Vocabulary Learning and Instruction
Volume 4, Issue 1, October 2015
http://vli-journal.org

Article

Second Language Vocabulary Assessment Studies: Validity Evidence and Future Directions

Rie Koizumi
Juntendo University
doi: http://dx.doi.org/10.7820/vli.v04.1.koizumi

Abstract

In this study, I review four papers by Stoeckel and Bennett; Shin; McDonald and Asaba; and McLean, Kramer, and Stewart. I will then summarize the validation evidence reported in each paper, in order to argue for the validity of the interpretations of the test scores as well as the uses of the tests considered in these four studies. This will help clarify areas of future research and strengthen the need for ties between specialists in the field of second language vocabulary assessment and general language assessment.

1 Introduction

The four studies I review are outstanding in terms of their research foci, designs, analyses, and interpretations for second language (L2) vocabulary assessment. However, the ultimate purposes of these studies are to understand how tests function and how test takers perform and to support or problematize test interpretations and uses. Therefore, it seems necessary to place the four studies in a wider framework of test validation. Validation frameworks help test developers/users understand the strengths and weaknesses of their tests (and research) and outline potential directions of study. As I argued in Koizumi (2015a), one strong candidate is an argument-based validation framework. This framework was presented systematically by Kane (1992, 2006) and is often employed when examining interpretations and uses based on test scores in L2 assessment studies (Chapelle & Voss, 2013). Kane's framework is rooted in Messick (1989) and American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA/APA/NCME, 1999), and test practitioners can implement validation more practically and systematically using this method (see Chapelle, Enright, & Jamieson, 2010, for the advantages of this approach over AERA/APA/NCME, 1999). An argument-based approach considers validity to be a matter of degree; therefore, tests can neither have perfect validity nor zero validity. Further, tests do not have high validity; instead, interpretations and uses based on test scores do. However, there are also stances for validation that oppose the argument-based approach (Markus & Borsboom, 2013; Newton & Shaw, 2014), and there are even variations among proponents of this approach (e.g., Bachman & Palmer, 2010; Kane, 2013; see Chapelle & Lee, 2013, for a review). In this review, I use Chapelle, Enright, and Jamieson (2008) as a

Please cite this article as: Koizumi, R. (2015). Second language vocabulary assessment studies: validity evidence and future directions. *Vocabulary Learning and Instruction*, *4* (1), 36–46. doi: 10.7820/vli.v04.1.koizumi

validation framework because their framework elaborates on the test constructs and test use (see Chapelle, Chung, Hegelheimer, Pendar, & Xu, 2010; Koizumi, Sakai, Ido, Ota, Hayama, Sato, & Nemoto, 2011; Purpura, Brown, & Schoonen, 2015, for examples).

There are two stages in the framework: (a) constructing an interpretive argument structure and making validity claims explicit, and (b) collecting evidence (i.e., backing) to support the interpretive argument structure and making a validity argument. Table 1 outlines the basic interpretive argument structure for L2 vocabulary tests. In this table, six inferences, warrants, and assumptions are used by way of an example. Each test developer/user can specify their framework. The number and descriptions of inferences, warrants, and assumptions in the structure can be modified (elaborated or simplified) across test contexts. For example, if the test is intended for educational purposes, all the inferences (from A to F) would come into play. If the test is to be used for research purposes, researchers can consider which inferences to include for their validation, based on the nature of the research. However, in the context of L2 vocabulary assessment, test scores are usually interpreted as reflecting a construct of L2 vocabulary, and examination seems needed for the Domain Definition inference to at least the Explanation inference (from A to D).

After those in charge of validation clarify the interpretive argument structure, in the second stage—that of collecting evidence to support the interpretive argument structure—they need to collect the evidence necessary to support the assumptions, which endorse a corresponding inference and warrant. When they finish collecting sufficient evidence, they develop a validity argument based on the interpretive argument structure and overall evidence, for example, "We argue that the interpretations and uses based on scores of this test are highly valid."

Ideally, researchers should obtain evidence from the Domain Definition inference and gradually progress to the higher inferences. However, this is difficult in practice, and they often find themselves collecting evidence separately for each inference in different phases and attempting to strengthen their validity argument gradually, along with more evidence. I will now briefly review and discuss each of the four abovementioned studies before summarizing the framework in the last section.

2 The New General Service List Test

Stoeckel and Bennett developed the New General Service List Test (NGSLT; Stoeckel & Bennett, n.d.) based on the NGSL (Browne, 2014) to measure written receptive knowledge and provide diagnostic profiles on the mastery of each level of the NGSL. The NGSLT has 100 multiple-choice format items (four options per item), all of which are in English. Stoeckel and Bennett administered the test to 238 students at Japanese universities and reported high reliability of the test and sections in each frequency band ($\alpha = .70$–$.93$); they also reported acceptably good quality of items in terms of discrimination and Rasch misfit analyses. They found a relatively high correlation ($r = .72$) between their test and the Test of English for International Communication (TOEIC). They further analyzed lexical profiles by showing the correct proportions for each band. They used 80% as the threshold for mastery and suggested that learners who fell in bands under 80% should review

Table 1. Inferences, Warrants, and Assumptions in the Interpretive Argument for an L2 Vocabulary Test

Inference	Warrant and assumptions (the latter numbered)
F. Utilization 	Students/teachers can use test results to make decisions pertaining to their learning/teaching. *Use of the test is beneficial for learning/teaching. (1) Meanings of test scores and score reports are clearly interpretable by learners and teachers. (2) Students/teachers are willing to use test results in their learning/teaching process. Students use diagnostic results to make decisions on how to study. *Students/teachers perceive this test positively. *The use of the test provides learning/teaching opportunities by offering feedback on relevant vocabulary. (3) *Students/teachers understand the degree to which they/their students have mastered the vocabulary; they also understand their own strengths and weaknesses. *The use of the test facilitates their learning/teaching process.
E. Extrapolation 	Test results are relevant to the L2 vocabulary learning context. (1) As intended, test scores are positively related to other indicators or test scores that reflect L2 skills or proficiency. (Test takers at different levels perform differently on the test, depending on their proficiency levels.)
D. Explanation 	Test scores reflect the aspects of a construct of L2 vocabulary in the L2 learning context. (1) Observed test-taking processes accord with test developers' expectations. (2) A factor structure of the entire test corresponds to what would have been predicted. (3) Means of the vocabulary level/band sections correspond to what would have been predicted. (4) As intended, test scores are positively related to the scores on tests assessing similar or different types of knowledge. (5) Test scores reflect the intended construct and do not radically overestimate or underestimate the test-takers' knowledge.
C. Generalization 	Test scores are consistent across test formats/test items. (1) The test includes a sufficient number of items and provides stable estimates of the test-takers' performance. (2) Test specifications are well defined so that parallel items can be created.

Table 1. (*Continued*)

Inference	Warrant and assumptions (the latter numbered)
B. Evaluation 	Test results provide students/teachers/test users with accurate information. (1) Item statistical characteristics are appropriate. (2) Score reports are accurate, clear, and specific.
A. Domain definition	Test items cover relevant knowledge representative of L2 vocabulary. (1) Test items cover critical vocabulary knowledge needed for L2 use. (2) A test item format elicits knowledge relevant to and representative of the target domain.

Note. Adapted from Chapelle et al. (2008) and Chapelle, Cotos, and Lee (2015), whose terms, expressions, and visual displays are also adopted. According to Chapelle et al. (2015), the Utilization inference can be divided into the Utilization and Ramification inferences; the former pertains to the usefulness of test scores and score reports, whereas the latter is related to the washback effects of tests on learning and teaching. The warrant and assumptions pertaining to the latter are denoted by *.

lexical items in these bands. By way of future research and development, they mentioned creating parallel test forms and bilingual versions of the test and developing a large pool of items calibrated on a single scale.

The greatest strength of their research lies in its analysis, shown in Table 1 of their paper. It shows how the NGSLT can provide diagnostic profiles in frequency bands, by comparing them with TOEIC and Vocabulary Size Test (VST) scores. The analysis shows that students with the same VST estimates have different profiles: Students may have mastered high-frequency levels and gradually have less knowledge in lower frequency levels (Examinees 1, 3, and 4). Others may not have mastered high levels but had a substantial degree of knowledge in lower frequency levels, which can be called "fluctuating profiles" (as termed by Stoeckel and Bennett in their presentation; Examinees 2, 5, and 6). The patterns of learning displayed by the latter type of learner with fluctuating profiles are not usually expected. The results also show that the latter type tends to have lower TOEIC scores and, therefore, lower L2 proficiency. A similar trend was reported in Ota, Kanatani, Kosuge, and Hidai (2003). They reported that some junior high school students belonged to the latter type and tended to have lower proficiency and that one student developed their speaking ability more slowly than students not in the latter type. These may suggest that unstable knowledge without the firm basis of high-frequency vocabulary can inhibit the smooth development of L2 proficiency. An investigation into the performance and development paths of these learners would be an important future research inquiry. Recently, researchers have started to examine the types of profiles for four skills (reading, listening, writing, and speaking) and the underlying causes of uneven profiles (Ginther, Yan, & Potts, 2015; Huhta, Alderson, Nieminen, & Ullakonoja, 2015; Koizumi, 2015b). I believe that expected and unexpected patterns of vocabulary profiles could be strong predictors of uneven profiles and this type of inquiry could contribute to uncovering the nature of language proficiency.

For the NGSLT to be useful for learners and teachers, researchers may need to consider two points. First, if researchers are to claim the superiority of the

NGSLT over the VST in terms of diagnostic functions, they would need to present subscores for the VST for each frequency level so that a comparison can be made of the sensitivity of diagnosis between the two tests. If the NGSLT could provide clearer patterns or more useful information, then more people would use it. Second, learners and teachers will need a template of NGSLT score reports. Otherwise, some test users would only count the total number of correct items and not inspect the lexical profiles; this could result in misuse of the test. The effectiveness of score reports for learners and teachers is another possible research topic that can be pursued (see Doe, 2015; Jang, Dunlop, Park, & van der Boom, 2015; Sawaki & Koizumi, 2015).

3 Psychometric Properties of Word Association Tests with Regard to Adolescent EFL Learners

Shin developed and analyzed a modified version of the Word Association Test (WAT) with the aim of assessing vocabulary depth in adolescent learners of English in Korea. The original format (Read, 1993) requires test takers to select four answers (paradigmatic, syntagmatic, and/or analytic) out of eight options. Shin's format requires test takers to select one answer (paradigmatic or syntagmatic) out of six. She administered the test to 121 sixth-grade elementary school students. She reported high reliability for the entire test ($\alpha = .83$) when she deleted items with low discrimination. She also reported a moderate correlation between the WAT scores and the scores of paradigmatic and syntagmatic relationships with a reading test ($r = .55$–$.63$), and a strong correlation ($r = .83$) between two factors of paradigmatic and syntagmatic relations in a factor structure, which shows a moderate fit to the data (e.g., root mean square error of approximation $= 0.09$).

Her study is valuable because a test with a modified format—especially in a new population of test takers—requires new validation, even though some relevant evidence from previous studies can be utilized. While she obtained positive evidence in terms of internal consistency, moderate relationships with a reading measure, and a factor structure, more evidence or explanations will be necessary in terms of her rationale, for example, for changing the format and for the organization of all the test items, including why the test has more items in paradigmatic relations than syntagmatic ones (i.e., 7 vs. 17 items). Shin may also want to compare her results to those of previous studies that used the original WAT (e.g., Read, 1993; Schoonen & Verhallen, 2008). For example, Batty (2012) administered the original WAT with 145 items to 530 learners of English at a Japanese university. He reported that the best fitting model was a bifactor model wherein the primary factor underlying all the items was a vocabulary factor and wherein the two factors of synonym (paradigmatic) and collocate (syntagmatic) were uncorrelated. Shin's and Batty's differing models may suggest the effects on the factor structure of (a) a format change, (b) a change in the number of items, (c) and a change in different test-taker populations; herein lies the scope for future avenues of research.

4 ''I Don't Know'' Use and Guessing on the Bilingual Japanese VST: A Preliminary Report

McDonald and Asaba examined the effects of including the "I don't know" option on test scores and analyzed how test takers select their responses when they are unsure of the answers. They used the bilingual version of the VST with up to 14,000-word family levels, with a stem in L2 English, four options in L1 Japanese, and the extra "I don't know" option. McDonald and Asaba interviewed four of 308 Japanese university students who took the VST. They reported an increased use of the "I don't know" option in lower frequency levels. They also noticed that when test takers do not know the answers, they make both uninformed and informed guesses, the latter of which are based on "true partial knowledge," "false partial knowledge," or "test strategy use." The students used all the types of uninformed and informed guesses, except in the case of one higher proficiency student who did not use uninformed guesses or guesses based on false partial knowledge. McDonald and Asaba computed the scores based on five scenarios: (a) "scores without guesses," (b) "scores with true partial knowledge-informed guesses," (c) "scores with all partial knowledge-informed guesses," (d) "scores with all informed guesses," and (e) "scores with all guesses." They showed that the scores varied substantially from (a) to (e), especially in the case of the lower proficiency students. For example, the lowest proficiency student had a difference of 30 points, which can be translated into a difference of 3,100-word family estimates. Of the five types of scores, McDonald and Asaba argued that scores with true partial knowledge-informed guesses (i.e., b) were most relevant to the construct. Therefore, the VST's overestimation of the test-takers' vocabulary size could threaten validity, and because of this, we should interpret the results with caution.

McDonald and Asaba's qualitative investigation using the bilingual version of the VST is very beneficial in terms of understanding students' test-taking processes and the drawbacks of the multiple-choice format. Based on the abovementioned five types of scores, one interesting direction for further research would be to devise strategies for producing confidence intervals (CIs), that is, the range of vocabulary size estimates that a single test taker can obtain. When using scores without guesses (i.e., a), the lower CI limit would indicate the minimum vocabulary size estimate when the VST includes the "I don't know" option, or if the test instructions urge test takers to avoid making random guesses, or if the instructions warn students that they will be penalized for selecting an incorrect option. The upper CI limit would indicate the maximum vocabulary size estimate obtained when test takers are encouraged to make guesses (i.e., e). These CIs would include scores with true partial knowledge-informed guesses (i.e., b), which will be gained through knowledge of the target word as well as guesses based on correct partial knowledge. A two-stage testing style may be needed to derive CIs for a paper-based test format, without conducting interviews: The first stage should contain the "I don't know" option or an instruction discouraging test takers from guessing; in the second stage, test takers should be encouraged to guess. The use of computers would enable easier administration wherein, in the second stage, only the items for which test takers selected "I don't know" appear. The presentation of CIs pertaining to vocabulary size estimates will improve test-users' understanding of their

overestimation risks and discourage them to take the results at face value. However, I admit that these procedures should be undertaken after consideration of the numerous other factors that come into play, and they require careful investigation before they can be implemented. Examples of the other factors that come into play are individual differences in whether or not test takers select "I don't know," the nature of the test instructions, and the stakes of the test (see Zhang, 2013, for details). Further, since guessing is not discouraged in the original VST (Nation, 2012), the use of the abovementioned procedures might alter the originally intended construct of the VST.

5 An Empirical Examination of the Effect of Guessing on VST Scores

McLean, Kramer, and Stewart examined the degree to which test takers guess before selecting answers in the VST. They administered the 1,000- to 8,000-word family frequency levels of the original VST to 3,373 students at Japanese universities, who had a wide range of *Hensachi* or T-scores. After excluding items with low discrimination, they used the Rasch and three-parameter logistic (3PL) models to investigate the effects of guessing. They showed that even low-ability learners guessed approximately 25% of the items correctly, and they obtained 41.7–85.5% of the correct scores in the 8,000-word family frequency level by using guessing strategies. They also reported a better fit of the 3PL model compared to the Rasch model; considering this, I wonder why they did not compare the two-parameter logistic (2PL) model with the 3PL model, because a gradual increase in parameters from the Rasch to the 2PL model and from the 2PL to the 3PL model (rather than from the Rasch to the 3PL model) would indicate differences in model fit more clearly and enhance our understanding of the underlying processes of guessing.

Similar to McDonald and Asaba, McLean et al. focused on the effects of guessing on the overestimation of vocabulary size, but from a quantitative perspective. Although the two studies used different versions of the VST, respectively—bilingual and monolingual—the results provide evidence that is rather negative: the VST has consistent overestimation issues.

Similar to the case with McDonald and Asaba, a possible direction for further study would be to devise methods of obtaining CIs by taking into account guessing percentages. Another direction could be an empirical examination into how many levels beyond test-takers' existing vocabulary levels are acceptable for test takers to take the VST, using the current data. McLean et al. showed that in the 8,000-word family frequency level, even the highest proficiency group (with a *Hensachi* score of 73) used guessing strategies at as high as 41.7%. For practical purposes, they could conduct similar analyses in the 1,000- to 7,000-word levels according to test-takers' size estimate groups as well as *Hensachi* groups, possibly by using different full marks across levels although the guessing percentages derived may not be very precise because McLean et al. omitted some 6,000 and 7,000 level items due to the low discrimination. If they set a permissible percentage for guessing (e.g., 25% or 30%), they could decide whether or not to endorse

Table 2. Evidence Provided by the Four Studies in Relation to the Interpretive Argument

Inference/Test	NGSLT (Stoeckel and Bennett)	Modified WAT (Shin)	Bilingual version of the VST (McDonald and Asaba)	Original VST (McLean et al.)
F. Utilization	*	*	*	*
E. Extrapolation	1. *r* with TOEIC	*	*	1. [B] Differences between proficiency groups
D. Explanation	*	2. Factor structure; 4. *r* with a reading test	5. [N] Large effects of guessing on test processes and scores	5. [N] Large effects of guessing on test scores 2. Item dimensionality; 3. [B] Increase in difficulty level
C. Generalization	1. Reliability (2. Creating parallel forms, bilingual versions, and an item bank)	1. Reliability	*	1. [B] Reliability; measurement invariance
B. Evaluation	1. Item quality	1. Item discrimination	*	1. [B] Item quality
A. Domain Definition	1. Content representativeness supported by the use of NGWL	*	1. Content representativeness supported by the translation of VST	1. [B] Content representativeness

Note. () = Mentioned as future research in the paper. * = Urgent need for examination. [B] = Evidence from Beglar (2010). [N] = Negative evidence. This table assumes an interpretive argument structure wherein the four studies intend to use the vocabulary tests for pedagogical purposes; however, other simplified structures such as those without the Utilization inference are also possible.

Beglar's (2010) suggestion that learners should take the VST with up to two levels higher than their vocabulary size.

6 Discussion and Conclusion

The four qualitative and quantitative studies reviewed here deserve substantial credit for their contributions to our existing knowledge on L2 vocabulary assessment studies and for providing positive and negative evidence for test validation. One issue for improvement is to clearly show how their studies fill the research gaps that exist in previous ones. This would further clarify the significance of their papers and substantially enhance contributions to the field.

Table 2 summarizes evidence, as reported by each paper, as well as puts forth areas for future examination, according to each test. I have also added evidence derived from Beglar (2010) because of its comprehensive coverage of the various aspects of validity. An examination of validation involves the entire process of examining test scores, interpretations, and uses from various perspectives; consequently, this task is often beyond the scope of a single study. Therefore, we need to accumulate evidence in relation to a validation framework so that other studies can refer to the evidence in the framework and observe the progress of validation while considering their individual research contexts. While I believe that L2 vocabulary researchers will benefit from using the argument-based approach to validity as a frame of reference, they may feel overwhelmed by the numerous detailed processes involved. However, this task can be shared with language assessment experts, giving rise to mutually beneficial research and practices (Koizumi, 2015a).

Acknowledgements

I am deeply indebted to Alastair Graham-Marr of Abax Publishing for his assistance and Yo In'nami for his invaluable comments. This work was partially supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research (C) [grant number 26370737].

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA/APA/NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford, U.K.: Oxford University Press.

Batty, A. (2012). Identifying dimensions of vocabulary knowledge in the Word Associates Test. *Vocabulary Learning and Instruction*, *1*, 70–77. Retrieved from http://vli-journal.org/wp/vli-v01-1-2187-2759/

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

Browne, C. (2014). A New General Service List: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction*, 3(2), 1–10. Retrieved from http://dx.doi.org/10.7820/vli.v03.1.browne

Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing*, *27*, 443–470. doi:10.1177/0265532210367633

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Diagnostic assessment with automated writing evaluation: A look at validity arguments for new classroom assessments. *Language Testing*, *32*, 385–405. doi:10.1177/0265532214565386

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3–13. doi:10.1111/j.1745-3992.2009.00165.x

Chapelle, C., & Lee, H.-W. (2013, July). *What is argument-based validation?* Paper presented at the 35th Language Testing Research Colloquium, KCCI (Korea Chamber of Commerce and Industry) Building, Seoul, Korea.

Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In A. Kunnan (Ed.), *The companion to language assessment* (Vol. III, pp. 1079–1097). Hoboken, NJ: Wiley-Blackwell.

Doe, C. (2015). Student interpretations of diagnostic feedback. *Language Assessment Quarterly*, *12*, *110–135*. doi:10.1080/15434303.2014.1002925

Ginther, A., Yan, X., & Potts, J. (2015, March). *The relationship between TOEFL and GPA: The case of Chinese students*. Paper presented at the 37th Language Testing Research Colloquium, Eaton Chelsea Toronto, ON, Canada.

Huhta, A., Alderson, J. C., Nieminen, L., & Ullakonoja, R. (2015, March). *Diagnostic profiling of foreign language readers and writers—Exploring the usefulness of Latent Profile Analysis in diagnostic assessment research*. Paper presented at the 37th Language Testing Research Colloquium, Eaton Chelsea Toronto, ON, Canada.

Jang, E. E., Dunlop, M., Park, G., & van der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing*, *32*, 359–383. doi:10.1177/0265532215570924

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *12*, 527–535. Retrieved from http://dx.doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. doi:10.1111/jedm.12000

Koizumi, R. (2015a). Book review: Vocabulary knowledge: Human ratings and automated measures. *Language Testing*, *32*, 124–126. doi:10.1177/0265532214541233

Koizumi, R. (2015b). Factor structure and four-skill profiles of the TOEIC® tests among Japanese university learners of English. *ARELE (Annual Review of English Language Education in Japan)*, *26*, 109–124.

Koizumi, R.,Sakai, H., Ido, T., Ota, H., Hayama, M., Sato, M., & Nemoto, A. (2011). Toward validity argument for test interpretation and use based on scores of a diagnostic grammar test for Japanese learners of English. *Japanese Journal for Research on Testing*, *7*, 99–119. Retrieved from http://www7b.biglobe.ne.jp/~koizumi/Koizumi_research.html

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: National Council on Measurement in Education/American Council on Education.

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

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Thousand Oaks, CA: Sage.

Ota, H., Kanatani, K., Kosuge, A., & Hidai, S. (2003). *Eigo ryoku wa donoyoni nobite yukuka* [How English ability is developed: Exploring English acquisition process of junior high school students]. Tokyo: Taishukan.

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics. *Language Learning*, *65*(Suppl. 1, Special issue: Currents in language learning series: Improving and extending quantitative reasoning in second language research), 37–75. doi:10.1111/lang.12112

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*, 355–371. doi:10.1177/026553229301000308

Sawaki, Y., & Koizumi, R. (2015, March). *Japanese students' and teachers' perception and use of score reports for two large-scale EFL tests*. Paper presented at the 37th Language Testing Research Colloquium, Eaton Chelsea Toronto, ON, Canada.

Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing*, *25*, 211–236. doi:10.1177/0265532207086782

Stoeckel, T., & Bennett, P. (n.d.). *The New General Service List Test (NGLST)*. Retrieved from http://www.newgeneralservicelist.org/ngsl-levels-test/

Zhang, X. (2013). The *I don't know* option in the Vocabulary Size Test. *TESOL Quarterly*, *47*, 790–811. doi:10.1002/tesq.98