Vocabulary Learning and Instruction
Volume 10, Issue 2, December 2021
http://vli-journal.org

Correlations of Modalities of Written Vocabulary Knowledge to Listening and Reading Proficiency: A Comparison

Jeffrey Stewart[a], Stuart McLean[b], and Aaron Olaf Batty[c]
[a]*Tokyo University of Science;* [b]*Momoyama Gakuin University;* [c]*Keio University*

Abstract

In recent years, there has been increasing debate and research regarding which modality of vocabulary knowledge has the strongest correlation to reading, with particular focus on distinctions between testing L2 form and L2 meaning, and between recall of answers from memory and recognition of answers from fixed options. However, relatively little attention has been paid to find out which modality has the strongest correlation to listening ability. A recent meta-analysis by Zhang and Zhang (2020) indicated that meaning recall was the superior predictor of reading proficiency. Although their results showed that form recall had the highest correlation to listening, the difference between form recall and meaning recall was statistically insignificant. The present study uses data from McLean et al. (2020) of learner responses to 1000-item vocabulary tests employing written tests of meaning recall, form recall, meaning recognition and Yes/No modalities, sampling them with replacement to create thousands of 100-item tests using a bootstrapping approach. The test scores were then correlated to measures of listening and reading proficiency for comparison. The results indicated that for written tests, meaning recall, form recall, meaning recognition and form recognition had the strongest correlations to both reading and listening, in descending order. All comparisons were statistically significant.

Keywords: Vocabulary, vocabulary testing, meaning recall, listening, reading

1 Background

In recent years, there has been increasing debate in the field of vocabulary learning and instruction regarding which modality of vocabulary knowledge has the strongest correlation to reading proficiency (Stewart et al., 2021; Stoeckel et al., 2021; Webb, 2021). A meta-analysis by Zhang and Zhang (2020) indicated that meaning recall, wherein learners recall and write or type L1 word meaning from memory after encountering the L2 written form, had a stronger correlation than both form recall (see the meaning, produce the L2 word form from memory) and the commonly used meaning recognition format (see the L2 word form, select a correct definition of the word from a list of fixed options, e.g., multiple-choice tests).

Please cite this article as: Stewart, J., McLean, S. and Batty, A. O. (2021). Correlations of modalities of written vocabulary knowledge to listening and reading proficiency: A comparison. *Vocabulary Learning and Instruction, 10*(2), 55–63. https://doi.org/10.7820/vli.v10.2.Stewart

Shortly before Zhang and Zhang's meta-analysis was released, McLean et al. (2020) published an additional study of the relationship between modalities of vocabulary and reading proficiency using a bootstrapping approach. Bootstrapping involves sampling a population with replacement in order to reach better estimates of confidence intervals and determine how the estimates from replicate experiments could be distributed (Kulesa et al. 2015). In the McLean et al. (2020) study, learners took 1,000-item tests in each modality, which were continually sampled with replacement to produce thousands of test forms and correlated to scores on the Reading section of the Test of English for International Communication (TOEIC®) test (https://www.ets.org/toeic). The study found further evidence that meaning recall was superior to meaning recognition as a predictor of reading ability. Written meaning recall was also found to have statistically and significantly higher correlations to reading when compared to form recall and form recognition modalities of vocabulary knowledge, as measured by L1-L2 multiple choice tests and written form recognition Yes/No tests, respectively.

Relatively less conclusive findings have been reported regarding written modalities of vocabulary knowledge as they relate to L2 listening ability. As listening is essentially a receptive aspect of language proficiency, theoretically meaning recall vocabulary knowledge could also be a strong predictor of listening ability. Despite this, Zhang and Zhang's meta-analysis more tentatively suggested that Form Recall[1] could possibly have a higher correlation to listening [$r = 0.63$ (95% CI = 0.53 – 0.72)] when compared to meaning recognition [$r = 0.50$ (95% CI = 0.41 – 0.58)].

Although it seems plausible that an L2 to L1 modality of vocabulary knowledge such as meaning recall could have stronger correlations to the receptive forms of language proficiency such as listening, Zhang and Zhang's finding that form recall could potentially be the superior predictor of listening ability was in alignment with a competing theory proposed by Laufer and Goldstein (2004) and Laufer et al. (2004). These studies found that written form recall was the most difficult form of vocabulary knowledge. Laufer et al. (2004) combined these four forms of vocabulary knowledge into a single unidimensional scale under the Rasch model. Under this framework, the mastery of "stronger" (i.e., more difficult) forms of vocabulary knowledge would imply mastery of "weaker" (i.e., less difficult) forms of vocabulary knowledge. This means that stronger forms of knowledge could have stronger correlations to other aspects of language proficiency more generally, even in cases where weaker forms of knowledge may appear to have stronger theoretical links to them, such as receptive vocabulary knowledge and its relationship to reading. However, Zhang and Zhang's results were inconclusive. This was because although the difference between meaning recall and meaning recognition was statistically significant, the difference between form recall [$r = 0.63$ (95% CI = 0.53 – 0.72)] and meaning recall [$r = 0.58$ (95% CI = 0.54 – 0.62)] was not. Therefore, there is still doubt as to whether written receptive form recall or written receptive meaning recall is the better predictor of listening.

An important caveat must be mentioned when discussing modalities of vocabulary knowledge in relation to listening ability. It is very likely that an auditory vocabulary test using spoken forms of words would have the strongest correlation to listening ability; evidence suggests that learners' written and spoken receptive vocabulary knowledge differ significantly (Masrai, 2020; Milton et al., 2010, 2013; Milton & Hopkins, 2006; Mizumoto & Shimamoto, 2008; Uchihara & Harada, 2018). Summarizing the results of such research, Zhang and Zhang's (2020) meta-analysis found that auditory-modality vocabulary tests had an average correlation of 0.6 to listening, compared to 0.52 for orthographic (written) modality tests. Although the difference was statistically insignificant, this lends further support to the hypothesis.

However, the existing spoken receptive levels tests have their limitations, most notably in their item formats. The Aural Lex (Milton & Hopkins, 2006) utilizes a spoken Yes/No format that does not require learners to demonstrate knowledge of the form-meaning link. The spoken receptive meaning recognition (multiple-choice) Listening Vocabulary Levels Test (LVLT) is only available for Japanese (McLean et al., 2015), Chinese (Zhang & Graham, 2020) and Vietnamese (Ha, 2021) learners, as tests that expect to measure spoken receptive meaning-recognition lexical knowledge should present answer options in the learners' L1 so as not to confound L2 written receptive and spoken receptive ability. However, while future research should strive to utilize appropriate tests, it is not always possible.

Cautions about using written tests of vocabulary knowledge to predict forms of proficiency that involve spoken vocabulary should be heeded. As Beglar (2010) advised, using written meaning recognition format tests such as the VST to measure listening vocabulary size "is not recommended as reading and listening vocabulary sizes can vary considerably" (p. 114). However, there may be value in determining if there is a single vocabulary test type that has, on average, the highest correlation to a wide range of L2 competencies, including listening, reading, writing and speaking. While it is highly unlikely that one test format will be superior in all situations, it may be possible to identify the best trade-off in terms of practicality and efficacy in situations where researchers wish to measure proficiency on multiple skills, but only have limited time to test vocabulary knowledge. Perhaps such an item format could be identified as an ideal overall measure of general L2 vocabulary knowledge.

In this paper, we will re-examine the data from Mclean et al. (2020) by correlating the aforementioned aspects of written vocabulary knowledge (meaning recall, form recall, meaning recognition and Yes/No) to listening proficiency as measured by the Listening section of the TOEIC test, using the same bootstrapping method as the original paper. A major limitation of this study is that the examined data set does not include the testing of spoken forms of words. Nor does it include measures of the productive skills of spoken and written proficiency. However, the results may provide evidence that undermines or lends further support to some of Zhang and Zhang's statistically insignificant findings and, at a minimum, shed some light on how written modalities of form, meaning, recall and recognition can affect predictions of listening and reading ability and confirm if differences exist depending on whether reading or listening is

tested. Future studies can examine if the findings related to these modalities are generalizable to the same modalities as applied to auditory tests of spoken word forms.

2 Method

One hundred and three learners took four 1,000-item vocabulary tests of the third 1,000 most frequent words according to the New General Service List (NGSL; Browne et al., 2013). Each of the four tests examined a different modality of vocabulary knowledge: meaning recall, meaning recognition, form recall and form recognition (as measured by a Yes/No test). The 1,000 items were sampled with replacement to create 1,000 100-item tests, the scores of which were correlated to the learners' scores on the Listening section of the TOEIC test. For further details on these tests, the tested sample of learners, and test administration, please refer to McLean et al. (2020).

3 Analysis

One thousand 100-item tests were generated by sampling with replacement from the McLean et al. (2020) data set for each of the modalities in question: written meaning recall, written meaning recognition, written form recall and written Yes/No. These test scores were then correlated to learners' TOEIC Listening scores, resulting in 4,000 correlations. This process was then repeated in order to compare the same conditions of correlations to TOEIC Reading scores. Following the procedure of McLean et al. (2020), a factorial 2 × 4 analysis of variance (ANOVA) was then conducted with the predictor variables "Channel" (listening or reading proficiency) and "Modality" (vocabulary knowledge type). The results of this ANOVA, descriptive statistics and post-hoc tests of the conditions can be seen below in Tables 1–3.

Channel (listening or reading proficiency) had a statistically significant effect on correlations, with correlations to listening being slightly lower than correlations to reading. However, this difference was very slight, with average correlations only 0.012 higher for reading, and a partial eta-squared of 0.119, indicating a negligible effect size. The effect of modality (type of vocabulary knowledge) on correlations to reading and listening proficiency was substantially higher, with a large partial eta-squared of 0.892. The interaction effect between channel and modality was statistically significant, but, as with channel, very slight in terms of effect size.

The examined forms of vocabulary knowledge displayed the same ranks of correlational strength to both listening and reading proficiency. As can be seen in Table 2, the written meaning recall had the strongest correlation to both listening and reading, at $r = 0.765$ and 0.778 respectively. For both forms of proficiency, meaning recall was followed by form recall, meaning recognition and Yes/No, in that order.

The differences between correlations, where each r value was a case, were significant for all conditions, as can be seen in Table 3. This includes

Table 1. Factorial ANOVA of Vocabulary Knowledge to Language Proficiency with predictor Variables Modality (Vocabulary Knowledge Type) and Channel (Proficiency Type)

Cases	Sum of squares	*df*	Mean square	*F*	*p*	η^2	ηp^2
Channel	0.277	1	0.277	1076.769	<0.001	.014	.119
Modality	17.054	3	5.685	22083.241	<0.001	.869	.892
Channel × Modality	0.248	3	0.083	320.506	<0.001	.013	.107
Residuals	2.057	7,992	2.574e-4				

Note. Type III Sum of Squares.

Table 2. Descriptive Statistics of Means of Correlation (*r*) by Proficiency Type and Written Vocabulary Knowledge Type

Channel	Modality	Mean	SD	*N*
Listening	Form Recall	0.745	0.016	1,000
	Form Recognition	0.650	0.017	1,000
	Meaning Recall	**0.765**	0.012	1,000
	Meaning Recognition	0.671	0.019	1,000
Reading	Form Recall	0.742	0.017	1,000
	Form Recognition	0.659	0.016	1,000
	Meaning Recall	**0.778**	0.012	1,000
	Meaning Recognition	0.699	0.018	1,000

Note. Highest correlations per channel are indicated in bold.

comparisons between written form recall and written meaning recall and listening proficiency. Written meaning recall's correlation to listening was 0.765 compared to 0.745 for form recall, with a Cohen's *d* effect size of 1.41.

4 Discussion

The findings show that correlations of forms of written vocabulary knowledge to listening skills are very similar to their correlations to reading proficiency, albeit slightly lower. In contrast to Zhang and Zhang's (2020) findings, meaning recall, rather than form recall, was the best predictor of listening proficiency. Furthermore, unlike in Zhang and Zhang's study, the difference between the two was statistically significant ($p < 0.001$), with a Cohen's *d* effect size of 1.41. Both Zhang and Zhang and McLean et al. (2020) found that meaning recall was superior to meaning recognition as a predictor of reading ability. Although Zhang and Zhang reported the same relationship with regard to listening ability, the result was statistically insignificant. The current study reconfirms the finding to a statistically significant degree, at least with regard to written tests. A caveat to this claim is that Zhang and Zhang considered meta-analytic data while this current study presents correlations where each acts as a case in parametric testing.

Table 3. Post Hoc Comparisons—Channel × Modality

		Mean difference	*SE*	*t*	*P Tukey*
Listening, Form Recall	Reading, Form Recall	0.003	7.175e-4	4.056	0.001
	Listening, Form Recognition	0.095	7.175e-4	132.818	<0.001
	Reading, Form Recognition	0.086	7.175e-4	120.260	<0.001
	Listening, Meaning Recall	−0.020	7.175e-4	−27.260	<0.001
	Reading, Meaning Recall	−0.032	7.175e-4	−45.044	<0.001
	Listening, Meaning Recognition	0.074	7.175e-4	103.216	<0.001
	Reading, Meaning Recognition	0.046	7.175e-4	63.872	<0.001
Reading, Form Recall	Listening, Form, Recognition	0.092	7.175e-4	128.762	<0.001
	Reading, Form Recognition	0.083	7.175e-4	116.205	<0.001
	Listening, Meaning Recall	−0.022	7.175e-4	−31.316	<0.001
	Reading, Meaning Recall	−0.035	7.175e-4	−49.099	<0.001
	Listening, Meaning Recognition	0.071	7.175e-4	99.160	<0.001
	Reading, Meaning Recognition	0.043	7.175e-4	59.817	<0.001
Listening, Form Recognition	Reading, Form Recognition	−0.009	7.175e-4	−12.557	<0.001
	Listening, Meaning Recall	−0.115	7.175e-4	−160.078	<0.001
	Reading, Meaning Recall	−0.128	7.175e-4	−177.861	<0.001
	Listening, Meaning Recognition	−0.021	7.175e-4	−29.602	<0.001
	Reading, Meaning Recognition	−0.049	7.175e-4	−68.945	<0.001
Reading, Form Recognition	Listening, Meaning Recall	−0.106	7.175e-4	−147.521	<0.001
	Reading, Meaning Recall	−0.119	7.175e-4	−165.304	<0.001
	Listening, Meaning Recognition	−0.012	7.175e-4	−17.045	<0.001
	Reading, Meaning Recognition	−0.040	7.175e-4	−56.388	<0.001
Listening, Meaning Recall	Reading, Meaning Recall	−0.013	7.175e-4	−17.783	<0.001
	Listening, Meaning Recognition	0.094	7.175e-4	130.476	<0.001
	Reading, Meaning Recognition	0.065	7.175e-4	91.133	<0.001
Reading, Meaning Recall	Listening, Meaning Recognition	0.106	7.175e-4	148.259	<0.001
	Reading, Meaning Recognition	0.078	7.175e-4	108.916	<0.001
Listening, Meaning Recognition	Reading, Meaning Recognition	−0.028	7.175e-4	−39.344	<0.001

Note. *p*-value adjusted for comparing a family of 8. All vocabulary knowledge modalities are written form.

McLean et al. (2020) proposed two theories for why meaning recall could outperform meaning recognition as a correlate of reading, which contrasted with the arguments from some researchers that meaning recognition could have more applicability to the skills required for reading (Laufer & Aviad-Levitzky, 2017). First, fixed options could introduce a source of error not contained in recall item formats due to guessing effects. Second, working under a continuum/cline theory of vocabulary strength (Laufer et al., 2004; Stewart et al., 2012), stronger forms of vocabulary knowledge could subsume weaker forms, meaning a learner who has mastery over form recall would also have mastery over meaning recall and meaning recognition, meaning form recall could therefore

have equal or better correlations not only to productive language skills but also receptive language skills such as reading and listening.

However, the present findings indicate that even if accurate with regard to meaning recall and meaning recognition, this theory does not appear to extend more broadly to form recall modalities. When placed in comparison to one another, recall of meaning outperforms recall of form, and both recall modalities examined in this study outperform both recognition modalities, regardless of whether form or meaning is tested. This implies that the distinction between recall and recognition exerts a stronger influence on correlations to other aspects of language proficiency than the distinction between form and meaning.

The results suggest that when possible, recall format tests are preferable measures of vocabulary knowledge when attempting to relate vocabulary knowledge to other language proficiency skills, particularly for research that must rely on sensitive measurement instruments. Although the differences in correlation between formats may appear to be small, these differences can become important in hypothesis testing, particularly in experiments that are underpowered, a common occurrence in the field of second language acquisition (e.g., Nicklin & Vitta, 2021). More sensitive and reliable measures can help to guard against Type II error, where researchers falsely reject correct hypotheses due to a lack of statistical significance.

With regard to testing learner vocabulary in pedagogical contexts, the tradeoff of the higher predictive validity and internal reliability (McLean et al., 2020) of recall item formats must be balanced with the relative difficulty in administering and scoring recall-format tests relative to recognition-format tests. Browser-based vocabulary tests such as Vocableveltest.org (McLean & Raine, 2018), which can be completed at computer terminals or on learners' smartphones, can help mitigate many of these issues. In addition to compiling learners' answers in digital format, the software can automatically score whitelisted and blacklisted answers and flag novel responses for manual grading. With each successive administration of the test, fewer and fewer responses require manual scoring. The results of the current study also suggest that form recall tests represent a tradeoff that further simplifies the marking of recall responses while maintaining a relatively strong correlation to receptive language proficiency, as form recall also outperforms meaning recognition measures as a correlate to receptive language proficiency. Scoring can be further simplified if learners can provide the L2 word form rather than a meaning as it greatly limits the number of possible answers. It is also possible that, when used as a correlate to general language proficiency, form recall's somewhat lower correlations to receptive language proficiency could be balanced by a higher correlation to productive language skills. Future research should attempt to replicate these results regarding form, meaning, recall and recognition modalities using an auditory test format of spoken word forms, as the modalities examined in the current study were all used written forms. It could also be beneficial for future studies to examine the correlations of meaning recall and form recall in relation to speaking and writing, in order to examine and compare their overall predictive power with regard to general language proficiency.

Note:

1. In Zhang and Zhang (2020) form, recall includes dictation modalities where learners have to dictate the word they hear. Dictation item formats do not require learners to demonstrate knowledge of the spoken form meaning link.

References

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

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

Ha, H. T. (2021). A Rasch-based validation of the Vietnamese version of the Listening Vocabulary Levels test. *Language Testing in Asia*, *11*(1), 1–19.

Kulesa, A., Krzywinski, M., Blainey, P., & Altman, N. (2015). Sampling distribu-tions and the bootstrap. *Nature Methods, 12*, 477–478. https://doi.org/10.1038/ nmeth.3414

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., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202–226. https://doi.org/10.1191/0265532204lt277oa

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

Masrai, A. (2020). Exploring the impact of individual differences in aural vocabulary knowledge, written vocabulary knowledge and working memory capacity on explaining L2 learners' listening comprehension. *Applied Linguistics Review*, *11*(3), 423–447. https://doi.org/10.1515/applirev-2018-0106

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research, 19*(6), 741–760.

McLean, S. & Raine, P. (2018). *VocabLeve ltest.org [Online program].* 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

Milton, J., Alexiou, T., & Mattheoudakis, M. (2013). Knowledge of spoken form. In: J. Milton & T. Fitzpatrick (Eds.), *Dimensions of vocabulary knowledge* (pp. 13–29). Palgrave Macmillan.

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *Canadian Modern Language Review/ La Revue Canadienne Des Langues Vivantes*, *63*(1), 127–147. https://doi.org/10.3138/cmlr.63.1.127

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. del Mar Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–98). Multilingual Matters.

Mizumoto, A., & Shimamoto, T. (2008). A comparison of aural and written vocabulary size of Japanese EFL university learners. *Language Education & Technology*, *45*, 35–51. https://doi.org/10.24539/let.45.0_35

Nicklin, C., & Vitta, J. P. (2021). Effect-driven sample sizes in second language instructed vocabulary acquisition research. *The Modern Language Journal, 105*(1), 218–236. https://doi.org/10.1111/modl.12692

Stewart, J., Batty, A., & Bovee, N. (2012). Comparing multidimensional and continuum models of vocabulary acquisition: An empirical examination of the Vocabulary Knowledge Scale. *TESOL Quarterly*, *46*(4), 695–721. https:// doi.org/10.1002/tesq.35

Stewart, J., Stoeckel, T., McLean, S., Nation, P., & Pinchbeck, G. G. (2021). 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

Uchihara, T., & Harada, T. (2018). Roles of vocabulary knowledge for success in English-medium instruction: Self-perceptions and academic outcomes of Japanese undergraduates. *TESOL Quarterly*, *52*(3), 564–587.

Webb, S. (2021). A different perspective on the limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, *43*(2), 454–461. https://doi.org/10.1017/S0272263121000449

Zhang, P., & Graham, S. (2020). Learning vocabulary through listening: The role of vocabulary knowledge and listening proficiency. *Language Learning*, *70*(4), 1017–1053. https://doi.org/10.1002/tesq.453

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