

Critical Commentary

WHAT THE RESEARCH SHOWS ABOUT WRITTEN RECEPTIVE VOCABULARY TESTING

A REPLY TO WEBB

Jeffrey Stewart 

Tokyo University of Science

Tim Stoeckel 


University of Niigata Prefecture

Stuart McLean 

Momoyama Gakuin University

Paul Nation 

Victoria University of Wellington

Geoffrey G. Pinchbeck 

Carleton University

In response to our *State-of-the-Scholarship* critical commentary (Stoeckel et al., 2021), Stuart Webb (2021) asserts that there is no research supporting our suggestions for improving tests of written receptive vocabulary knowledge by (a) using meaning-recall items, (b) making fewer presumptions about learner knowledge of word families, and (c) using appropriate test lengths. As we will show, this is not the case. However, we think questions and concerns he raises reflect those of many who have used these tests until now without controversy, and we appreciate the opportunity to explain these issues in greater detail.

To begin, we think Webb has more common ground with our position than he may realize. We agree with many of his statements, and do not state otherwise in Stoeckel et al. (2021). For example, we agree that few if any vocabulary test makers have claimed their

We are grateful to Norbert Schmitt, Henrik Gyllstad, Christopher Nicklin, Joseph Vitta, Nick Bovee, and Dale Brown for their comments on earlier versions of this article.

Correspondence concerning this article should be addressed to Tim Stoeckel, University of Niigata Prefecture, 471 Ebigase, Higashi-ku, Niigata City, Niigata, Japan 950-8680. Email: stoeckel@unii.ac.jp

© The Author(s), 2021. Published by Cambridge University Press

tests should be used as substitutes for reading tests; we agree that despite this, vocabulary tests typically do show good correlations with reading; and we agree that despite that, such tests should not be used as evidence of reading comprehension. These matters are not in dispute. However, there are remaining points of disagreement to address.

TEST USE

Noting that “the premise on which their article was written is that the intended purpose of the VLT and VST is to measure vocabulary knowledge for the purpose of reading,” Webb appears to dispute that this was ever the case or an intention of the test makers. Webb further asserts that it is “not the intended purpose” of the tests “to accurately reveal the degree to which learners may reach key lexical coverage figures” (p. 458), track growth, or suggest vocabulary learning goals. However, while it is certainly true that these tests have not been sufficiently *validated* for these purposes, the fact that these applications factored into test makers’ *intentions* during their creation is clear from their own statements, as can be seen in Table 1.

Furthermore, while space constraints prevent a comprehensive list, the majority of uses of these tests in SLA literature are for purposes such as those previously mentioned rather than merely to check learner knowledge of words without reference to other considerations. The desire of researchers to use them these ways is sympathizable. As Webb notes, in general, vocabulary test scores do indeed correlate with other constructs, and if they could be employed only to measure vocabulary knowledge without any other such inferences permitted, their usefulness would be quite limited. Indeed, while we would welcome such restrictions, were Webb’s cautions about appropriate test use strictly followed it would all but mark the end of use of tests such as the Vocabulary Levels Test (VLT) and Vocabulary Size Test (VST) as variables in research published in journals such as *SSLA*.

We believe the long-standing confusion regarding the intended purposes of these tests stems from the fact that often, at the times of their creation, these tests did not have narrowly specified uses (Norbert Schmitt, personal communication, May 6, 2021) and, as we and a number of our colleagues would argue, many still do not have them today (Schmitt et al., 2020). Thus, it is understandable that teachers and researchers would use them for a wider variety of purposes than appropriate. We hope Stoeckel et al. (2021) acts as a caution against this.

Our own view is that while we do not support the notion that vocabulary comprehension alone is sufficient for reading comprehension (McLean, 2021), vocabulary knowledge is uncontroversially a *component* of reading, which can be useful as one of several variables in studies of reading proficiency. Furthermore, while vocabulary knowledge alone is not sufficient for reading comprehension, testing vocabulary mastery can at least ensure that lack of vocabulary knowledge is not an impediment for readers of given texts (Nation, 2009, p. 52). However, in all such applications, ideally as closely as possible test items should approximate how vocabulary is encountered in text (Schmitt et al., 2020). This leads us to the first suggestion in our original article, regarding choice of item format.

ITEM FORMAT

In addition to demonstrating that fixed options incorrectly increase estimates of vocabulary size (Gyllstad et al., 2015; McLean et al., 2015; Stewart & White, 2011) research has

TABLE 1. Some stated purposes of size and levels tests of written receptive vocabulary knowledge

Purpose	Citations from test creators and validation studies
Assessing vocabulary knowledge needed for reading	<p>“For instance, a vocabulary test claiming to test written receptive form-meaning link knowledge (i.e., the vocabulary knowledge needed for reading) could be administered alongside a reading comprehension test.” (VLT; Schmitt et al., 2020, p. 117)</p> <p>“Users of the test need to be clear what the test is measuring and not measuring. It is measuring written receptive vocabulary knowledge, that is the vocabulary knowledge required for reading.” (VST; Nation, 2012, para. 6)</p> <p>The UVLT “is a measure of receptive vocabulary knowledge indicating the degree to which test takers may be able to understand the meanings of words that they encounter in written text.” (Webb et al., 2017, p. 57)</p>
Selecting reading materials	<p>[The VLT] “can be utilized by teachers and administrators in a pedagogical context to inform decisions concerning whether an examinee is likely to have the lexical resources necessary to cope with certain language tasks, such as reading authentic materials.” (Schmitt et al., 2001, p. 56)</p> <p>“to know at what level learners should begin reading, it is useful to measure their receptive vocabulary size.... The [VLT] developed by Schmitt, Schmitt and Clapham (2001) provides a means of doing this.” (Nation, 2009, p. 52)</p> <p>“Their average raw score was 27.7/30, indicating that they had mastered that level (Schmitt, Schmitt, & Clapham, 2001) and should have little difficulty understanding all of the running words in the treatments.” (Webb, 2008, p. 234)</p>
Tracking growth	<p>Determining “the rate that words and lexical items are acquired” is one of the problems that can be solved with tests like the VLT. (Beglar & Hunt, 1999, p. 131)</p> <p>A “reason for measuring vocabulary size is to be able to chart the growth of learners’ vocabularies.” (VST; Nation & Beglar, 2007, p. 9)</p> <p>“Measuring knowledge of five sequenced word frequency levels should help teachers (and learners) to see the extent of vocabulary learning progress.” (UVLT; Webb et al., 2017, p. 55)</p>
Goal setting	<p>“[T]o set specific goals, it is essential to know if learners need to focus on high-frequency, academic, technical, mid-frequency or low-frequency words. This is best decided on by diagnostic testing using the Vocabulary Levels Test, or by size testing using the Vocabulary Size Test.” (Nation, 2013, p. 570)</p>

also consistently shown that, all else being equal, recall formats are more reliable than meaning-recognition formats testing the same words when learners are asked to attempt every item (McLean et al., 2020; McLean et al., 2016; Stewart, 2012; Stoeckel et al., 2019). The relatively poor discrimination of meaning-recognition items makes experiments using them more prone to Type II error, where hypotheses are rejected due to seemingly statistically insignificant results. Furthermore, there is a growing consensus in the literature (e.g., Grabe, 2009, p. 23; Kremmel & Schmitt, 2016, p. 378; Nation & Webb, 2011, pp. 219, 285–286) that meaning-recall represents an appropriate threshold of lexical knowledge for reading because, as in fluent reading, word meaning must be retrieved from memory rather than identified in a list of options.

Webb appears to contest this position by citing Laufer and Aviad-Levitzky (2017), who gave learners the meaning-recognition based VST, a parallel meaning-recall test, and a reading test. Departing from the VST's specifications (Nation, 2012), they had instructed learners to skip items testing words that they did not believe they knew. Perhaps as a consequence of this change, there was no statistically significant difference between correlations of the two tests to the reading measure (.91 and .92). Despite the insignificant result, the authors argued that meaning-recognition was the better predictor of reading ability. (Webb expressed surprise that we did not include this study in our review. As noted in our original paper, we excluded studies that allowed learners to skip unknown words on the meaning-recognition measure because research demonstrates that examinees use the option to skip differentially, which impacts the relationship between recognition and recall scores [Stoeckel et al., 2016].)

To better identify the differences in correlations between these modalities and reading proficiency, subsequent research by McLean et al. (2020) involving meaning-recall and meaning-recognition formats and reading proficiency used a bootstrapping approach to mitigate Type I and II errors. Both meaning-recall and meaning-recognition were tested bilingually for direct comparisons. By sampling with replacement for thousands of iterations, McLean et al. demonstrated that with very little overlap, meaning-recall outperformed meaning-recognition as a predictor of reading proficiency. Webb notes that meaning-recognition was also correlated to reading proficiency. We do not dispute this; as we note in the preceding text, all vocabulary tests will correlate to reading to at least some extent. However, the goal of our paper was to suggest improvements to vocabulary tests. As McLean et al. show, for a test of 30 items, meaning-recall outperforms meaning-recognition in average correlations to reading 0.74 to 0.65 ($d = -3.622$; see Figure 1), a distinction that becomes even clearer for tests with more items (Figure 2). Such differences in variables can have substantial impacts on models, so researchers should take note.

Nor are such findings restricted to the previously mentioned study. A meta-analysis by Jeon and Yamashita (2014) also found that meaning-recall was the better predictor, but could not attain statistical significance due to the small number of meaning-recall studies examined (7). However, a more recent meta-analysis by Zhang and Zhang (2020) including 21 studies using meaning-recall and 14 using meaning-recognition found that mean correlations between meaning-recall and reading proficiency ($r = .66$ [.58, .71]) are significantly stronger than those between meaning-recognition and reading proficiency¹ ($r = .53$ [.49, .57]). Debate about uses of tests such as the VLT and the VST aside, the research seems clear: if one *does* desire to measure vocabulary as it relates to reading, meaning-recall appears to be the better option.

Although he acknowledges the risk of overestimation in meaning-recognition, Webb argues that meaning-recall could underestimate vocabulary size, suggesting meaning-recognition provides more “sensitivity” in scoring. Research in which learners are orally interviewed about their answers on meaning-recognition items shows that despite higher mean scores, the format is highly insensitive, with learners choosing the items they do for a variety of disparate reasons, including construct-irrelevant ones such as test strategies and blind guessing (Gyllstad et al., 2015; McDonald, 2015).

It is true that meaning-recall tests such as Aviad-Levitzky et al. (2019) that demand answers with perfect L2 English spelling of target word synonyms can depress scores for

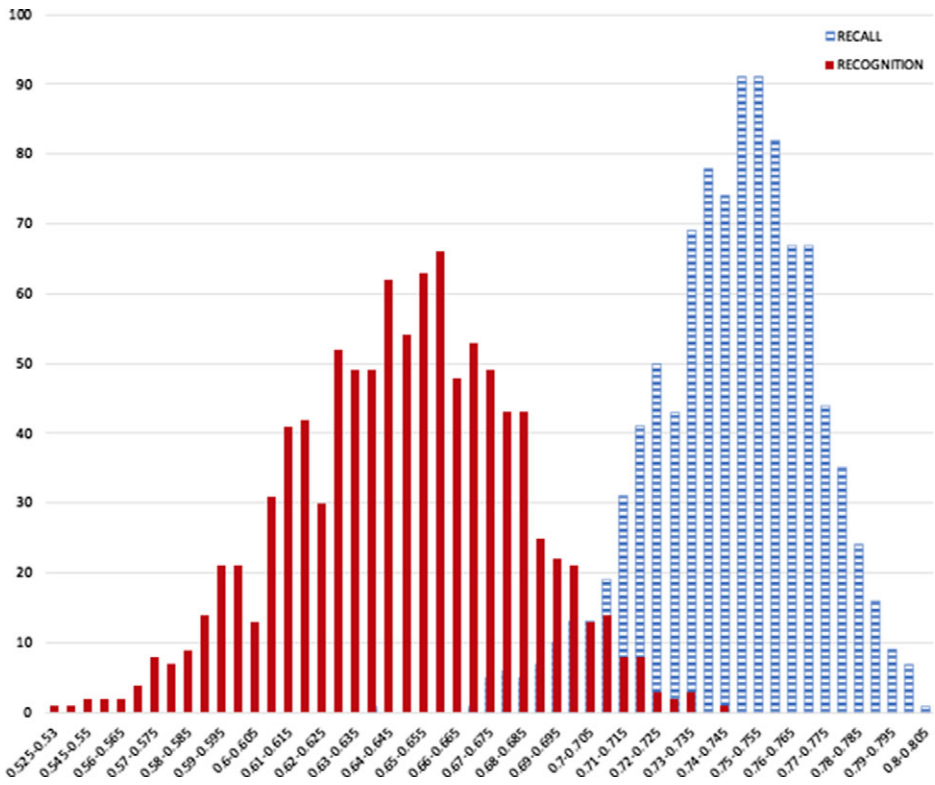


FIGURE 1. Histograms of bootstrapped correlations of meaning-recall and meaning-recognition to reading proficiency, 30 items (adapted from McLean et al., 2020).

reasons unrelated to learners’ understanding of meaning, particularly given English’s complex orthography. However, an advantage of recall tests is that unlike fixed-option meaning-recognition tests, researchers retain learners’ free responses, which can then be examined and graded as leniently as desired. Although a common complaint of this procedure is the time required to mark answers, online resources such as www.vocableveltest.org greatly expedite this process. Novel choices are presented to the researcher and can then be whitelisted or blacklisted during initial scoring, allowing for automated scoring of those same responses thereafter.

Moving on to practical matters, Webb expresses concern that meaning-recall tests may take longer to administer to learners. However, the time required need not be prohibitive. Recent research by McLean et al. (2020) indicates that meaning-recall increases test time by roughly 28%, meaning a 10-minute meaning-recognition test would still require less than 13 minutes to complete using meaning-recall. Webb further argues that monolingual tests may be more appropriate for many ESL settings. While this is a legitimate concern in many contexts, as mentioned already, online resources such as www.vocablevelstest.org can simplify this process by permitting the inclusion of multiple L1s as possible answer options.

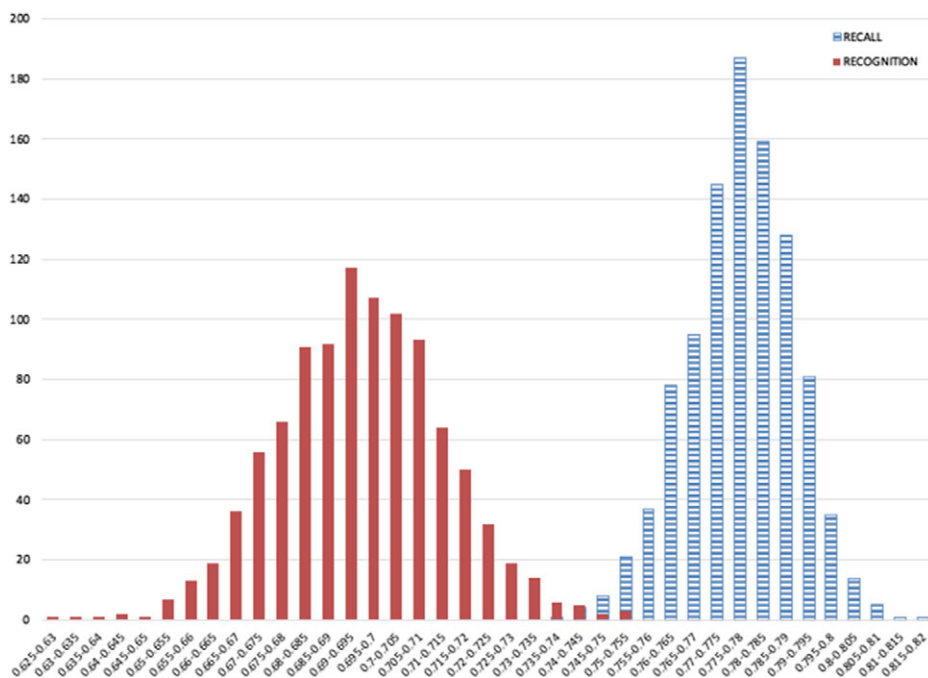


FIGURE 2. Histograms of bootstrapped correlations of meaning-recall and meaning-recognition to reading proficiency, 100 items (adapted from McLean et al., 2020).

These arguments in favor of meaning-recall do not mean meaning-recognition tests have no value at all. In cases in which learners do not have access to computers or cell phones, multiple-choice tests may have advantages in classroom contexts where teachers need fast results. Scoring of multiple L1s is possible in meaning-recall tests, but involves greater initial overhead and greater complexity in scoring standards. While more research is necessary, it is possible meaning-recall tests could underestimate knowledge when it is difficult to express meaning. On balance however, research shows meaning-recall is the preferred option for tests measuring vocabulary knowledge for reading.

LEXICAL UNIT

Just as the choice of item format for a test should consider the purpose of the test and the learners taking the test, the choice of a lexical unit should also take account of such considerations. Bauer and Nation's (1993) level six word family (WF6) is too inclusive for some purposes and for many learners, and we need to develop tests that use more appropriate word family levels for the high-frequency and initial mid-frequency vocabulary. Use of WF6 in tests assumes that learners know most or all family members to the same level of knowledge that the target word was tested at. This assumption is unsupported by research with L2 learners of English from a range of proficiency levels (McLean, 2018; Stoeckel et al., 2020; Ward & Chuenjundaeng, 2009).

Webb argues lemma-based instruments require testing more words. However, from a statistical standpoint this is incorrect: for size tests, precision is a function of sample size rather than population size (Smith, 2004), so keeping item numbers constant does not affect accuracy. For levels tests, we agree more levels may be desirable for lemma-based instruments, but this need not be a serious concern. As Webb has observed, learners need only complete test levels at their proficiency level (Webb et al., 2017).

Webb cautions against reaching conclusions on this matter until more research is available. However, currently the prudent choice is to assume less derivational knowledge on the part of learners, not more. The available evidence suggests that learners well beyond beginner level have trouble recalling the meaning of some derivational forms of known basewords (see Brown et al., 2020, *in press*; McLean, 2021 for recent reviews). Tests relying on smaller lexical units will still be effective for learners regardless of proficiency, but the same cannot currently be ensured for WF6-based tests.

TARGET WORD SAMPLE SIZE

Webb argues that ideal item counts for size and levels tests are “not straightforward” on the grounds that “the greater the number of good test items, the more accurately a test should help to assess knowledge” (p. 458). It is true that good items have higher discrimination, reducing tests’ standard error of measurement. However, although multiple-choice items can be screened and improved, all else being equal, tests with recall items of the same words demonstrate superior quality (McLean et al., 2020; Stewart, 2012). Furthermore, in regard to size estimation, regardless of item quality an axiom of inferential statistics is that the larger the sample size, the more reliable the vocabulary size estimate,² and even theoretically perfect item discrimination does not obviate the need for sufficient sample sizes in this regard (Gyllstad et al., 2015, 2020b). In his response, Webb calls for examining how test performance is affected by manipulating disputed variables. Just such a study was conducted by Gyllstad et al. (2020b). An example of the difference item counts make to accuracy in size estimation is illustrated in Figure 3.

As explained in Stoeckel et al. (2021), research indicates that size estimation based on item response theory (IRT) can help address concerns about test length. Research by Culligan (2008) and Gibson and Stewart (2014) illustrates how IRT-based computer adaptive tests can tailor to learners’ ability levels, mitigating the need for many items either far above or below learner ability. Although it is still advisable to test sufficiently at appropriate difficulty levels, IRT can greatly shorten tests of words with wide ranges of frequencies and difficulties, such as the VST.

CONCLUSION

Webb concluded by expressing his belief that no empirical evidence exists to support our positions regarding meaning-recall items, smaller lexical units, and appropriate test lengths. We hope the research cited in this commentary puts his concerns to rest and makes the evidence for our positions clear. However, we wholeheartedly agree with Webb’s call for further research regarding our suggestions, and hope this dialogue inspires more such studies.

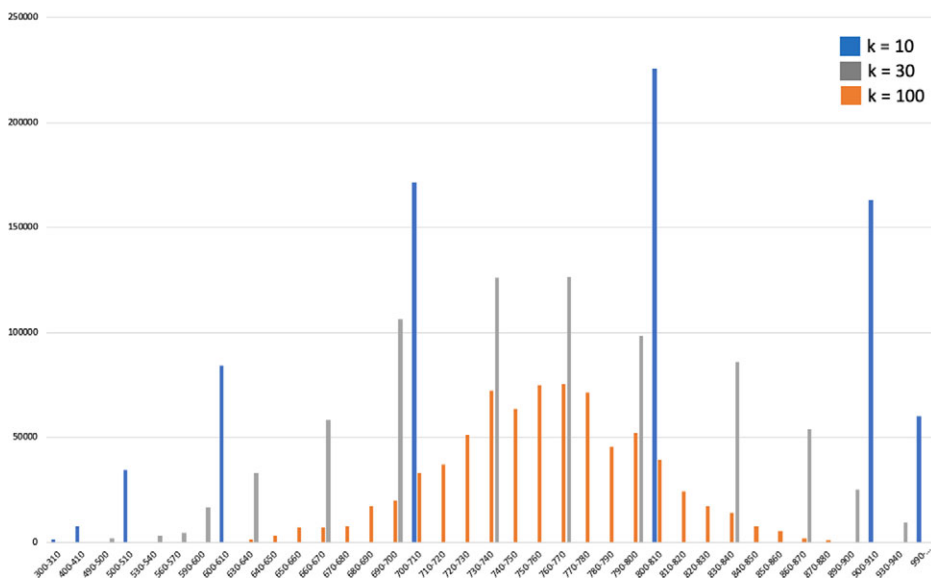


FIGURE 3. Monte Carlo study of vocabulary size estimates using tests of 10, 30, and 100 items (adapted from Gyllstad et al., 2020b).

Note: The true number of words known by this learner is 750.

To use a contemporary term, commentaries such as Stoeckel et al. (2021), McLean (2021), Stewart (2014), and Schmitt et al. (2020) “problematize” widely used vocabulary tests. Problems are rarely welcomed with open arms. Much in the same way dated statistical standards can attain a semblance of authority through the precedent of their past use, standards for instruments used in research can take on an air of unimpeachability when they have been used unquestioned for so long. However, it is important not to confuse what is familiar with what is preferable. Leaving precedent unquestioned can prevent appropriate scrutiny of past research.

As a final thought, it should be noted that initially each of the aforementioned characteristics of conventional vocabulary tests (i.e., fixed-response, meaning-recognition tests with relatively few randomly sampled items per level based on word families) were established with little empirical evidence, and based on early and underdeveloped perspectives of validation (Norbert Schmitt, personal communication, May 6, 2021). Validations of newer tests that have inherited these characteristics have rarely attempted to examine their underlying assumptions. While we appreciate Webb’s calls for further evidence, we hope that going forward similar scrutiny is applied to older standards as is now applied to the increasing calls for updated ones.

NOTES

¹Form-recall also outperformed meaning-recognition in both this study and in Mclean et al. (2020).

²We have found that treating size tests as polls of proportions of known words results in slightly better confidence interval estimation than test SEM, despite the latter accounting for item variance (Gyllstad et al., 2020a).

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, 345–368. <https://doi.org/10.1080/15434303.2019.1649409>
- Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, 6, 253–279. <https://doi.org/10.1093/ijl/6.4.253>
- Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, 16, 131–162. <https://doi.org/10.1177/026553229901600202>
- Brown, D., Stewart, J., Stoeckel, T., & McLean, S. (in press). The coming paradigm shift in the use of lexical units. *Studies in Second Language Acquisition*.
- 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*. Advance online publication. <https://doi.org/10.1093/applin/amaa061>
- Culligan, B. (2008). Estimating word difficulty using yes/no tests in an IRT framework and its application for pedagogical objectives. (Unpublished doctoral dissertation). Temple University, Japan.
- Gibson, A., & Stewart, J. (2014). Estimating learners' vocabulary size under item response theory. *Vocabulary Learning and Instruction*, 3, 78–84. <http://www.vli-journal.org/issues/03.2/issue03.2.full.pdf#page=82>
- Grabe, W. (2009). *Reading in a second language*. Cambridge University Press.
- Gyllstad, H., McLean, S., & Stewart, J. (2020a). [Unpublished raw data comparing confidence intervals for vocabulary size estimates produced by a poll of a proportion and the standard error of measurement.]
- Gyllstad, H., McLean, S., & Stewart, J. (2020b). Using confidence intervals to determine adequate item sample sizes for vocabulary tests: An essential but overlooked practice. *Language Testing*. Advance online publication. <https://doi.org/10.1177/0265532220979562>
- 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, 278–306. <https://doi.org/10.1075/itl.166.2.04gy1>
- Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, 64, 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, 377–392. <https://doi.org/10.1080/15434303.2016.1237516>
- 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, 729–741. <https://doi.org/10.1111/modl.12431>
- McDonald, K. (2015). The potential impact of guessing on monolingual and bilingual versions of the vocabulary size test. *Osaka JALT Journal*, 2, 44–61. <http://www.osakajalt.org/journal/>
- McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, 39, 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 operationalized. *Reading in a Foreign Language*, 33, 126–140. <https://nflrc.hawaii.edu/rfl/item/528>
- 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, 26–35. <http://vli-journal.org/wp/wp-content/uploads/2015/10/vli.v04.1.2187-2759.pdf#page=31>
- McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, 37, 389–411. <https://doi.org/10.1177/0265532219898380>
- McLean, S., Stewart, J., & Kramer, B. (2016, September 12–14). A comparison of multiple-choice and yes/no test formats with a meaning-recall knowledge criterion [Paper presentation]. Vocab@Tokyo. Tokyo, Japan.
- Nation, I. S. P. (2009). *Teaching ESL/EFL reading and writing*. Routledge.
- Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Heinle.
- 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>
- Nation, P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

- Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31, 9–13. https://jaltpublications.org/lt/issues/2007-07_31.7
- 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, 109–120. <https://doi.org/10.1017/S0261444819000326>
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18, 55–88. <https://doi.org/10.1177/026553220101800103>
- Smith, M. H. (2004). A sample/population size activity: Is it the sample size of the sample as a fraction of the population that matters? *Journal of Statistics Education*, 12. <https://doi.org/10.1080/10691898.2004.11910735>
- Stewart, J. (2012). A multiple-choice test of active vocabulary knowledge. *Vocabulary Learning and Instruction*, 1, 53–59.
- Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, 11, 271–282. <http://vli-journal.org/issues/01.1/issue01.1.09.pdf>
- Stewart, J., & White, D. A. (2011). Estimating guessing effects on the vocabulary levels test for differing degrees of word knowledge. *TESOL Quarterly*, 45, 370–380. <https://doi.org/10.5054/tq.2011.254523>
- Stoeckel, T., Bennett, P., & McLean, S. (2016). Is “I Don’t Know” a viable answer choice on the Vocabulary Size Test? *TESOL Quarterly*, 50, 965–975. <https://doi.org/10.1002/tesq.325>
- Stoeckel, T., Ishii, T., & Bennett, P. (2020). Is the lemma more appropriate than the flemma as a word counting unit? *Applied Linguistics*, 41, 601–606. <https://doi.org/10.1093/applin/amy059>
- Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, 43, 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, 461–469. <https://doi.org/10.1016/j.system.2009.01.004>
- Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, 20, 232–245. <https://nflrc.hawaii.edu/rfl/item/178>
- 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, 454–461. <https://doi.org/10.1017/S0272263121000449>
- Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test. *ITL - International Journal of Applied Linguistics*, 168, 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*. Advance online publication. <https://doi.org/10.1177/1362168820913998>