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

Article

A Review of Four Studies on Measuring Vocabulary Knowledge

Akiyo Hirai
University of Tsukuba
doi: http://dx.doi.org/10.7820/vli.v03.2.hirai

Abstract

The purpose of this paper is to review each of the following four vocabulary studies: (1) Estimations of Japanese University Learners' English Vocabulary Sizes Using the Vocabulary Size Test (VST), by Stuart McLean, Nicholas Hogg, and Brandon Kramer; (2) Local Item Dependence on the Vocabulary Levels Test (VLT) Revisited, by Tadamitsu Kamimoto; (3) Test Taking and DK Use on the VST, by Dawn Lucovich; (4) Estimating Learners' Vocabulary Size under Item Response Theory (IRT), by Aaron Gibson.

1 Estimations of Japanese University Learners' English Vocabulary Sizes Using the Vocabulary Size Test

McLean, Hogg, and Kramer investigated university students' vocabulary sizes using the VST (Nation & Beglar, 2007). They found that the VST seriously overestimated vocabulary size. They showed that even the low-level students maintained a score of approximately 20% correct at the 5,000 vocabulary level and higher, instead of decreasing linearly as the difficulty of the vocabulary level went up. A major cause for this result is the VST's four-option multiple-choice (MC) format, which allows for a high probability of guessing correctly. Another cause is students' familiarity with English loanwords in Japanese. Although many loanwords are relatively low frequency words, they can be easier for Japanese students because they are already familiar with the pronunciation and meanings of the words from Japanese.

Besides these major findings, the authors raised the interesting question of whether university departments' *hensachi*, or *T*-score, can be a useful proxy for English vocabulary size. A *hensachi* score is a standard score used to rank which percentile students or schools are in. *T*-scores always have an average of 50 and are calculated using the following formulas:

$$T\text{-score} = 10\ (z\text{-score}) + 50$$

$$[z\text{-score} = (\text{a raw score - average}) / \text{standard deviation}]$$

To answer this research question (RQ), the authors took the correlation between departments' *hensachi* scores, which are widely available on the Internet, and 3,427 students' vocabulary sizes. They found a relatively high correlation of .73, and

Please cite this article as: Hirai, A. (2014). A Review of Four Studies on Measuring Vocabulary Knowledge. *Vocabulary Learning and Instruction*, *3* (2), 85–92. doi: 10.7820/vli.v03.2.hirai

therefore concluded that the departments' *hensachi* are a fair indicator of Japanese university students' relative vocabulary size.

Such a high correlation is reasonably expected because the departments' *hensachi* represents students' general proficiency levels. Vocabulary knowledge is widely regarded as a major factor in English proficiency (e.g., Milton, 2009, pp. 175–176; Stæhr, 2008). In particular, the relationship between vocabulary size and reading ability, which is the main skill tested by most Japanese university entrance examinations, is the strongest among the four language skills. According to Stæhr (2008), as much as 72% of the variance in the ability to score an average mark or above on the reading test can be explained by vocabulary size. Thus, simply proving the correlation coefficient is not enough to show the significance of this study. Besides, their rich data deserve further explanation.

One idea to make the study more significant and informative is to draw a best-fit regression line from *hensachi* to vocabulary size and create a conversion table. It would be helpful in grasping students' vocabulary size if the conversion table provided vocabulary sizes that are equivalent to *hensachi* scores, starting from the lowest score, i.e., around 40, and progressing upward in small steps, until the highest, around 80. A teacher might want to know, for example, the mean vocabulary size that is equivalent to a departmental *hensachi* score of 60 when she wants to decide on a textbook to choose before the class starts. At present, *hensachi* scores are used for classifying only three proficiency groups, which seems to be a bit too vague and uninformative. A group with *hensachi* below 50 is categorized as the Low group, and approximately one-third of students fall into that category. This category should be further divided to identify lower-level students' vocabulary sizes. Since a large sample of data were collected from all over the country, a relatively accurate regression line for the conversion table may be obtained.

In relation to the validity of vocabulary size estimation derived from *hensachi* scores, the inclusion of VST scores for first-year or at most second-year university students alone may produce a more precise relationship between *hensachi* and vocabulary size. It is possible that third- and fourth-year students' vocabulary knowledge has changed since leaving high school. Another related issue concerns the correlation coefficients between the VST scores and other test scores, such as the *Test of English as a Foreign Language* (TOEFL) and *Test of English for International Communication* (TOEIC) tests. Because the sample sizes for these correlations are small, it would have been better to also provide the *hensachi* level and vocabulary size of the students.

Lastly, McLean, Hogg, and Kramer's study deserves credit for collecting such a large sample of students and revealing that Japanese university students' English vocabulary sizes are wide-ranging. They have a mean of 3,715 words, and the growth of their vocabulary knowledge is not parallel to the frequency lists. Thus, the selection of classroom material must be made with care, depending on their majors and needs, in order to provide appropriate instruction.

As in the case of the VST, the Vocabulary Levels Test (VLT; Nation, 1990) seems to have a similar tendency to overestimate test takers' vocabulary knowledge, which is pointed out by the next study.

2 Local Item Dependence on the Vocabulary Levels Test Revisited

The second study was done by Kamimoto, who claims that LID arises as a result of the matching format of the VLT. To examine the LID, he compared the original test of three 6 3 (i.e., 6 words and 3 definitions translated in Japanese) matching forms and the combined 18 9 (i.e., 18 words and 9 definitions) form.

LID should be avoided since it violates local independence, one of the assumptions of both classical and item response testing theories. Local independence means that "when the abilities influencing test performance are held constant, examinees' responses to any pair of items are statistically independent" (Hambleton, Swaminathan, & Roger, 1991). In this regard, the format of the VLT seems to be a problem since after one or two prompts have been answered from the six options, and the number of options for the last definition is reduced. In other words, these definitions and options influence each other. Thus, to reduce the local dependence, Kamimoto increased the options and definitions by three times as many as were on the original form.

However, if the matching format is the source of the LID, the 18 9 form is still regarded as a matching format and, strictly speaking, the LID still exists. Thus, as shown in Figure 1, it would be worth comparing the original 6 3 form with 6 1 or 4 1 form (i.e., MC formats). For example, as shown in Figure 1, one set of the 6 3 form can be divided into three sets of 6 1 forms for the same target words, and each set can be presented to different groups. Thus, six sets of the 6 3 form can be broken down as follows:

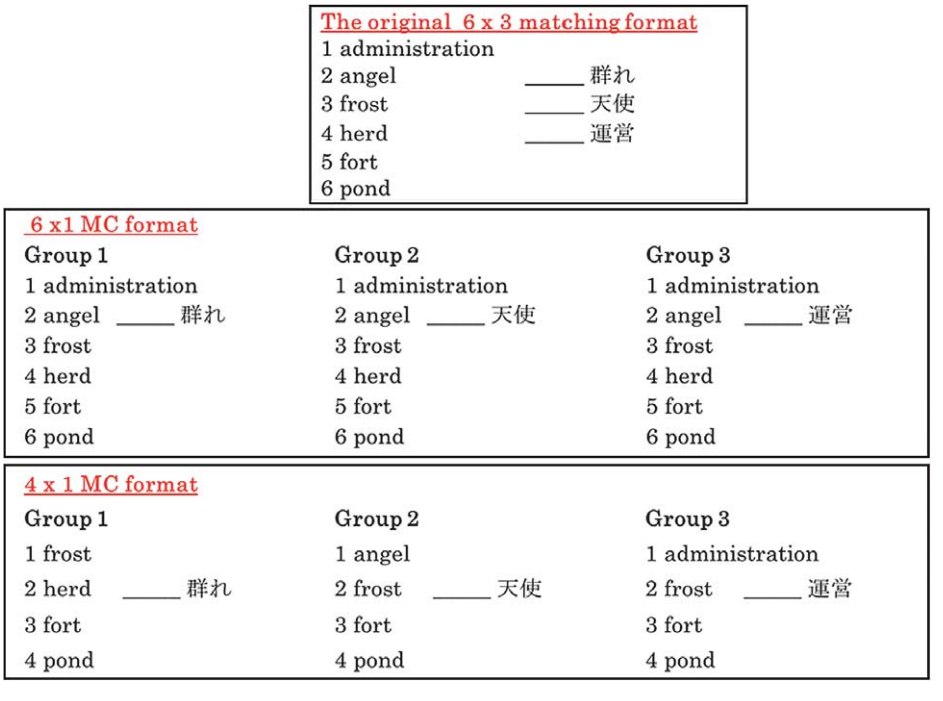

Figure 1. Alternative designs for comparing the original matching form with MC forms.

Group 1 takes six sets of the 6 1 form (the six options and the first prompt of Forms A and B); Group 2 takes six sets of the 6 1 form (the six options and the second prompt of Forms A and B); and Group 3 takes six sets of the 6 1 form (the six options and the third prompt of Forms A and B). A couple of weeks later, all the groups take the original 6 3 form (Forms A and B). By comparing the original form with the MC responses, guessing produced by the LID could be more clearly shown.

As for the Item Facility (IF) analyses of the 6 3 form and 18 9 form, the Type 2 IF index, which represents responses answered correctly on the 6 3 form but incorrectly on the 18 9, and the Type 3 IF index, which indicates responses answered incorrectly on the 6 3 form but correctly on the 18 9 form, are very informative. Through these analyses, the author successfully detected high guessing items, such as the words "spoil," "stool," and "deficiency," and pinpointed the causes of the high guessing of these items by examining the other options and definitions in the sets. Thus, the Types 2 and 3 IF analyses can be used diagnostically to replace problematic distractors in the original VLT test.

Following the detailed individual item analyses, the author summarized the responses categorized in the Type 2 IF index and found an average guessing rate of 19%. However, the responses categorized in Type 3 should also be summarized because these responses seem to come from purely random guessing. By comparing the guessing rates for these two types, the author could show more clearly how much more the 6 3 form allows for guessing than the 18 9 form does.

With such a high guessing rate (i.e., the mean rate of Type 2 responses) in the original VLT, Kamimoto concluded that items in clusters tend to be more dependent than the previous Rasch-based studies indicate (e.g., Schmitt, Schmitt, & Clapham, 2001). However, it is worth noting that the VLT the authors used for this study was the Japanese version of the original VLT, while Schmitt et al. (2001) used revised versions. Such a difference may have led to somewhat different results, besides the difference in the use of Rasch analysis. Although it is difficult to make clear whether the high guessing rate of 19% is due to LID or simply the difference in the probability of guessing from 6 options or 18 options, this study is significant in that it shows clearly a high level of guessing that is hard to ignore.

As the above two studies pointed out, test format is a major influence on the item difficulty of a trait measured. Both the MC format and the matching format were found to allow test takers to answer correctly using various types of guesswork, such as informed guessing and uninformed (i.e., random) guessing. From the viewpoint of the washback of a test, a good item is one that encourages test takers to interact with it actively, making them try hard to retrieve their target knowledge. In this regard, the act of informed guessing, the process in which test takers use their partial knowledge while guessing, is preferable to uninformed guessing. The next study deals with reducing only the latter type of guessing.

3 A Qualitative Analysis of the Options on the VST

Lucovich has suggested the addition of a "don't know" (DK) option to the current four-option items of the VST to reduce test takers' uninformed guessing.

To investigate the effect of the DK option, she addressed the following RQs: (1) How did two Greek advanced learners of English (as examples of non-Japanese/non-English test takers) determine their answers on the VST? (2) How did they qualitatively perceive and use the DK option? (3) Did the two Greek participants differ in the choice of the DK option from the first language (L1) American English and L1 Japanese users of English?

Regarding RQs (1) and (2), the two participants took two versions of the 100-item VST (one without the DK option and one with the DK option), and their responses were coded through the test takers' retrospective think-aloud protocol. Lucovich reported that 4% of the items were categorized as "uninformed guesses," which implies that even with the DK option included, it is difficult to eliminate uninformed guessing completely. However, the effect of the DK option was still clear because 12% of the items were coded as "DK usage." In addition, the interview data reveal that the test takers chose the DK option ONLY when they did not have any knowledge of the correct answer. In this way, Lucovich successfully showed that a large proportion of purely uninformed guessing could be avoided with the DK option. By categorizing the answers of the VST and using an interview technique, Lucovich shed light on psychological aspects of test-taking behavior that could not be revealed by quantitative methods alone.

However, to make her study more interpretable, a couple of points are recommended. First, the focal point of this study is the difference in test-taking strategy between tests with or without the DK option. Thus, it would be better to show the coding results of both versions of the test separately for comparison. Second, the four figures of the results of the two tests by the two participants are informative. Thus, for the sake of comparison, each participant's two figures should be placed next to each other. Moreover, all the figures should mark off the same divisions on the scale of the *x*-axis. In her draft, those divisions are slightly different across the four figures (for example, "0, 3, 6, 9, 15" and "0, 3, 6, 10, 13, 16" for Participant A and "0, 3, 6, 9, 15" and "0, 3, 5, 8, 11, 13, 16" for Participant B). In addition, the figures need to be explained. Likewise, careful layout of the tables and figures would help readers to interpret them correctly.

Concerning RQ (3), Lucovich reported that the two Greek participants used more DK options than the intermediate-proficiency L1 Japanese users of English she had analyzed in her previous study (Lucovich, 2013). It would have been helpful to have identified whether this difference was the result of differences in proficiency, personality, or test-taking conditions. She also should have provided more information about the participants in the previous study in the Introduction of this current study if such a comparison was to be made.

With regard to the limitations of the study, as Lucovich acknowledges, it is difficult to generalize from the results of this study because there were only two Greek participants, and they had similar proficiency levels. Thus, to determine the factors affecting the use of the DK option, research that is more comprehensive is required, perhaps by combining the current study with the previous one in which the L1 Japanese and L1 English users participated, and by adding some less proficient learners of English.

Zhang (2013) argues that whether the DK option works well or not may depend on the level of the stakes. If test takers' scores reflect their grade, they may not dare to choose the DK options unless some penalties are added. If, however, the scores are used for placement or diagnostic purposes, without affecting test takers' grades, the DK option may work well, and a more accurate estimation of their vocabulary size would be obtained. Therefore, it might be an interesting topic to explore what kinds of instructions and/or penalties are necessary to make the DK option function properly, even in a high-stakes situation.

4 Estimating Learners' Vocabulary Size under Item Response Theory (IRT)

The last study, by Gibson, is about an alternative method for estimating vocabulary size using IRT. The equating of several test forms using IRT and the estimation of all the difficulties of the items on the same scale merit attention because learners' vocabulary knowledge can be estimated using a smaller number of words than could be done by a test without using IRT. In addition, words fit to the test taker's ability can be selected in a non-random way for a test as long as these words have been calibrated.

Gibson deserves particular credit for the following three points. First, he adopted a counterbalanced design in a careful manner. The five forms to be pre-tested each semester were shuffled and distributed randomly to students, even within a single class. Consequently, the sample size and test takers' ability levels were equal across the test forms, which made more precise estimation of items possible. Second, he tried to make IRT scores (i.e., logit values) more interpretable for teachers and students. He calculated the estimated number of words known out of 150 words to be taught in a semester, based on "the probability of correct guess by ability level." For example, a student with a logit value of zero, which is 50% probability of answering the total number of words correctly, is estimated to know 75 words out of 150 words. The third point is that he used a productive vocabulary test called "Active MC," in which students identified the first letter of a word by looking at L1 definitions, the part of speech, the second and third letters in the word, and the number of letters contained in the entire word (Stewart, 2012a). Assessing students' productive vocabulary knowledge is more valid than assessing receptive knowledge because the test takers are students in English conversation classes who learn these words productively. In addition, the use of a receptive vocabulary test such as the VST and VLT would overestimate learners' mastery of words as mentioned by three studies above.

Although estimating learners' vocabulary size by using IRT has great potential, there are some drawbacks. First of all, under IRT, all the words need to be calibrated beforehand, which would take time. Second, related to the first point, students' vocabulary size can only be estimated with the words calibrated. Therefore, a learner's total vocabulary size cannot be measured as it can be by using the frequency-based VLT or VST. In this study, the total vocabulary size is 150 words, not 1,000 or 2,000 words. However, it is not impossible to estimate the total size out of a frequency level of 2,000 if all 2,000 words have been calibrated. One way to do this is to expand the item bank that stores calibrated items bit by bit.

Whenever a pretest is given to students at the beginning of each semester, some new items can be mixed with the calibrated items, similar to the way that *Educational Testing Service* (ETS) includes some unscored listening or reading items in TOEFL tests. Later, these items can be calibrated by anchoring them to the rest of the calibrated items.

Since the aim of Gibson's study is to outline the equating procedures of an in-house testing project, it is difficult to understand the whole picture of the equating project without reading other relevant papers, such as Stewart (2012a, 2012b) and Stewart and Gibson (2010). However, the project seems to be promising and certainly helps teachers who want to minimize testing time but still get an accurate evaluation.

5 Conclusion

One important issue commonly raised in these studies was the overestimation of vocabulary size under the current VLT and VST. Some measures may need to be taken to deal with this problem, such as improving the format or individual distractors, adding a "DK" option, penalizing random guesswork, deducting a certain percentage from the score obtained (i.e., estimated vocabulary size), or using a different method such as IRT. Given that it is impossible to create one single perfect test, as these studies have suggested, we need to improve or evaluate currently available vocabulary tests such as the VLT and VST by considering the purpose of their use and aspects of their assessment, including validity, reliability, washback, and practicality.

References

Hambleton, R.K., Swaminathan, H., & Roger, H.J. (1991). *Fundamentals of item response theory (Vol. 2)*. London, UK: Sage.

Lucovich, D. (2013). The inclusion of "I don't know" on the Vocabulary Size Test. *Tokyo JALT Journal*, *1*, 28–31.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Clevedon OH: Multilingual Matters.

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.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*, 55–88. doi:10.1177/026553220101800103

Stæhr, L.S. (2008). Vocabulary size and the skills of listening, reading and writing. *Langauge Learning Journal*, *36* (2), 139–152, doi:10.1080/09571730802389975

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.

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