Vocabulary Learning and Instruction
Volume 2, Issue 1, September 2013
http://vli-journal.org

Validating a Pictorial Vocabulary Size Test via the 3PL-IRT Model

Wen-Ta Tseng
National Taiwan Normal University
doi: http://dx.doi.org/10.7820/vli.v02.1.tseng

Abstract

The paper presented a newly conceived vocabulary size test based on pictorial cues: Pictorial Vocabulary Size Test (PVST). A model-based (1-2-3 parameter logistic item response theory model comparisons) approach was taken to check which model could absorb the most information from the data. Junior high school and primary school students participated in the study ($N = 1{,}354$). Subjects' ability estimates and item parameter estimates were computed based on expected *a posteriori* (EAP) method, one type of Bayesian method. BILOG-MG 3 was adopted to execute parameter estimates and model comparisons. The results showed that the 3PL-IRT model best fit the empirical data. It was then argued that test takers' English vocabulary size could be best captured under the 3PL-IRT model, as not only the discrimination parameter, but also the guessing parameter has a fundamental role to play in consideration of the test format adopted in the PVST. The article concluded that the PVST could have positive washback effects on test development and English vocabulary instruction.

1 Introduction

Vocabulary size is a key indicator of lexical ability and language proficiency. A wide range of research has consistently indicated that certain and different vocabulary sizes are necessary to complete different language tasks (e.g. Adolphs & Schmitt, 2003; Hazenberg & Hulstijn, 1996; Laufer, 1988; Nation & Waring, 1997). For example, 2,000–3,000 word families are required for basic daily conversation (Laufer, 1988), and 5,000 word families are the threshold to embark on independent reading of authentic texts (Laufer, 1988). Furthermore, to match the lexicon of a native university graduate, a vocabulary size approaching 20,000 word families is needed (Goulden, Nation, & Read, 1990). Carter (1998) remarks that non-native speakers need to obtain 1,000 word families per year to catch up with the level of an educated native speaker. Although it is not a realistic aim for most foreign learners to fully reach the level of an educated native speaker, it is nevertheless essential for foreign learners to commit themselves to sustained vocabulary study in order to reach the vocabulary requirements for even daily conversation and the modest reading of authentic materials.

Given the significance of acquiring sufficient English vocabulary size, the Ministry of Education in Taiwan has also paid close attention to the significant role vocabulary size plays in English language learning. In the year of 2003, the Ministry of Education in Taiwan published the 1,000 Words List (now 1,200 Words List),

Please cite this article as: Tseng, W.-T. (2013). Validating a pictorial vocabulary size test via the 3PL-IRT model. *Vocabulary Learning and Instruction*, *2* (1), 64–73. doi: 10.7820/vli.v02.1.tseng

which are all high-frequency words. The reason for compiling the word list was to create a list that can serve as a guideline for different textbook publishers to refer to when developing texts for primary and junior high school students (Ministry of Education, 2007).

2 Aims

The existence of the word list notwithstanding, it is however noted that thus far there are few attempts made to develop a reliable and valid vocabulary size test on the basis of this word list. To be specific, none of the currently available vocabulary size tests referenced to the word list have been developed under Item Response Theory (IRT) (See below for more information). Thus, the purpose of this project is to develop a multiple-choice vocabulary size test via IRT. In particular, in this initial validation study, it is intended to examine the extent to which the discrimination and guessing parameter should be included for estimating test takers' vocabulary size.

2.1 Operational definition of the trait to be measured

The Pictorial Vocabulary Size Test (PVST) is designed as a diagnostic test, requiring test takers to recognize and match the correct target word forms with a series of pictorial cues. This test format is designed in such a way that the required knowledge of grammar and reading can be minimized and thus can avert the potential risks of measuring something other than the ability to recognize the meanings of the target word forms. Likewise, the use of pictorial cues for test takers to elicit correct word meanings can also receive its validity and has been underpinned by vocabulary researchers such as Nation (1990, 2001). As he states,

> In recognition tests, we want to see if the learners know the meaning of a word after they hear or see it. In such tests the learners hear or see an English word and then. . . . (c) choose one from a set of pictures, mother-tongue words, or English synonyms or definition. (Nation, 1990, pp. 79–80)

The use of pictorial cues as indications of meanings of words, it is argued, is particularly useful regarding the target population considered in the current test project. Considering the need that primary school students can comprehend the test format without problems, it is believed that colorful and vivid pictures which unequivocally represent the underlying concepts of target words can help them, if they know the answers, match the correct word forms with the pictures (meanings) immediately. Arguably, this test feature enables the PVST to obtain strong contextual validity through "an instance of the meaning" (Nation, 2001, p. 85) without recourse to *reading* sentence stems. In other words, the use of pictures can not only establish the contextual validity of the test but also avoid the possibility that the ability to recognize the meanings of word forms may be contaminated by the ability to read sentence stems.

3 Specifications of the pictorial vocabulary size test

(i) Inferences: To estimate Taiwanese students' receptive vocabulary size knowledge through pictorial cues during their primary and junior high education.

(ii) Decisions: (a) To determine the extent to which the target students have succeeded or failed to master the words that are prescribed by the Ministry of Education in Taiwan.
(b) To track the trajectory of the target students' vocabulary size over time.

(iii) Impacts: The test is a low-stakes test. It is designed in a way that the target students' motivation toward learning the 1,200-word list can be greatly enhanced though taking the test.

3.1 Format of the test

To ensure the statistical property of local independence assumed in the IRT, the author adopted a multiple-choice format: a picture prompt with four options. A sample item is shown in Figure 1.

In total, 180 items were sampled from the 1,200-word list and pooled as an initial item bank, roughly one-in-seven. The 180 items were then further divided into two forms, with 90 items included in each form.

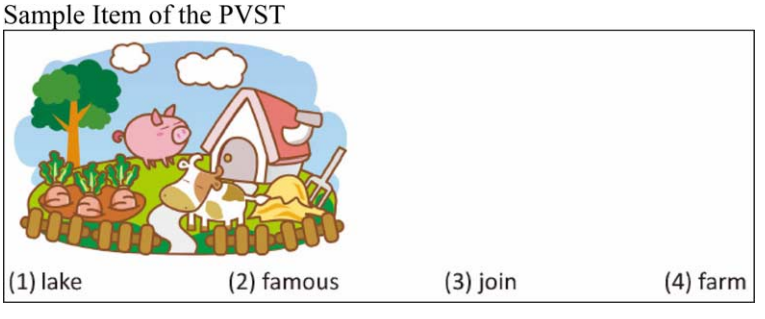

Figure 1. Sample item of the PVST.

4 Methods

4.1 Participants

The participants came from four junior high schools and three primary schools in Taiwan. In total there were 1,354 students participating in the study. The average time spent learning English of this sampling ranged between 0 and 5 years.

4.2 Software

BILOG-MG 3 was adopted to execute parameter estimates and model comparisons. BILOG-MG 3 was a popular IRT program designed for analyzing

dichotomous response test data. It was developed by Zimowski, Muraki, Mislevy, and Bock (2003).

4.3 Item response theory

A model-based (1-2-3 parameter logistic item response theory model comparisons model comparisons) approach was followed to check which model could absorb the most information from the data. The 1PL-IRT model is primarily concerned with the calibration of item difficulty parameter. The 2PL-IRT model take into account not only item difficulty parameter but also item discrimination parameter. Like the Rasch model, no allowance is made for the guessing behaviors of examinees in the 2PL-IRT model. By contrast, the 3PL-IRT model incorporates yet another item parameter – guessing – into the 2PL-IRT model. The reason for including the guessing parameter into the model is to take into account the likelihood of guessing behaviors as likely observed on the examinees at the lower end of the ability continuum.

Subjects' ability to estimate and item parameter estimate were computed based on expected *a posteriori* (EAP) method, a type of Bayesian methods. In practice, the three IRT models are often used to score examinees on a latent-trait in conjunction with three different scoring methods: (1) maximum likelihood (MLE), (2) maximum *a posteriori* (MAP), and (3) EAP. (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). The scoring procedure of MLE is an iterative searching process to find out the value of theta (θ) – latent trait – that can maximize the likelihood of an examinee's response pattern for all the items taken. MLE method has several psychometric strengths. With large samples, it is unbiased and efficient, and the measurement errors with this estimator are normally distributed. However, MLE estimation fails to achieve a solution on all-correct or all-wrong response patterns due to its statistical property. Hence, to solve the underlying problems of MLE scoring method, the other two scoring estimators are further proposed. Both MAP method and EAP method are developed under Bayesian Theory, which does not need to rely on large samples but instead takes advantage of any prior information of test items or test takers. Through the help of some prior information, the problem of MLE's inability to find the θ that maximizes the likelihood function can be delicately addressed. In other words, all-correct and all-wrong response patterns can be estimated by the two alternatives. However, IRT literature suggested that EAP is a better estimation method than MAP, due to EAP's convenience and speed of achieving solution for θ estimation, "EAP is perhaps the most easily implemented scoring strategy across a range of IRT models and testing contexts" (Embretson & Reise, 2000, p. 182).

5 Results

Table 1 reports the model fit indices for the three IRT models. The results showed that the 3PL-IRT model consistently obtained the lowest values of −2Log-likelihood, *Akaike information criterion* (AIC), and *Bayesian information criterion* (BIC) fit indices on the two test forms, suggesting that the 3PL-IRT model absorbed the most of information from the data. Table 2 reports the results of

Table 1. Fit Indices of –2LL, AIC, and BIC for the Three IRT Models

	Form 1			Form 2		
	–2LL	AIC	BIC	–2LL	AIC	BIC
1PL Model	95,109.59	95,291.59	95,781.08	87,182.50	87,364.50	87,853.99
2PL Model	92,661.68	93,021.68	93,989.90	83,816.42	84,176.42	85,144.64
3PL Model	91,561.59	92,101.59	93,553.92	83,086.99	83,626.99	84,079.32

Table 2. Contrast of Fit Indices of the Three IRT Models

	Form 1			Form 2		
	$\Delta - 2LL$	ΔAIC	ΔBIC	$\Delta - 2LL$	ΔAIC	ΔBIC
1PL – 2PL	2,447.91	2,269.91	1,791.18	3,366.08	3,188.08	2,709.35
2PL – 3PL	1,100.09	920.09	435.98	729.43	549.43	1,065.32

model comparisons among the three models. Regarding $\Delta - 2LL$, which refers to the contrasted value of -2Log-likelihood between two IRT models, it could be seen that all the $\Delta - 2LL$ values were above 113.145, the threshold to reach statistical significance, and that the values of AIC and BIC became smaller as the parameters of models increased. In summary, the results showed that the 3PL-IRT model fit significantly better than the 2PL-IRT model and the 2PL-IRT model fit significantly better than the 1PL-IRT model, which suggested that difficulty, discrimination, and guessing parameters should be all kept in the model to precisely measure the targeted population's vocabulary size.

Based on the results calibrated by the 3PL IRT model, reliability of the two test forms was computed, respectively. The results showed that both of the test forms obtained high-reliability indices up to 0.95, indicting a high level of measurement precision across the latent trait continuum.

Figure 2 indicates the nonlinear relationship between expected number correct and theta when the two test forms are combined into a whole test. According to Figure 2, around 45 test items – one-fourth of the whole test – could be answered correctly by the test takers with theta less than -2.5, whereas when theta was greater than 1, nearly all the 180 items could be answered correctly.

Hence, to instruct on how to calculate vocabulary size, a straightforward three-step approach can be taken. First, obtain theta on the horizontal axis. Second, find the expected number correct which corresponds to the theta on the nonlinear curve. Third, calculate the proportion of expected number correct out of 180 items and multiply that proportion by 1,200 words. For instance, for the theta less than -2.5, 45 is the corresponding value on the curve. Then, divide 45 by 180 and multiply by 1,200. This leads to 300 words. Likewise, for the theta at 0, 150 is the intersecting value on the curve. Divide 150 by 180 and multiply by 1,200. Hence, test takers whose theta is found at 0 should know 1,000 words.

It can be seen from Table 3 that test takers with theta less than -1.5 recognize fewer than 400 words, whereas those with theta larger than 0 could have acquired more than 1,000 words. For the theta values between -1.5 and 0, there is

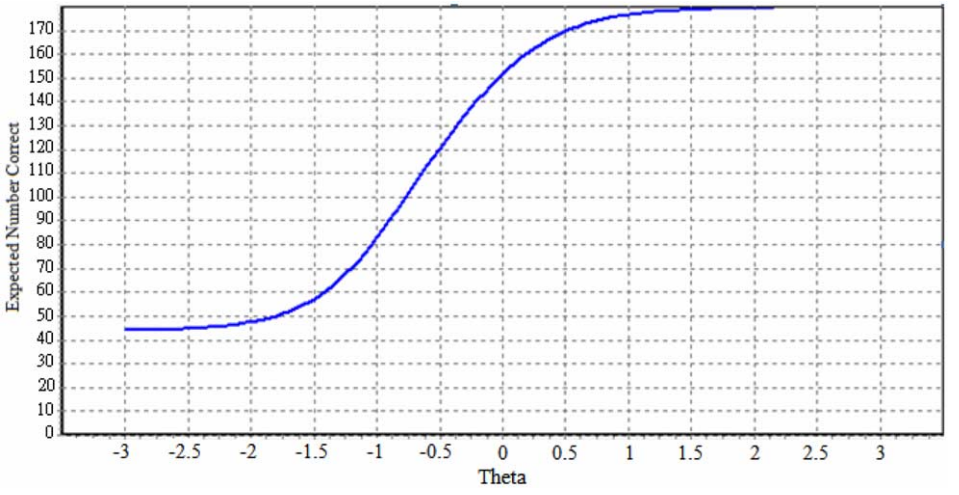

Figure 2. Expected number correct corresponding to theta level.

Table 3. Correspondence Table of Theta, Expected Number Correct, and Vocabulary Size

Theta	Expected number correct (items)	Vocabulary size (words)
−3	45	300
−2.5	45	300
−2	50	333
−1.5	58	387
−1	85	567
−0.5	120	800
0	150	1,000
0.5	160	1,067
1	178	1,187
1.5	180	1,200

a linear correspondence between theta estimates and words, as every 0.5 unit increase on theta corresponds to an increase of around 200 words; meanwhile such a linear mapping is not observed at the two ends of the nonlinear curve.

Table 4 further shows the means and standard deviations of the three IRT parameters for the entire set of 180 items. The mean difficulty of the whole test was −0.518; the mean discrimination, 2.26; the mean guessing, 0.210.

Table 4. Means and Standard Deviations of the Three IRT Parameters

	Mean	SD
Difficulty	−0.518	0.417
Discrimination	2.260	0.687
Guessing	0.210	0.086

As each test item had four options, the threshold for incurring guessing was set at 0.25. Following this criterion, it was found that there were in total 34 items whose guessing parameter went beyond 0.25. The item which was most easily subject to guessing was Q10, as shown in Figure 3. The guessing parameter observed on Q10 was found to be 0.418, and its associated item characteristic curve is shown in Figure 4. The *c* value marked in Figure 4 indicates the probability of guessing by Q10.

Figure 3. Test item: Q10.

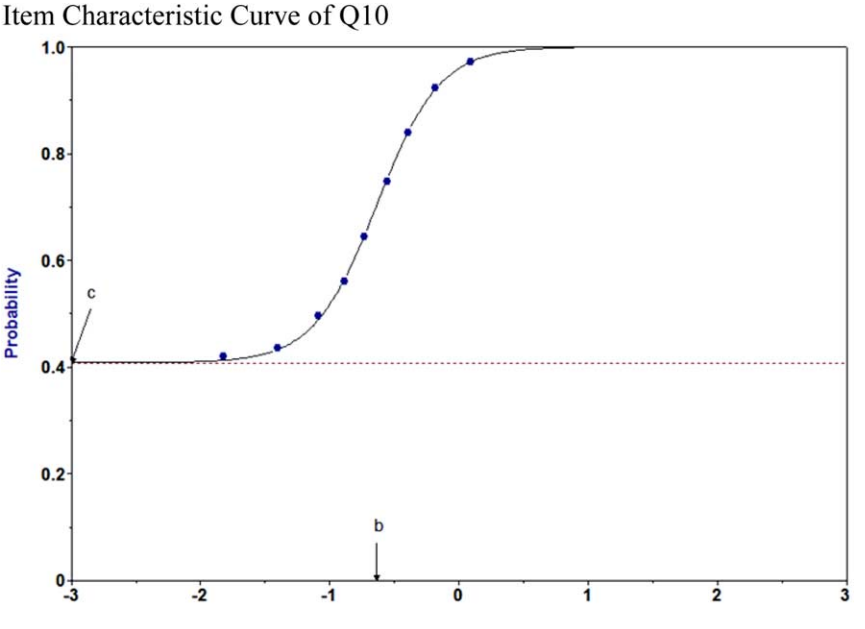

Figure 4. Item characteristic curve of Q10.

6 Discussion

The current study utilizes an IRT model-driven approach to developing and validating a newly conceived English vocabulary size test. The targeted word list is derived from an EFL educational context. It includes the most high-frequency words in any word bank such as BNC. Due to the high-frequency property of the word list, pictorial cues are designed as the prompts for test takers to connect the word meanings to the corresponding word forms. The results of the study demonstrate that the newly conceived PVST can be successfully developed and validated by the 3PL-IRT model. The 3PL-IRT model acquires the best fit to the empirical data. On the validation basis of the 3PL-IRT model, the test forms of the PVST can also obtain a high level of measurement precision.

In consideration of the mean difficulty, discrimination, and guessing parameters, the PVST can be considered moderately easy for test takers to accomplish, highly discriminating in differentiating test takers' vocabulary size, and efficiently capable of modeling the chances of item guessing phenomena. In other words, it is argued that test takers' English vocabulary size can be best captured under the 3PL-IRT model, as not only the discrimination parameter, but also the guessing parameter has a fundamental role to play regarding the multiple-choice test format adopted in the PVST.

Measuring EFL learners' English vocabulary size is possible via both classical test theory (CTT) and IRT (Stewart & White, 2011). Under CTT, the estimation of vocabulary size is made possible through a balanced sampling from different frequency bands of a word list. Although this classical testing procedure appears intuitive and straightforward, it does not account for the discrepancy between raw scores and item responses. This is because CTT models raw scores, whereas IRT models item responses. In the case where test takers obtain the same raw scores, they will be thought to have the same vocabulary size. However, it is very likely that the same raw scores are derived from different item responses. This is often observed on long tests. Using raw scores to estimate vocabulary size may hence incur a high chance of miscalculating the *true* value to be measured. In a typical vocabulary size test that adopts the multiple-choice format such as the PVST and the Vocabulary Levels Test (Nation, 1990), the overall length of the test tends to go beyond 100 items. Modeling item responses rather than raw scores in such tests not only greatly increases the likelihood of capturing the true value of vocabulary size, but also makes it possible to model the guessing phenomena of the test items. Measuring English vocabulary size via IRT clearly possesses significant psychometric strengths over CTT.

Developing an IRT-based vocabulary size test has important theoretical, instructional, and social implications and consequences. Theoretically, it is enlightening to see the extent to which IRT can be applied to modeling the development of the whole test. Pedagogically, Ryan (1997), on discussing the ability to recognize word form, argues that "failures at the word level can severely hamper reading ability, and reading ability is a key skill in using English for academic or professional purposes" (p. 187). Hence, the PVST can be used to diagnose or identify both primary and junior high school Taiwanese learners' strengths and

weaknesses regarding the words in the list. Then, appropriate and effective vocabulary learning programs can be further established.

7 Conclusion

In conclusion, the results of this validation study suggest that the 3PL-IRT Model is a superior psychometric model over both the 1PL- and 2PL-IRT model in the data-model-fitting process. A conversion formula transforming IRT scores to vocabulary size estimates is also successfully acquired. Further equating work should be undertaken to ensure the equivalence of the two test forms developed. The article concludes that the PVST can have positive washback effects on test development and English vocabulary instruction.

Acknowledgement

This project was funded by National Science Council Research Grant (99-2410-H-003-081-MY2). The author is indebted to Prof. Yuh-show Cheng for her critical feedback on the design of the pictorial cues as well as the distracters during the process of test development.

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics, 24,* 425–438. doi:10.1093/applin/24.4.425

Carter, R. (1998). *Vocabulary: applied linguistic perspective* (2nd ed.). London: Routledge.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Goulden, R., Nation, I.S.P., & Read, J. (1990). How large can a receptive vocabulary be?. *Applied Linguistics, 11*, 341–363. doi:10.1093/applin/11.4.341

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Norwell, MA: Kluwer Academic.

Hazenberg, S., & Hulstijn, J.H. (1996). Defining a minimal receptive second language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics, 17,* 145–163. doi:10.1093/applin/17.2.145

Laufer, B. (1988). What percentage of text-lexis is essential for comprehension? In C. Laurén & M. Nordmann (Eds.), *Special language: From humans to thinking machines* (pp. 316–323). Clevedon: Multilingual Matters.

Ministry of Education. (2007). *Basic 1200 English words.* Taipei: Ministry of Education.

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

Nation, I.S.P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press. doi:10.1017/CBO9781139524759

Nation, I.S.P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy*. Cambridge: Cambridge University Press.

Ryan, A. (1997). Learning the orthographical form of L2 vocabulary – A receptive and a productive process. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: description, acquisition, and pedagogy*. Cambridge: Cambridge University Press.

Stewart, J., & White, D.A. (2011). Estimating guessing effects on the vocabulary levels test for differing degrees of word knowledge. *TESOL Quarterly*, 56, 370–380.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, E. (2003). BILOG-MG 3. In du Toit, M. (Ed.), *IRT from SSI*. Lincolnwood: Scientific Software International. Inc.