Studies in Second Language Acquisition, **43** (2021), 454–461
doi:10.1017/S0272263121000449

Critical Commentary

A DIFFERENT PERSPECTIVE ON THE LIMITATIONS OF SIZE AND LEVELS TESTS OF WRITTEN RECEPTIVE VOCABULARY KNOWLEDGE

Stuart Webb

University of Western Ontario

Stoeckel, McLean, and Nation's article, *Limitations of Size and Levels Tests of Written Receptive Vocabulary Knowledge* (2021), discusses whether the Vocabulary Size Test (VST; Coxhead et al., 2015; Nation & Beglar, 2007) and the Vocabulary Levels Test (VLT; Nation, 1983; Schmitt et al., 2001; Webb et al., 2017) are effective at measuring the vocabulary knowledge necessary for the purpose of reading. Stoeckel et al. suggest that these tests are likely to overestimate receptive vocabulary knowledge and that there are three ways in which the tests could be improved. The first way to improve the tests is by moving from a recognition format to a recall format. The second way is to move from using word families as the lexical unit to using lemmas. Their third suggestion is to increase the number of target items in the tests. Stoeckel et al. conclude that existing size and levels tests lack the accuracy necessary for many specified testing purposes.

Although it is useful to look at different ways to improve on measures of lexical knowledge, there is little research evidence supporting the claims made by Stoeckel et al., and there are several aspects of their article that should be considered further. First, 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.[1] However, the VLT was developed to reveal to teachers where they should focus vocabulary learning (Nation, 1983, 1990, 2008; Nation & Webb, 2011; Read, 2000; Webb & Nation, 2017; Webb et al., 2017). The VST was developed to measure L2 learner knowledge of the most frequent 14,000 word families as a whole (Nation & Beglar, 2007), and was later expanded to measure both nonnative and native speakers' knowledge of the most frequent 20,000 word families as a whole (Coxhead et al., 2015). There is currently no research indicating that the tests are not working well for these purposes. Neither test was developed and validated for the purpose of predicting reading comprehension. In fact, Beglar (2010,

I am grateful to Frank Boers, Batia Laufer, Ruslan Suvorov, Elke Peters, Ana Pellicer-Sánchez, and Paul Nation for their helpful comments on earlier drafts of this article.
Correspondence concerning this article should be addressed to Stuart Webb, University of Western Ontario, Faculty of Education, 1137 Western Road, London, ON N6G 1G7, Canada. E-mail: swebb27@uwo.ca

p.114) reports that "test-takers' responses provide only a rough indication of how well they can read, so the VST should not be viewed as a substitute for a reading test."

From Stoeckel et al.'s article, we might assume that the VLT and VST do not work well for the purpose of reading. However, this does not appear to be the case. Qian (1999) and Qian (2002) found significant correlations of .78 and .74 between the scores on Nation's (1983) version of VLT and reading comprehension. Stæhr (2008) found a significant correlation of .83 between scores on Schmitt et al.'s (2001) version of the VLT and reading comprehension. Laufer and Ravenhorst-Kalovski (2010) reported a significant correlation of .80 between VLT (Schmitt et al., 2001) scores and reading comprehension. In one study examining the relationship between scores on the VST and different types of reading comprehension questions, Chen and Liu found smaller significant correlations ranging from .35 to .49 between these variables. It should be noted that Chen and Liu only included scores on the first 10 frequency levels of Nation and Beglar's (2007) VST. Because the VST was developed and initially validated to measure knowledge of a greater number of frequency levels, it is possible that using only part of the test reduces the validity and reliability of these findings. It would be useful for future studies to examine the relationship between reading comprehension and the most recent version of the VLT (Webb et al., 2017) and complete versions of the VST (Coxhead et al., 2015; Nation & Beglar, 2007) to determine whether the results are consistent with earlier findings. It would also be useful to investigate the degree to which reading comprehension is associated with scores on different vocabulary tests. For example, research could examine the relationships between reading comprehension scores and scores on receptive tests of form-meaning connection such as the VLT and VST, tests of productive vocabulary knowledge such as Lex30 (Meara & Fitzpatrick, 2000), tests that include multiple formats (e.g., Computer Adaptive Test of Size and Strength: Aviad-Levitzky et al., 2019), and tests that measure other aspects of vocabulary knowledge such as the Word Part Levels Test (Sasao & Webb, 2017) and Guessing from Context Test (Sasao & Webb, 2018).

Perhaps Stoeckel et al. are pointing to the fact that in studies of L2 vocabulary, the scores from both tests have been provided to indicate whether participants may be able to understand the L2 input encountered in different learning conditions (e.g., Feng & Webb, 2020; Horst et al., 1998). Providing the scores of vocabulary tests that have gone through rigorous development and validation procedures in research is useful because it helps to provide a clearer picture of the vocabulary knowledge of participants. It may also reveal the degree to which prior vocabulary knowledge was a factor in learning (e.g., Peters, 2020; Webb & Chang, 2015). However, the scores of these tests should not be considered to indicate the degree to which materials are understood. There is no research that indicates that tests of vocabulary knowledge can determine comprehension. Comprehension tests are needed for that purpose.

Much of the justification used for the claims made in Stoeckel et al. is based on the extent to which tests may distinguish the lexical coverage of text. Studies of lexical coverage have used carefully controlled research designs that tend to involve replacing low-frequency words with pseudowords to determine the relationship between lexical coverage and comprehension (e.g., Hu & Nation, 2000; Van Zeeland & Schmitt, 2013). Lexical profiling studies have reported the vocabulary knowledge necessary to reach the lexical coverage of materials that may indicate that the text might be understood (e.g., Nation, 2006; Webb & Macalister, 2013). However, it is important to note that meeting lexical coverage figures associated with comprehension does not ensure that materials will be understood. In fact, knowing all the words in spoken and written input (100% lexical coverage) does not ensure that the input will be understood (Hu & Nation, 2000; Martinez & Murphey, 2011; Schmitt et al., 2011). There are many factors that affect comprehension, and while vocabulary knowledge of the words encountered in input may be the most important factor (Laufer & Sim, 1985), many other factors also play a role (Grabe, 2009). In fact, research that has investigated the degree to which the lexical profiles of materials are associated with reading comprehension indicates that there may only be a small correlation between the two variables (Webb & Paribakht, 2015). Moreover, individual differences among the vocabulary knowledge of L2 learners in a class or in a sample of participants is likely to lead to varying levels of lexical coverage and varying degrees of comprehension. Thus, while research on lexical coverage has been extremely useful in revealing the importance of vocabulary for comprehension (e.g., Hu & Nation, 2000; Laufer, 1989; Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011) and vocabulary learning targets associated with understanding of different materials (e.g., Dang & Webb, 2014; Nation, 2006; Webb & Rodgers, 2009), its value may relate primarily to theory rather than to practice. Claiming that vocabulary levels and size test scores are likely to determine reading comprehension is a misinterpretation of both the intended purposes of the tests, as well as the findings of studies of lexical coverage.

TEST FORMAT

The VST and VLT use meaning recognition formats. Stoeckel et al. argue that the tests would be improved through using a meaning recall format. Surprisingly, the only study (Laufer & Aviad-Levitzky, 2017) to explicitly investigate the relationships between meaning recall, meaning recognition, and reading comprehension using one of the tests was not discussed. Laufer and Aviad-Levitzky (2017) explicitly investigated whether meaning recognition items from the VST or meaning recall items for the same words were more closely related to reading comprehension. They found that both test formats were highly correlated with reading comprehension ($r = .91$ for meaning recall and $r = .92$ for meaning recognition). However, in contrast to Stoeckel et al., they argued that meaning recognition is a better predictor of reading comprehension than meaning recall because discriminating between distractors in meaning recognition formats may better reflect the processes that readers use to infer unfamiliar vocabulary when reading. Stoeckel et al. justified the value of meaning recall in part by reporting that it was found to have a significantly higher correlation with reading proficiency than meaning recognition in a recent study by McLean et al. (2020). However, it is important to note that McLean et al. also found that both test formats were relatively highly correlated with reading proficiency (Pearson correlations between meaning recall and meaning recognition and reading proficiency in a 30-item test were .74 and .65, respectively) and that the test items used in the study were not from either the vocabulary size or levels tests (they did however follow a similar construction procedure to VST items). Moreover, the meaning recall format examined in McLean et al. was bilingual (test takers provide the L1 meaning when cued with the L2 form) rather than monolingual (test takers provide the L2 meaning when cued with the L2 form). Readers should question the validity of comparisons of monolingual and bilingual test formats because the former can be used in both EFL and ESL

https://doi.org/10.1017/S0272263121000449 Published online by Cambridge University Press

contexts while the latter can only be used in EFL contexts in which all students share the same L1 and the teacher is also proficient in the learners' L1.

There have also been many other studies investigating the different test formats used to measure knowledge of form-meaning connection (e.g., Nakata, 2016; Smith & Karpicke, 2014). The study that has most rigorously investigated these formats for L2 learners was conducted by Laufer and Goldstein (2004). Laufer and Goldstein showed that four common test formats indicate different degrees in knowledge of form-meaning connection; form recall is the most demanding and represents the greatest strength of knowledge while meaning recognition is the least demanding and represents the smallest strength of knowledge. Thus, if we were to compare the sizes of gains in vocabulary knowledge using the four tests, we should expect the highest scores to occur for meaning recognition with scores gradually decreasing in size for form recognition, meaning recall, and form recall in that order. Stoeckel et al. argue that meaning recognition test formats overestimate knowledge. However, the same argument could be used to claim that meaning recall formats underestimate knowledge. The value of the different test formats should be the degree to which they indicate knowledge for the intended purpose. Justification for the use of a meaning recognition format used in the VLT is that it is sensitive to learner knowledge and is easy to complete and grade (Nation, 1983). Using recall formats in tests designed to measure vocabulary levels and size might have a large effect on test administration and grading. Because recall formats are more difficult, it would likely take longer to complete a test, thereby reducing test practicality. In addition, grading would not only take longer but can also present challenges with how to evaluate incorrect spelling, grammatical forms, and unexpected responses that indicate partial knowledge, which can have a negative impact on reliability. Changing to a less user-friendly test format might thus have the consequence of reducing its perceived value to teachers (i.e., reducing face validity). Nation and Webb (2011) also suggested that using meaning recognition formats in diagnostic tests such as the VLT is useful for teachers because it reveals knowledge that could be further developed. Research investigating the advantages and disadvantages of the different formats for their intended users (teachers and learners) would be useful to further clarify the value of the different test formats.

LEXICAL UNIT

Stoeckel et al. argue that the levels and size tests could be improved through changing the lexical unit from word families to lemmas. There has been a great deal of discussion recently about whether lemmas or word families are best suited for measuring receptive knowledge of vocabulary (e.g., Brown et al., 2020; Kremmel, 2016; Laufer et al., 2021; McLean, 2018; Nation & Webb, 2011). The value of using word families in tests is that by measuring knowledge of morphologically unrelated words (e.g., *care*, *know*, *run* rather than *care*, *careful*, *careless*), tests assess L2 learning of different words without evaluating knowledge of the morphological system. The value of using lemmas in tests is that by measuring knowledge of derivatives and headwords separately, tests may provide a more precise measurement of lexical knowledge (Kremmel, 2016). Evaluating knowledge using a lemma-based test might be most sensible for beginners who are unable to recognize the similarities among morphologically related words. However, one disadvantage of using lemma-based tests is that there are far more lemmas than word families to

https://doi.org/10.1017/S0272263121000449 Published online by Cambridge University Press

measure. The most frequent 1,000 and 3,000 word families are made up of 3,281 and 9,132 lemmas, respectively (Nation, 2016). Although these different lemmas will vary in frequency, the much greater number of lemmas than word families would require measuring lexical knowledge with a much greater number of test items, thereby reducing test practicality. In addition, because there are many morphologically related lemmas, there is bound to be inclusion of morphologically related items when using lemmas as the lexical unit. For example, *care*, *careful*, *carefully*; *consider*, *considerable*, *considerably*, *consideration*; *differ*, *difference*, *different*; *employ*, *employee*, *employer*, *employment*; and *important, importance, and importantly* are a few of the many morphologically related lemmas within the most frequent 2,500 lemmas of Brezina and Gablasova's (2015) New General Service List. Teachers and intermediate and advanced learners might question the value of tests that measure knowledge of morphologically related words, thereby reducing face validity. To provide a more transparent measurement of vocabulary knowledge, teachers could use the Word Part Levels Test (Sasao & Webb, 2017) to assess knowledge of the derivational system together with a test of form-meaning connection. There might also be greater benefit for research and pedagogy in developing a test designed to measure knowledge of derivations than to modify existing tests that appear to be working correctly.

NUMBER OF ITEMS

Stoeckel et al. argue that there are an insufficient number of items in tests because they will be unable to accurately reveal the degree to which learners may reach key lexical coverage figures. However, this is not the intended purpose of the tests and is likely more relevant to research than pedagogy. The question of how many items should be included in a vocabulary size or levels test is a good one although not as straightforward as was presented. In general, the greater the number of good test items, the more accurately a test should help to assess knowledge (Haladyna & Rodriguez, 2013). Creating more precise tests should be a goal so there is merit to this claim. However, aspects of practicality such as time for test administration, time for grading, and test taker fatigue should also be considered. If there is insufficient time to administer and grade a test, then it will likely have little value to teachers. It would be useful to investigate how tests with different numbers of items meet pedagogical needs.

CONCLUSION

In this commentary I have argued that the way to move forward with the development of receptive tests of vocabulary levels and size is through research involving the efficacy of those tests. The transparency and replicability of research articles is the foundation of the research process. It enables readers to evaluate research methods as well as the validity of interpretations that can be made on the basis of test results. Moreover, it allows researchers to conduct further studies to follow-up earlier findings. When there are differences of opinion, further research provides the opportunity to clarify findings.

I agree with Stoeckel et al. that it is important to try to improve existing measures of vocabulary knowledge. However, a major problem with Stoeckel et al.'s article is that it did not provide any empirical evidence indicating that size and levels tests of written receptive vocabulary knowledge are not working correctly. Because there is no research that has shown that any of the suggested changes to these tests would improve on their validity and reliability, it is premature to dismiss or reject the existing versions of these tests. Neither was there any empirical evidence provided to support any of the three conclusions that the VST or the VLT can be improved through (a) changing the test formats from meaning recognition to meaning recall, (b) increasing the number of items, and (c) changing the lexical unit from word families to lemmas. This is extremely worrying because we should expect to find evidence-based conclusions. Support should be provided by the findings of multiple studies that have (a) investigated the use of the tests to reveal their shortcomings, and (b) examined how test performance was affected through manipulating the three variables (test format, number of items, lexical unit). Schmitt et al. (2020) encouraged more rigorous vocabulary test development and validation. This should involve conducting studies of existing tests to investigate the degree to which they are working correctly for learners in a variety of L2 learning contexts. Through further validation researchers can determine whether a test is working correctly for its intended purpose, and sufficiently meeting the needs of teachers, learners, and researchers. Research can also examine the degree to which different variables can be manipulated to improve test performance. There would also be great value in creating new tests that tap into other aspects of vocabulary knowledge.

NOTES

[1] Stoeckel et al. also report that purposes of the levels and size tests are tracking vocabulary growth and goal setting. However, their conclusions were based almost entirely on whether the tests are appropriate for the purpose of reading. The tests have been used in research to track growth and suggest vocabulary learning goals (e.g., Webb & Chang, 2012). However, because this is not the purpose for which the tests were developed and initially validated, and there are no studies that have investigated the degree to which they are effective at tracking vocabulary growth and goal setting, there is no reason to reject or dismiss the tests for these purposes. Instead, there is a stronger argument to question the accuracy of the results of studies such as Webb and Chang (2012) that used one of the tests for these purposes.

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.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*, 101–118.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, *36*, 1–22.

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

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*, 121–135.

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes*, *33*, 66–76.

Feng, Y., & Webb, S. (2020). Learning vocabulary through reading, listening, and viewing: Which mode of input is most effective? *Studies in Second Language Acquisition*, *42*, 499–523.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond *A Clockwork Orange*: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*, 207–223.

https://doi.org/10.1017/S0272263121000449 Published online by Cambridge University Press

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403–430.

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions, *TESOL Quarterly*, *50*, 976–987.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters.

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.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*, 399–436.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*, 15–30.

Laufer, B., & Sim, D. D. (1985). Taking the easy way out: Non-use and misuse of clues in EFL reading. *English Teaching Forum*, *23*, 405–411.

Laufer, B., Webb, S., Yohanan, B., & Kim, S. K. (2021). How well do learners know derived words in a second language? The effect of proficiency, word frequency, and type of affix. *ITL - International Journal of Applied Linguistics. Advance online publication*. https://doi.org/10.1075/itl.200020.lau

Martinez, R., & Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly*, *45*, 267–290.

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*, 823–845.

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.

Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, *28*, 19–30.

Nakata, T. (2016). Effects of retrieval formats on second language vocabulary learning. *International Review of Applied Linguistics in Language Teaching*, *54*, 257–289.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*, 12–25.

Nation, I. S. P. (1990). *Teaching and Learning Vocabulary.* Heinle and Heinle.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*, 59–82.

Nation, I. S. P. (2008). *Teaching vocabulary: Strategies and techniques*. Heinle.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company.

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

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Heinle.

Peters, E. (2020). Factors affecting the learning of single-word items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 125–142). Routledge.

Qian, D. D. (1999). Assessing the roles of depth and breadth of knowledge in reading comprehension. *Canadian Modern Language Review*, *56*, 282–308.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*, 513–536.

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Sasao, Y., & Webb, S. (2017). The word part levels test. *Language Teaching Research*, *21*, 12–30.

Sasao, Y., & Webb, S. (2018). The guessing from context test. *ITL - International Journal of Applied Linguistics*, *169*, 115–141.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*, 26–43.

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.

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.

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, *22*, 784–802.

https://doi.org/10.1017/S0272263121000449 Published online by Cambridge University Press

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*, 139–152.

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.

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*, 457–479.

Webb, S. A., & Chang, A. C.-S. (2012). Second language vocabulary growth. *RELC Journal*, *43*, 113–126.

Webb, S., & Chang, A. C.-S. (2015). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition*, *37*, 651–675.

Webb, S., & Macalister, J. (2013). Is text written for children appropriate for L2 extensive reading? *TESOL Quarterly*, *47*, 300–322.

Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.

Webb, S., & Paribakht, T. S. (2015). What is the relationship between the lexical profile of test items and performance on a standardized English proficiency test? *English for Specific Purposes*, *38*, 34–43.

Webb, S., & Rodgers, M. P. H. (2009). The vocabulary demands of television programs. *Language Learning*, *59*, 335–366.

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics*, *168*, 34–70.

https://doi.org/10.1017/S0272263121000449 Published online by Cambridge University Press