function words in context, also tests contextualized vocabulary use some instances. Although ETS does not publicize a detailed test specification, it has been estimated that 10-12 items in this section, and therefore 10-12% of the Reading section, can be considered to test vocabulary explicitly (Hilke, Aizawa, & Maeda, 2018). It is possible that these items could have inflated correlations to the vocabulary tests somewhat. Additionally, meaning-recognition items used in this study were bilingual rather than the monolingual English modality of the original VST.

The predictive power of the modalities of vocabulary knowledge was assessed using linear Pearson correlations. However, as can be seen in the scatterplots in the Supplementary File, a tendency for some learners with low TOEIC Reading scores to nevertheless attain high scores on the multiple-choice modality of meaning-recognition led to a curve in the data, perhaps due to guessing effects (Stewart, 2014). Although the meaning recognition modality still had poorer correlations to reading proficiency than recall modalities, meaning the ultimate findings of this study remained unchanged, the use of a second order polynomial fit improved the modality's predictive power somewhat. This finding suggests that prediction formulas for multiple choice vocabulary tests may benefit from the use of non-linear models.

Future studies should explore these associations with other measures of reading proficiency, and monolingual tests to investigate the generalizability of the findings of the present study. Finally, as some researchers in the field of SLA (e.g., Larson-Hall & Herrington, 2010; Plonsky, 2013) have been advocating in recent years, researchers in the field of L2 assessment should make wider use of robust statistical methods such as bootstrapping in order to make better and more substantial inferences from their data.

References

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, *16*(2), 131-162. https://doi.org/10.1191/026553299666419728

Browne, C., Culligan, B., & Phillips, J. (2013). The new general service list. Retrieved from www.newgeneralservicelist.org.

Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, *35*(1), 3-25. https://doi.org/10.1177/0265532216676851

Coxhead, A., Nation, P., & Sim, D. (2015). Measuring the vocabulary size of native speakers of English in New Zealand secondary schools. *New Zealand Journal of Educational Studies, 50*(1), 121-135. https://doi.org/10.1007/s40841-015-0002-3

Culligan, B. (2015). A comparison of three test formats to assess word difficulty. *Language Testing*, *32*(4), 503-520. https://doi.org/10.1177/0265532215572268.

Educational Testing Service. (2017). *Examinee handbook for the updated version of the TOEIC® Listening & Reading Test*. Retrieved from https://www.ets.org/s/toeic/pdf/examinee-handbook-for-toeic-listening-reading-test-updated.pdf

Educational Testing Service. (2018). *2018 report on test takers worldwide.* Retrieved from https://www.ets.org/s/toeic/pdf/2018-report-on-test-takers-worldwide.pdf

Glass, G. V., & Hopkins, K. D. (1996). Statistical methods in education and psychology. *Psyccritiques*, *41*(12), 1224.

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of *r*. *The Journal of Experimental Education*, *74*(3), 249-266. https://doi.org/10.3200/JEXE.74.3.249-266

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL-International Journal of Applied Linguistics*, *166*(2), 278-306. https://doi.org/10.1075/itl.166.2.04gyl

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303-317. https://doi.org/10.1017/S0272263199002089

Hilke, R., Aizawa, T., & Maeda, H. (2018). *TOEIC L&R tesuto chokuzen no gijutsu* [Methods for shortly before the TOEIC L&R test]. Tokyo, Japan: Alc Press Inc.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, *19*(3), 227-245. https://doi.org/10.1191/0265532202lt229oa

Hunston, S., Francis, G., & Manning, E. (1997). Grammar and vocabulary: Showing the connections. *ELT Journal*, *51*(3), 208-216. https://doi.org/10.1093/elt/51.3.208

Jeon, E. H., & Yamashita, J. (2014). L2 Reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160-212. https://doi.org/10.1111/lang.12034

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

LaFlair, G. T., Egbert, J., & Plonsky, L. (2015). A practical guide to bootstrapping descriptive statistics, correlations, *t* tests, and ANOVAs. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 46-77). NY: Routledge.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*(2), 202-220. https://doi.org/10.1177/1094428105284919

Larson-Hall, J., & Herrington, R. (2009). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, *31*(3), 368-390. http://dx.doi.org/10.1093/applin/amp038

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 126-132). https://doi.org/10.1007/978-1-349-12396-4_12

Laufer, B., & Aviad-Levitzky, T. (2017). What type of vocabulary knowledge predicts reading comprehension: Word meaning recall or word meaning recognition? *The Modern Language Journal*, *101*(4), 729-741. https://doi.org/10.1111/modl.12431

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399-436. https://doi.org/10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*(1), 33-51. https://doi.org/10.1177/026553229901600103

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, *48*(3), 365-391. https://doi.org/10.1111/0023-8333.00046

Lockhart, R. S. (1998). *Introduction to statistics and data analysis: For the behavioral sciences*. Macmillan.

McLean, S. (2017). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*(6), 823-845.

McLean, S. (2018). VocabLeveltest.org. [Online program]. Available from https://www.vocableveltest.org/

McLean, S., Kramer, B., & Stewart, J. (2015). An empirical examination of the effect of guessing on vocabulary size test scores. *Vocabulary Learning and Instruction*, *4*(1), 26-35.

Meara, P. (1990). Some notes on the Eurocentres Vocabulary Tests. In J. Tommola (Ed.), *Foreign language comprehension and production* (pp. 103-113). Turku: AFinLA Yearbook.

Meara, P. (1992). *EFL vocabulary tests*. New York: ERIC Clearinghouse.

Meara, P. (2010). *EFL vocabulary tests.* Retrieved from http://www.lognostics.co.uk/vlibrary/meara1992z.pdf

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142-154. https://doi.org/10.1177/026553228700400202

Mochida, K., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*(1), 73-98. https://doi.org/10.1191/0265532206lt321oa

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, *20*(2), 233. https://doi.org/10.2307/747758

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. NY: Heinle & Heinle Publishers.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, *63*(1), 59–82.

Nation, I. S. P. (2012). *The Vocabulary Size Test: Information and specifications*. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

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

Nation, I. S. P., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Retrieved from http://teslcanadajournal.ca/index.php/tesl/article/view/1135/954

Norris, J. M. (2015). Statistical significance testing in second language research: Basic problems and suggestions for reform. *Language Learning*, *65*(S1), 97-126. https://doi.org/10.1111/lang.12114

Nunnally, J. C. (1978). *Psychometric theory* (2nd Ed.). Hillsdale, NJ: McGraw-Hill.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*(4), 655-687. https://doi.org/10.1017/S0272263113000399

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878-912. https://doi.org/10.1111/lang.12079

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513-536. https://doi.org/10.1111/1467-9922.00193

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

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

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., Jiang, X., & Grabe, W. (2011). The Percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26-43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88.

Stewart, J. (2012). A multiple-choice test of active vocabulary knowledge. *Vocabulary Learning and Instruction*, *1*(1), 53-59.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST?. *Language Assessment Quarterly*, *11*(3), 271-282.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5). Boston, MA: Pearson.

Wasserman, S., & Bockenholt, U. (1989). Bootstrapping: applications to psychophysiology. *Psychophysiology*, *26*(2), 208-221. https://doi.org/10.1111/j.1469-8986.1989.tb03159.x

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

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215-229.