Vocabulary Learning and Instruction
Volume 3, Issue 2, September 2014
http://vli-journal.org

Article

Estimating Learners' Vocabulary Size under Item Response Theory

Aaron Gibson and Jeffrey Stewart
Kyushu Sangyo University
doi: http://dx.doi.org/10.7820/vli.v03.2.gibson.stewart

Abstract

Perhaps the most qualitatively interpretable vocabulary test score is an estimate of the total number of words the learner knows in the tested domain, such as a frequency word list, or vocabulary taught as part of a course curriculum. In cases where it is not possible to test the entire domain word-for-word, vocabulary tests such as the vocabulary levels test (Nation, 1990) and vocabulary size test (Beglar, 2010; Nation & Beglar, 2007) typically employ a polling method, in which total vocabulary size is inferred from a sample of tested words. A drawback of this method is that it assumes the tested words are randomly sampled from and therefore representative of the tested domain, which can affect test reliability in cases where there are many words in the domain that are far below or above learners' ability. This paper outlines an alternate method for estimating vocabulary size from a test score using item response theory, which allows estimation of total vocabulary size from a nonrandom sample of words well matched to learners' ability, resulting in tests of practical length with high reliability that can be used to estimate the total number of words a learner knows. Such a test scoring method, currently in use at a private university in southern Japan, is used as an example.

1 Background

Perhaps the most qualitatively interpretable vocabulary test score is one that estimates the total number of words the learner knows in the tested domain, such as a frequency word list, or vocabulary taught as part of a course curriculum. In cases where the tested domain includes very large numbers of words; however, it quickly becomes impractical to test learners on every word. Checklists, which simply ask learners to report their knowledge of words can be used, but such self-report measures may not be appropriate in cases where learners are assigned a grade, and there are concerns regarding the tested construct, which could be recognition of words rather than definitional knowledge (see Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001; Mochida & Harrington, 2006). Consequently, many popular vocabulary tests employ multiple-choice formats, or other formats that require the learner to provide the definition of the word, limiting the total number that can be tested within a typical class period.

In such cases, tests that report scores as total vocabulary size typically employ a polling method, in which total size is inferred from a sample of tested words. A

Please cite this article as: Gibson, A., & Stewart, J. (2014). Estimating learners' vocabulary size under item response theory. *Vocabulary Learning and Instruction*, *3* (2), 78–84. doi: 10.7820/vli.v03.2.gibson.stewart

drawback of this method is that it assumes the words are randomly sampled from the domain, which can affect test reliability in cases where there are many words in the domain that are far above or far below learners' ability. Under both classical test theory (Brown, 2005) and most common item response models (Parchev, 2004), tests have the highest reliability when the learner has a 50% probability of a correct guess. This may not be the case for tests of domains that include many high- or many low-frequency words however, leading to a "ceiling" or "basement" effects in scores.

Alternately, vocabulary size can be estimated under item response theory (IRT), which allows vocabulary size to be estimated on a common scale even in cases where learners take tests customized toward their respective levels of proficiency. This allows test makers to create tests with optimal internal reliability, even if the words used are not representative of the entire pool of words that their vocabulary size is estimated for. For example, the number of the first 5,000 most frequent words known by a low-level learner can be estimated by testing them on the items only from the first 2,000. Conversely, a high-level learner need only take a test of the 5,000 band to obtain comparable estimates, as so many higher frequency words are known by them that testing these items provides little information. This paper will outline this method, and describe how it was operationalized for test scoring at private university in southern Japan. It is hoped this paper can help curriculum organizers use similar scoring systems at their own institutions.

2 Context

As detailed in Fryer, Stewart, Anderson, Bovee, and Gibson (2011), curricula for first- and second-year English conversation classes at Kyushu Sangyo University ($n = 4{,}000$) include weekly lists of vocabulary selected from the 2,000 most common words in English as determined by analysis of the British National Corpus (BNC, 2007). The university's Language and Education Research Center (LERC) provides pre- and post-tests to assess student achievement on words taught each semester. Under the current curricula, first- and second-year students are tested on productive knowledge (Stewart, 2012a) of 150 words from the BNC 2,000 band each semester, for a total of 600 over the course of two years. Due to practical concerns however, only 30 of these words are tested as part of a larger vocabulary test at the end of each semester.

Semester tests are equated under the Rasch model, and until 2014, posttest scores were reported to students using a "scale score" based on Rasch ability estimates, with a score of 500 signifying the school average (for further details, see Stewart, 2012b). Although this method is convenient for test equating, a drawback of the logit-based scale score approach is diminished test score interpretability for both students and their teachers. Aside from relation to the school mean, what qualitative interpretation does a score, of, say, 600 signify for students, and in what way can it be interpreted by their teachers? Ultimately, what teachers desire is a precise estimate of the number of words on each semester word list known by students both before and after a semester of instruction rather than a statistical abstraction, regardless of how theoretically

sound the equating of scores is across test forms and student proficiency levels. The center resolved to use the following method to estimate the total number of words students learned each semester using the results of the relatively short tests.

3 Size Estimation under IRT

One of the simplest and most common IRT models, the Rasch model, is written as:

$$P_{ij} = \frac{e(\theta_i - b_j)}{1 + e(\theta_i - b_j)}$$

in which θ represents the ability level of the student, b represents the difficulty of the item on the same scale, and e is the natural log (2.171). Assuming the model fits, the formula can be used to determine the probability that student i will answer item j correctly.

Suppose a student with a theta ability level of 1 encounters the following words (corresponding hypothetical difficulties listed in parentheses):

union (1.25), *proper* (1.49), *mark* (−0.98), *separate* (1.32), *agreement* (1.1)

If the student's ability level is, for example, 1 logit, the differences between her ability and these (hypothetical) word difficulties are as follows:

$$-0.25,\ -0.49,\ +1.98,\ -0.32,\ -0.1$$

The logit differences can then be converted to probability of a correct answer to each test item using the above formula. Using a conventional Rasch model with a slope of 1, the probabilities of correctly providing a definition for each of the above words are as follows:

$$0.44,\ 0.38,\ 0.88,\ 0.42,\ 0.48$$

These probabilities can then be added together to estimate the expected value, i.e., the total number of these words that the student would be expected to know given their ability level, which in this case is approximately 2.59, or about half.

A useful feature of this model is that it is not necessary to test students on every word on a semester word list; with both the ability of a given student and the difficulty of each word on the semester list estimated on a common scale, it then becomes possible to calculate the probability the student knows each word on the list, including words that the individual student was not actually tested on (for an example, see Stewart & Gibson, 2010). Estimation of the total number of words known on a given list can be performed by determining the item pool score (Dorans, 2000), analogous to a computer adaptive test that estimates students' probability of success on hundreds of test items contained in an item bank after administering a subset of those test questions. To estimate the difficulties of words contained in semester word lists, multiple test forms are distributed to students, the sum of which test all the words contained on the list.

All of these test forms and their unique items are statistically equated, and logit difficulty estimates are calculated on a common scale for every word in the curriculum.

4 Method

Doing the above, however, requires giving multiple tests to multiple groups of students, in order to calculate the difficulties of all the respective words. As the class time available for tests must be kept to a minimum, it was necessary for word tests to be divided in such a way that each individual student wrote a subset of no more than 30 test questions. Therefore, items testing all 600 "breadth" words taught as part of the institution's first- and second-year English conversation classes were piloted using a total of 20 separate pretest forms over the course of two academic years and four semesters. Ten pretests for first-year students and ten pretests for second-year students (five for each semester) were created, each with an average of 15 items unique to the form and the remaining as common anchor items added for equating purposes.

For each semester of both years, five tests were first shuffled and then randomly distributed to participating students. This ensured each test form was administered to roughly equal sample sizes (an average of approximately 283 per form), and of roughly equal ability level (the school mean). After two years of data collection, semester test forms were then equated under IRT.

5 Analysis and Results

Data analysis was performed using the *ltm* package for R (Rizopoulos, 2006). Three item response models were tested for fit to the data: a 2-parameter logistic (2PL) model, in which item slopes are estimated separately; a constrained Rasch model, in which the item discrimination (slope) is presumed to be 1 for all items; and finally an "unconstrained" Rasch model, in which a single slope is empirically estimated for all items (typically close to the average slope of the same items under the 2PL model). This final model can not only offer superior model fit in cases where item slopes are relatively uniform, but also substantially higher or lower than 1.

Model fit comparisons conducted in *ltm* using the Akaike information criterion (Akaike, 1973) and Bayesian information criterion (Schwarz, 1978) indicated the unconstrained Rasch model had the best fit. The model's slope was estimated as 1.27, indicating that the tested items have relatively high discrimination.

After selecting the model, ability estimates were computed for the sample. Mean pretest item pool scores were calculated for low-, mid-, and high-level English conversation classes, in order to estimate the number of curriculum words known by students prior to instruction by class level. This was done using an excel spreadsheet (see Kim, 2004, for a comprehensive how-to guide). The spreadsheet calculates probabilities of knowing words by ability level, and sums totals to

			-3	-2.5	-2	-1.5	-1	-0.5	0
Enter Item Parameters Here			**5**	**8**	**14**	**22**	**34**	**47**	**63**
a = Item discrimination b = difficulty			**Probability of correct guess by ability level**						
			θ (ability level of student)						
	a	**b**	-3	-2.5	-2	-1.5	-1	-0.5	0
AC1.1.association	1.27	0.57	**0.01**	**0.02**	**0.04**	**0.07**	**0.12**	**0.20**	**0.33**
AC1.10.management	1.27	1.52	**0.00**	**0.01**	**0.01**	**0.02**	**0.04**	**0.07**	**0.13**
AC1.11.whose	1.27	-1.25	**0.10**	**0.17**	**0.28**	**0.42**	**0.58**	**0.72**	**0.83**
AC1.12.address	1.27	-0.83	**0.06**	**0.11**	**0.18**	**0.30**	**0.45**	**0.60**	**0.74**
AC1.121.structure	1.27	1.25	**0.00**	**0.01**	**0.02**	**0.03**	**0.05**	**0.10**	**0.17**
AC1.122.divide	1.27	1.49	**0.00**	**0.01**	**0.01**	**0.02**	**0.04**	**0.07**	**0.13**
AC1.123.clock	1.27	-0.98	**0.07**	**0.13**	**0.22**	**0.34**	**0.49**	**0.65**	**0.78**
AC1.124.relation	1.27	1.32	**0.00**	**0.01**	**0.01**	**0.03**	**0.05**	**0.09**	**0.16**
AC1.125.encourage	1.27	1.10	**0.01**	**0.01**	**0.02**	**0.04**	**0.06**	**0.12**	**0.20**
AC1.126.go	1.27	2.45	**0.00**	**0.00**	**0.00**	**0.01**	**0.01**	**0.02**	**0.04**

Figure 1. Spreadsheet to calculate the expected value of the ability level of students from the difficulty of the tested items.

Table 1. Mean Logit Ability and Estimated Number of Words Known Prior to Instruction for First-year First Semester Low-, Mid-, and High-level Classes

Level	Mean logit ability	Estimated number of words known
Low	−0.165	57.69
Mid	0.161	68.19
High	0.609	82.83

produce expected values. Figure 1 provides an example of such a file, abbreviated for brevity.

Next, mean student ability was calculated for each class level. This information was used to determine how many of the 150 words students knew by level before beginning the semester (see Table 1).

The results indicate that on average, first-year students have productive knowledge of 63 of these 150 words taught in the first semester prior to instruction. By level, low-level students know approximately 58 words; mid-level students know approximately 68; and high-level students know approximately 83. These estimates will be used as ipsative benchmarks when posttest results are assessed in July 2014; progress on the word list will be measured in reference to student knowledge of the words before entering the course. Our goal is to design the posttest section in such a manner that an estimate of words known that does not differ from pretest estimates will result in a low score, and only students estimated as knowing nearly all the words on the list will be able to attain a perfect score. By doing this, we will be able to report scores on this section of the final test to students and their teachers in a manner that is highly interpretable and has clear relevance to content mastery.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Academiai Kiado.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the yes/no vocabulary test: Some methodological issues in theory and practice. *Language Testing*, *18*(3), 235–274. doi:10.1177/026553220101800301

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

The British National Corpus. (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium (version 3) [BNC XML Edition]. Retrieved from http://www.natcorp.ox.ac.uk/

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. Upper Saddle River, NJ: Prentice Hall Regents.

Dorans, N. J. (2000). Scaling and equating. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 135–157). Mahwah, NJ: Lawrence Erlbaum Associates.

Fryer, L. K., Stewart, J., Anderson, C. J., Bovee, H. N., & Gibson, A. (2011). *Coordinating a vocabulary curriculum: Exploration, pilot, trial and future directions.* JALT2010 Conference Proceedings, Tokyo, Japan.

Kim, J. (2004). *An excel manual for item response theory.* Retrieved from http://coeweb.gsu.edu/coshima/EPRS8410/Sarah_Project1%2012%203%202004.pdf

Mochida, A., & Harrington, M. (2006). The yes/no test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*(1), 73–98. doi:10.1191/0265532206lt321oa

Nation, I. S. P. (1990). *Teaching and learning vocabulary.* Boston, MA: Heinle & Heinle.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13. Retrieved from http://jalt-publications.org/tlt/issues/2007-07_31.7

Parchev, I. (2004). *A visual guide to item response theory.* Retrieved from http://www.metheval.uni-jena.de/irt/VisualIRT.pdf

Rizopoulos, D. (2006). *ltm:* An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. Retrieved from http://www.jstatsoft.org/v17/i05

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

Stewart, J. (2012a). A multiple-choice test of active vocabulary knowledge. *Vocabulary Learning and Instruction*, *1*(1), 53–59. doi:10.7820/vli.v01.1.stewart

Stewart, J. (2012b). The LERC vocabulary program: Score gains for first and second year students. *Kyushu Sangyo University Language Education and Research Center Journal*, *7*, 87–93.

Stewart, J., & Gibson, A. (2010). Equating classroom pre and post tests under item response theory. *SHIKEN*, *14*(2), 11–18. Retrieved from http://jalt.org/test/ste_gib1.htm