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

Article

An Empirical Examination of the Effect of Guessing on Vocabulary Size Test Scores

Stuart McLean[a], Brandon Kramer[b] and Jeffrey Stewart[c]
[a]*Kansai University;* [b]*Momoyama University;* [c]*Kyushu Sangyo University*
doi: http://dx.doi.org/10.7820/vli.v04.1.mclean.et.al

Abstract

The Vocabulary Size Test (VST) was created to provide a reliable estimate of a second language learner's written receptive vocabulary size, measuring from the most frequent fourteen 1,000 word families of the spoken subsection of the British National Corpus. While some have recommended that users should limit the amount of the test taken to only slightly above a student's level, others argue that learners should take every level of the test. However, this raises concerns that correct responses on lower frequency levels could largely be attributed to guesses rather than vocabulary knowledge. In this paper we analyze a data set of 3,373 Japanese university students' responses to the first eight levels of the original VST under the 3PL model, in order to determine the minimum expected score on the test for learners of low ability, examine the proportion of low-level students' scores on the lowest frequency level tested that can be attributed to guessing under the 3PL model, and conduct a model fit comparison to determine whether the 3PL model offers a significantly better description of the data than the Rasch model. The results indicate that a substantial portion of lower level learners' scores on items testing low-frequency words can be attributed to guessing and support the position that students should not sit every level of the test. The authors recommend using the results of the 3PL analysis in order to determine which sections of the test learners of different proficiency levels should sit.

1 Background

1.1 Introduction

The Vocabulary Size Test (VST) was created to provide a reliable estimate of a second language learner's written receptive vocabulary size, measuring from the most frequent fourteen 1,000 word families of the spoken subsection of the British National Corpus (Nation & Beglar, 2007). As the original VST tests 140 of these 14,000 most common words, vocabulary estimates are derived by multiplying the raw score on the test by 100, under the assumption that each correct answer can be considered equivalent to knowledge of each tested word.

Nation (2012), Karami (2012), Nguyen and Nation (2011), and Coxhead, Nation, and Sim (2014) argue that learners should attempt every level of the test, as

Please cite this article as: McLean et al. (2015). An empirical examination of the effect of guessing on vocabulary size test scores. *Vocabulary Learning and Instruction*, *4* (1), 26–35. doi: 10.7820/vli.v04.1.mclean.et.al

learners may know some low-frequency words in a "Slumdog Millionaire" effect. Beglar (2010) and Elgort (2013), however, recommend that learners should not take more than two levels above their ability. In support of this position, Stewart (2014) further argues that we should limit the levels of the test that are administered because use of a four-option multiple-choice format implies a 25% chance that a learner can guess correctly. Therefore, correct answers by lower level learners at more challenging levels of the test could likely be attributed to random guessing. Furthermore, a new version of the VST (Coxhead, Nation, and Sim, 2014) samples the first 20,000 words, initially with 10 items and currently only 5 items per 1,000-word band. This increases the likelihood that a sizeable portion of learner size estimates could be attributed to guessing.

To what degree do guesses by lower level learners on multiple-choice items testing less frequent words indicate knowledge of them, and to what degree do they constitute luck? Stewart (2014) suggested that the three-parameter logistic (3PL) model (Birnbaum, 1968), an item response model similar to the Rasch model, could be useful in helping to empirically determine the degree to which random guesses inflate scores. Although it does not account for guessing behaviors by higher proficiency learners, the 3PL model estimates the probability that low-level learners can correctly guess more difficult items entirely by chance. In this paper, we analyze a data set of 3,373 Japanese university students' responses to the first eight levels of the original VST under the 3PL model. The purpose of this paper is to determine the minimum expected score on the test for learners of low ability, examine the proportion of students' scores on the lowest frequency level tested that can be attributed to guessing under the 3PL model, and conduct a model fit comparison to determine whether the 3PL model offers a significantly better description of the data than the Rasch model.

1.2 The Rasch Model and the 3PL Model: A Brief Primer

In order to explain the method of analysis used in this paper, this section briefly compares and contrasts the Rasch model and the 3PL model. It should be noted that the 3PL model is used in this study and compared to the Rasch model for analytical purposes, rather than as an endorsement of test scoring or vocabulary size estimation under the 3PL model. For further information on both models, please consult De Ayala (2009).

The VST was validated by Beglar (2010) using the Rasch model (Rasch, 1960). Under the Rasch model, the likelihood that a given test taker will correctly answer a given test item is modeled as a logistic function of the difference between the learner's ability level (the person parameter) and the difficulty of the test item (the item parameter). This relationship is depicted in Figure 1. The horizontal axis indicates learner ability expressed in "logits." The vertical axis indicates the probability of a correct answer given the learner's ability level. Finally, the "s"-shaped line depicts the Rasch model's logistic function, which describes the relationship between ability and the probability of a correct answer on a given item.

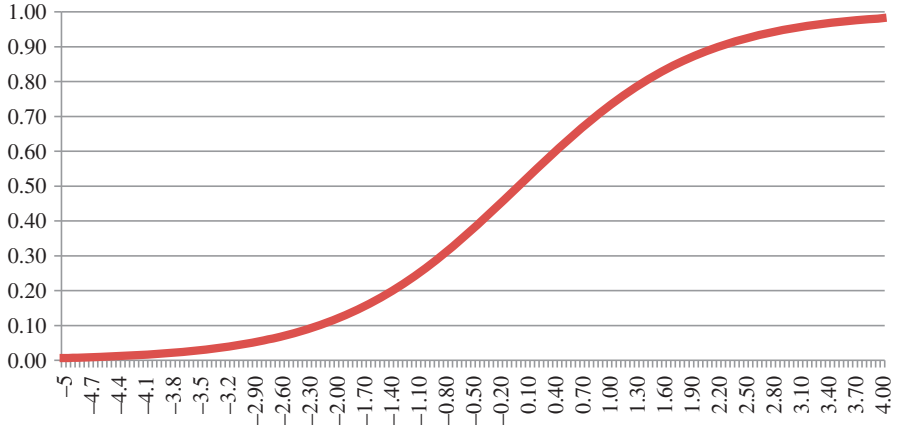

Figure 1. An Example of a Test Item Fit to the Rasch Model.

The Rasch model assumes minimal guessing; given a low enough level of ability, a learners' probability of answering given items should approach 0. However, this may not be the case with multiple-choice items. In cases where lower level learners do not know the answer to test items, they may guess the answer from the available options. Because of this, the probability of a correct guess may never actually fall to near 0. The 3PL item response model accounts for this with a ''pseudo-guessing'' parameter, which models minimum probabilities of correct answers that can be attributed to guessing, and, due to the use of a multiple-choice format, no longer drop with lower levels of ability. Figure 2 depicts a four-option multiple-choice test item, in which the minimum probability of a correct answer never falls below 0.25, even for learners of very low ability.

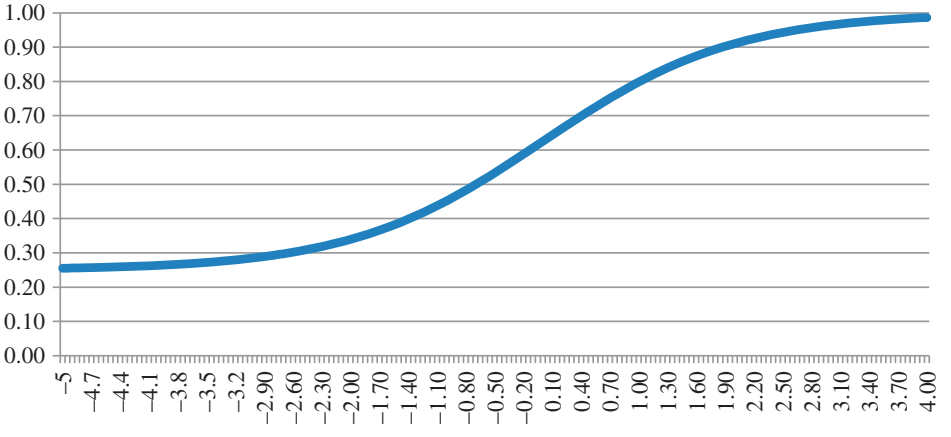

Figure 2. An Example of a Test Item Fit to the 3PL Model.

Although we typically use Rasch models in our work and research, we believe the 3PL model has useful analytical properties in regards to the validity of vocabulary size estimates made using the VST. As vocabulary size estimates derived from the test scores are based on the assumption that answering a multiple-choice question correctly implies knowledge of the tested word, a concern regarding the VST is that a substantial component of test scores – and therefore of its suggested vocabulary size estimates – could be attributable to guessing that is unrelated to proficiency, thereby inflating estimates. As the 3PL models a "flat" guessing rate for very low-level learners that is no longer related to changes in ability, if the VST has better fit to the 3PL model than to the Rasch model, it could help determine whether, and at what point, raw scores on the test cease to decrease with lower levels of proficiency, indicating the minimum vocabulary size estimates the test will provide for low-level learners.

2 Methodology

2.1 Instrument

This study utilized the first eight 1,000-word levels of Nation and Beglar's (2007) VST.

2.2 Participants

This study used a cross-sectional design with data from 3,373 Japanese university students collected through a snowball sampling approach, all of whom attempted to answer every question on the test as per Nation's (2012) instructions. The participants were from a range of Japanese universities representing a full range of abilities, as determined by examination of *hensachi* rank scores. A *hensachi* is a score assigned to individual students or school departments based on student performance on a national test standardized to the national mean across five subjects. A *hensachi* of 50 represents the mean, where one standard deviation above or below is represented by 60 or 40, respectively. The scores can range from 20 to 80, but 95.4% of all university departments fall between 30 and 70 (Newfields, 2006). Department *hensachi* scores were obtained from Benesse, a large testing company in Japan <http://shinken.zemi.ne.jp/hensachi>.

McLean, Hogg, and Kramer (2014) demonstrated that the mean *hensachi* of a participant's department was a good predictor of participants' ability, with a significant difference in VST scores between participants from three *hensachi* groups: $\geq$61, 51–60, and $\leq$50 ($F(2, 3{,}424) = 1{,}383.14$, $p < .001$, $\eta^2 = 45$). Further, participants' department *hensachi* scores correlated strongly with VST scores ($r = .73$, $p < .001$). Figure 3 shows that while the most common department *hensachi* of participants in the present study was 49 ($n = 1{,}030$), the majority of the data came from participants with average or above average department *hensachi* of 50 or above ($n = 1{,}940$), with an overall mean department *hensachi* score of 53.2 for all participants, which is somewhat above the national mean of 50.

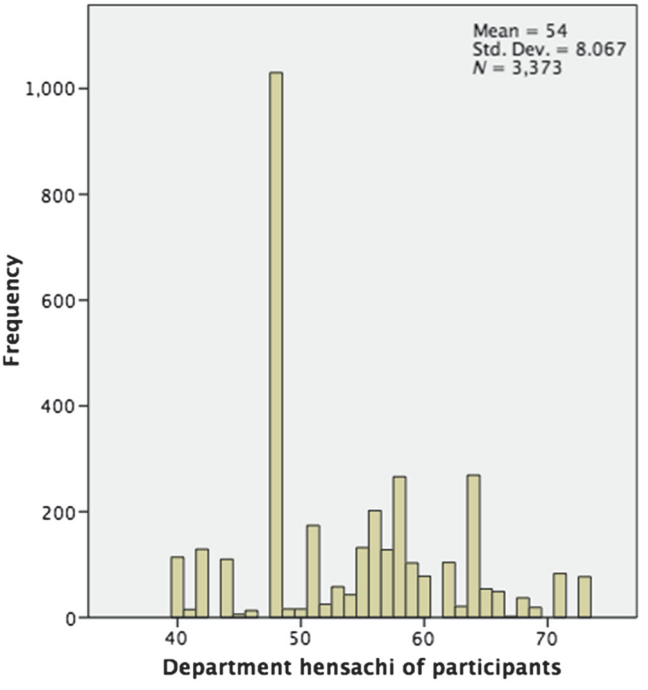

Figure 3. Histogram of Participants' Department *Hensachi*.

3 Data Analysis

3.1 Comparison of Models

The data set was analyzed with the software program IRTPro 2.1 (Cai, Thissen, & du Toit, 2011) under the Rasch model (from a technical standpoint, a two-parameter logistic model with slopes constrained to 1) and the 3PL model. However, due to the model failing to converge initially, it was necessary to remove the seven poorest items. These seven items had near-zero point-biserial correlations and negative slopes under a preliminary analysis using the two-parameter logistic model, meaning less able participants were *more* likely to answer these items correctly than more able participants, in violation of the model's assumption that guessing indicates a minimum probability of a correct answer, rather than a maximum. These items were item 10 (*basis*), 14 (*nil*), 29 (*rove*), 55 (*threshold*), 58 (*cavalier*), 65 (*bristle*), and 68 (*gimmick*). Two of these items (*basis* and *rove*) were reported by Beglar (2010) to perform poorly and underfit the Rasch model as well (standardized infit values $> + 2.00$), and Rasch analyses presented by McLean (2013) and McLean and Kramer (2014) found that all seven of the above items underfit the Rasch model (it should be noted that this study utilized the original version of the VST, and the version of the VST presently available online at Nation's website <http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-14000.pdf> includes an edited version of item 10, *basis*).

Table 1. Mean Raw Scores and Ability Estimates for Four Schools

Hensachi	Mean Score	3PL Mean Ability	Rasch Mean Ability
40	24	−1.48	−1.00
48	31	−0.66	−0.44
51	36	−0.07	−0.03
73	52	1.35	1.13

Note: The mean ability estimate for both models is 0.

Mean raw scores and scale score estimates under both models for schools with four separate *hensachi* ranks are listed in Table 1.

It is possible to graph the results of the Rasch model fit to determine what raw test score a learner at a given ability level would receive. In the Test Characteristic Curve depicted in Figure 4, the horizontal axis indicates student ability under the Rasch model, and the vertical axis indicates the raw score a learner at that ability level would be expected to receive on the test. Under this model, learners with lower ability levels than the university students tested in this study (for example, junior high or elementary school students) would be expected to receive lower test scores, with learners of very low ability expected to know almost none of the words, and therefore receive a score near 0.

However, the 3PL model's Test Characteristic Curve tells a somewhat different story (Figure 5). Under this model, a learner of very low ability would still have an expected score of 18 out of 73 or approximately 25%. As a sum score–scale score conversion table generated by IRTPro indicates that −3.349 is the lowest level of ability that the analyzed test form is capable of estimating, under the 3PL model this would appear to represent the practical minimum raw score estimate for the test, regardless of the ability level of the learner.

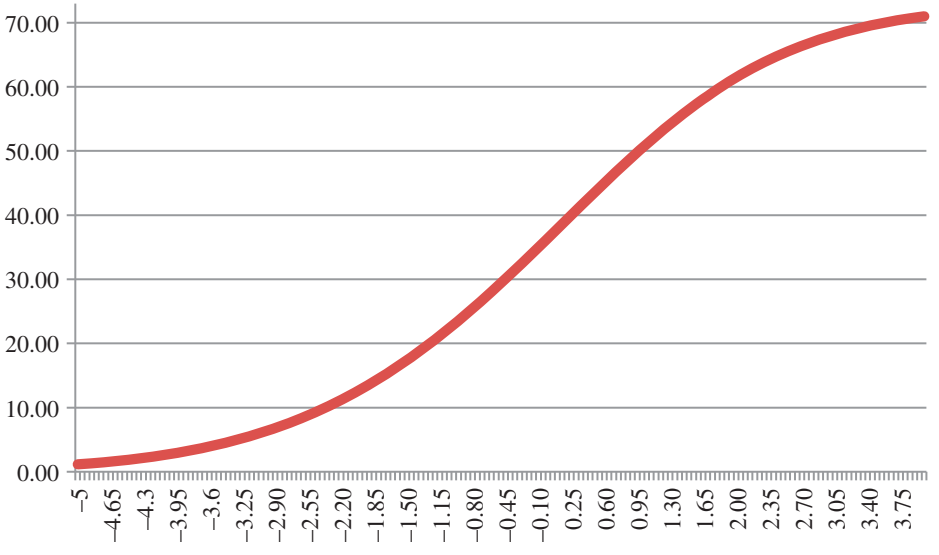

Figure 4. Test Characteristic Curve of the VST Data Set ($k = 73$) Fit to the Rasch Model.

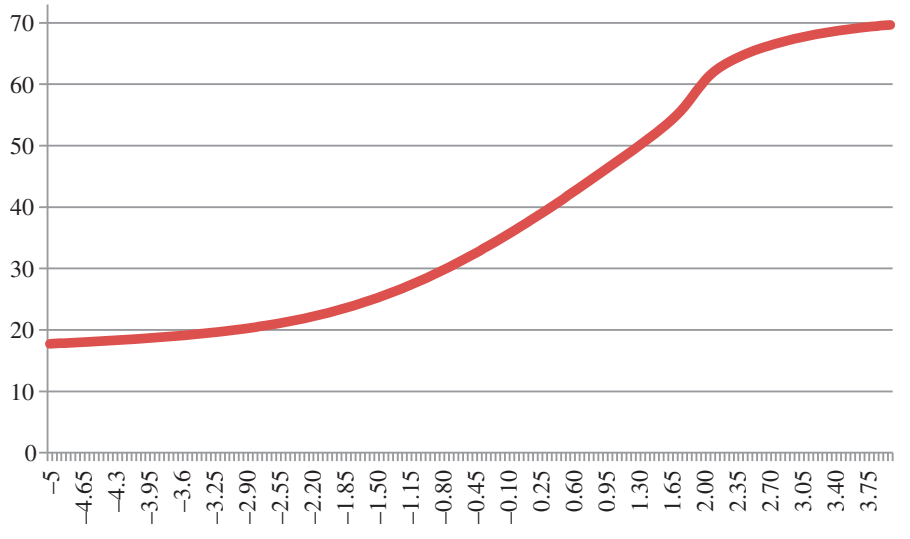

Figure 5. Test Characteristic Curve of the Data Set ($k = 73$) Fit to the 3PL Model.

3.2 Effect of Guessing on Items Testing Lower Frequency Words

A drawback of the current study is that only VST items testing the first eight 1,000-word levels were examined, and guessing would presumably be most prevalent with the lower frequency bands of the test. A further drawback is that poor items were removed from the 6k and 7k levels of the test prior to analysis, meaning it is not possible to estimate scores out of 10 for these levels. However, it is possible to examine students' scores out of 10 on the lowest frequency band tested, the 8k level, and calculate the proportion of their scores that could be attributed to uninformed guesses under the 3PL model.

As can be seen in Table 2, the mean score on the 8k level for departments with a *hensachi* rank score of 40 is 2.71 out of 10, just slightly above the estimated minimum score on this level for low-level learners. Given this, it appears that approximately 85.5% (2.32 / 2.71 $\times$ 100) of the mean raw score on this frequency band for students at this proficiency level (and therefore 85.5% of the 271 words in the band they would be estimated to know) could be attributed to uninformed guesses under the 3PL model. Even for the highest level departments tested, under the 3PL model, 41.7% of raw scores at this level can be attributed to guessing

Table 2. Mean Raw Scores Out of 10 on 8k Level of Test by Department *Hensachi*

Hensachi	Score on 8k Level	Minimum estimated score (3PL)	% of score below guess threshold
40	2.71	2.32	85.5
48	3.11	2.32	74.6
51	3.58	2.32	64.8
73	5.57	2.32	41.7

unrelated to proficiency. Although untested in this study, one can only assume that guesses would comprise even higher proportions of scores at even lower frequency levels (9k–20k levels).

3.3 Model Fit

The two item response models present differing descriptions of the VST in relation to the effect of uninformed guesses. Which item response model comes closer to describing the nature of the data? When assessing model fit, it should be noted that statistical models with more parameters (such as the 3PL) will nearly always demonstrate superior model fit to at least a negligible degree. However, if the difference is slight, the more complex model may simply overfit the data in a way that would not be generalizable to another data set (Zucchini, 2000). For this reason, IRT model fit is often assessed with fit statistics for nested models such as the Akaike Information Criterion (AIC; Akaike, 1998) or the Bayesian Information Criterion (BIC; Schwarz, 1978), which "penalize" additional parameters (see Table 3). However, under both statistics, values remain lower for the 3PL model, indicating superior fit. Given this, it appears the 3PL model provides a more accurate description of the test data.

Table 3. Model Fit Statistics

Statistics based on loglikelihood	Rasch	3PL
−2loglikelihood	275036.58	270066.76
AIC	275182.58	270504.76
BIC	275629.6	271845.82

4 Conclusion

Under the 3PL model, which had superior fit to the data, very low-level learners would be expected to receive a score of approximately 2.32 out of 10 due to guessing on the lowest frequency band of the test examined (the 8k level). As students in departments with a *hensachi* near the national mean of 50 had a mean score of 3.58 out of 10 on this word level, it seems that while average students' scores out of 10 on this word level are above chance, the bulk of their scores could be attributed to guessing that is unrelated to vocabulary knowledge.

This study has a number of limitations. Most importantly, the most difficult six levels of the original test were not taken, making the 8K level the lowest frequency word level students were tested on. Furthermore, seven items with very low point-biserial correlations could not be analyzed, and the overall mean department *hensachi* score for the sample was 53, which indicates the tested sample had a slightly higher proficiency level than the national average. However, despite these shortcomings, for most students guessing unrelated to proficiency appeared to have a greater effect on test scores on the lowest frequency band than proficiency did. Given these findings, it is difficult to endorse the position that the entire test be

given to students of all proficiency levels, and that they be encouraged to guess on items they do not believe they know the answers to. These results support the recommendation by Beglar (2010) and Elgort (2013) that learners not sit levels of the test well above their ability.

We suggest that the results of a 3PL analysis can be of use in determining precisely which sections of the VST learners of various proficiency levels should sit, as it can indicate when test items are too difficult to provide information about a given population. However, in cases where teachers wish to use a vocabulary test for pedagogical purposes, we recommend that rather than using a size test such as the VST, teachers use levels tests such as the Listening Vocabulary Levels Test (McLean, Kramer, & Beglar, 2015), parallel reading levels tests, or the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham, 2001), which, due to the typically larger numbers of items for each frequency band, can provide them with richer detail about the number of pedagogically relevant, higher frequency words that students know.

Acknowledgement

This research was supported in part by a grant from the Japan Society for the Promotion of Science (No. 24720278).

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected papers of Hirotugu Akaike* (pp. 199–213). New York, NY: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4612-1694-0_15

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

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison.

Cai, L., Thissen, D., & du Toit, S. (2011). *IRTPRO 2.1 for windows*. Chicago, IL: Scientific Software International.

Coxhead, A., Nation, P., & Sim, D. (2014). Creating and trialling six forms of the Vocabulary Size Test. *TESOLANZ Journal*, *22*, 13–26.

De Ayala, R. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*, 253–272. doi:10.1177/0265532212459028

Karami, H. (2012). The development and validation of a bilingual version of the Vocabulary Size Test. *RELC Journal*, *43*(1), 53–67. doi:10.1177/0033688212439359

McLean, S. (2013, December). *Investigating university students' vocabulary sizes and the VST*. Paper presented at the Vocab@Vic Conference, Victoria University Wellington, New Zealand.

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the Vocabulary Size Test. *Vocabulary Learning and Instruction*, *3*(2), 47–55. doi:10.7820/vli.v03.2.mclean.et.al

McLean, S., & Kramer, B. (2014, September). *Investigating university students' vocabulary sizes and the VST*. Paper presented at 18th Annual Conference of the Japan Language Testing Association, Ritsumeikan, Japan.

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*. Advance online publication. doi:10.1177/1362168814567889

Nation, I. S. P. (2012, October 23). *The Vocabulary Size Test*. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13. Retrieved from jalt-publications.org/files/pdf/the_language_teacher/07_2007tlt.pdf

Newfields, T. (2006). Assessment literacy self-study quiz #1 [Suggested answers]. *Shiken*, *10*(2), 25–32. Retrieved from jalt.org/test/SSA1.htm

Nguyen, L. T. C., & Nation, P. (2011). A bilingual Vocabulary Size Test of English for Vietnamese learners. *RELC Journal*, *42*(1), 86–99. doi:10.1177/0033688210390264

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.* Copenhagen: Danish Institute for Educational Research.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing*, *18*(1), 55–88. doi:10.1177/026553220101800103

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464. doi:10.1214/aos/1176344136

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*, 271–282. doi:10.1080/15434303.2014.922977

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, *44*(1), 41–61. doi:10.1006/jmps.1999.1276