Developing Equivalent Forms of a Test of General and Academic Vocabulary

Phil Bennett
Miyazaki International College

Tim Stoeckel
Miyazaki International College

Reference Data:

Bennett, P., & Stoeckel, T. (2013). Developing equivalent forms of a test of general and academic vocabulary. In N. Sonda & A. Krause (Eds.), *JALT2012 Conference Proceedings*. Tokyo: JALT.

This paper outlines the development of a new vocabulary test that assesses written receptive knowledge of the words in the General Service List and the Academic Word List. The test is intended to enable the provision of diagnostic feedback and goal setting over the course of a program of study. To avoid a possible testing effect from repeated assessment, 4 forms of the test were created, each made to the same blueprint. The instrument was field-tested with 334 Japanese university students, and results were analyzed from a Rasch measurement perspective. The vast majority of test items demonstrate good technical quality, test reliability for the 4 forms ranges from .87 to .93, and the 4 test forms have been found to be equivalent for use with Japanese students, within 1 standard error.

本稿では、新たな語彙テストの作成過程の概略を述べる。このテストは、頻出基本単語リスト（GSL）と学術基本単語リスト（AWL）の書面における受容語彙知識を測定するものであり、高等教育および大学教育における学習過程を通して、診断的なフィードバックを与え、目標設定を容易にする目的で作られている。度重なる試験の施行から生じるテスト効果の可能性を回避するため、4形式のテストが作成されており、それぞれは同じ設計書（ブループリント）に基づいている。334名の日本人大学生を対象にこのテストを行い、結果はラッシュモデルで分析した。テスト項目の大多数は性質上正確であり、日本人学生を対象に使用した場合、4形式のテストの信頼性は.87から.93であり、1標準誤差以内であることが判明した。

Vocabulary, once a somewhat neglected aspect of language learning, has now gained a far more prominent position in the field of language acquisition. Several empirical studies have demonstrated high correlations between vocabulary knowledge and performance on tests of the four main language skills (Meara & Buxton, 1987; Milton, Wade, & Hopkins, 2010; Stæhr, 2008). From studies such as these, attempts have been made to estimate the required vocabulary sizes to achieve competence at various language tasks. These estimates show some variation, but the figure of 2,000 words has regularly been put forward as indicative of a "threshold" vocabulary size, without which little can be comprehended (Milton, 2009; Stæhr, 2008).

Vocabulary size is often measured in terms of the number of word families a learner knows. A word family is a headword plus its inflections and closely related derivations. Bauer and Nation (1993) developed a system for determining word family membership based on the criteria of frequency, productivity, predictability, and regularity to grade the affixes used to

produce inflected and derived forms. This system has been employed in the development of several important word lists (Bauman & Culligan, 1995; Coxhead, 2000; Nation, 2006). Studies investigating the number of word families necessary for comprehension of oral interaction beyond a very basic level have proposed figures in the 2,000-3,000 word family range (Milton, 2009; Schmitt, 2010), and learners are likely to require 4,500 word families or more to be able to comprehend a range of written text types and to achieve passing scores on higher level English examinations (Milton & Hopkins, 2006; Nation, 2006; Schmitt, 2010).

If these values are accepted, then language teachers have a benchmark against which to judge learner progress and set appropriate goals. The provision of clear goals that are perceived as important and challenging, yet attainable, is one of the key elements of goal-setting theory as described by Dörnyei (2001). Since most learners of English in either secondary or tertiary institutions follow courses that are at least a year in duration, commitment to learning could be enhanced if regular assessment and individualized vocabulary learning goals were included in language programs.

Word Lists: The Frequency Model and Specialized Needs

Frequency is the standard principle by which vocabulary is organized and sequenced for testing. It is widely recognized that a relatively small number of highly frequent words comprises a very large proportion of typical English texts (Nation, 2001), and the frequency model predicts that the more frequent a word is, the more likely learners are to recognize it (Brown, 2012; Meara, 1992). However, Zipf (as cited in Milton, 2009) has demonstrated that the effects of the model are limited at lower frequency levels. Aizawa's (2006) study of word recognition among Japanese university students found that, beyond the fourth 1,000-word band of English, differences in learners' recognition were no longer statistically significant and were in some cases inconsistent with the predictions of the frequency model. This, coupled with the fact that less frequent words offer progressively lower text coverage, suggests that at some point it would be more beneficial for learners to tailor their vocabulary learning to their individual needs than to study progressively less frequent word bands.

The Academic Word List (AWL; Coxhead, 2000) serves such a purpose for learners in academic settings. The AWL is a list of 570 word families that commonly occur in a range of academic texts. It was compiled as a focused set of lexical items for learners of academic English to study once the words on the General Service List (GSL; West, 1953) have been acquired. The GSL was developed originally to aid the writing of simplified texts for language learners but has also been used to define a minimum vocabulary threshold for comprehension of basic discourse. A frequency-ranked version of the GSL was compiled by Bauman and Culligan (1995). This revised list comprises 2,284 word families and can be divided into two sublists, covering approximately the first and second 1,000 words of English (hereinafter GSL1 and GSL2). While it has been criticized for its age and coverage (Hancioğlu, Neufeld, & Eldridge, 2008), the GSL has been shown to cover around 75% of the words in academic text (Coxhead, 2000) and 80-90% of texts in other genres (Nation, 2001). Taken together, the GSL and AWL provide coverage of around 86% of academic texts (Coxhead, 2000).

Vocabulary Testing Instruments

Two of the more well-known tests of word recognition are the Vocabulary Levels Test (VLT; Nation, 1983; Schmitt, Schmitt, & Clapham, 2001) and the Vocabulary Size Test (VST; Nation & Beglar, 2007). The VLT is primarily intended as a diagnostic

tool, providing feedback on gaps in learners' vocabularies at the 2,000, 3,000, 5,000, and 10,000 word-frequency bands, as well as in a band of words drawn from the AWL. The VST offers a measure of vocabulary size. It contains target items drawn from the first to the 14th thousand-word frequency bands of the British National Corpus. Scores on the VLT and VST are used to estimate the percentage of words known in each tested frequency band and overall vocabulary size, respectively (Beglar, 2010; Nation, 1983). These interpretations, which are derived directly from raw scores, are meaningful to learners and educators and have been used as measures in numerous studies of the relation between vocabulary knowledge and other aspects of second language learning (e.g., Laufer & Ravenhorst-Kalovski, 2010; Stæhr, 2008).

One limitation to both of these instruments is the lack of multiple forms. In their most recent incarnations, only two forms of each instrument have been made available. As a result, repeatedly using either instrument over the course of a program of study to monitor vocabulary growth risks a testing effect.

Equating Tests of Vocabulary Knowledge

When using multiple versions of a test to track vocabulary development, the equivalency of test forms must be established, or the scores need to be transformed to a common scale. However, the primary obstacle for equating L2 vocabulary tests has been meeting the requirement of population invariance, which demands that the equating function be identical for each significant subpopulation (Petersen, 2007). Schmitt et al. (2001) found establishing equivalency of two versions of the VLT to be untenable due to differences in English vocabulary knowledge stemming from learners' various L1 backgrounds.

Purpose

This paper introduces and describes the ongoing development of a new test of vocabulary knowledge. Our objective is to produce an instrument capable of tracking the development of threshold English vocabulary knowledge for Japanese students in academic contexts. To avoid the possibility of a testing effect, four forms of the test were made, each following the same blueprint. The goal was for these forms to be of equivalent difficulty such that raw scores could be used and interpreted interchangeably. By focusing our study on native Japanese speakers, we hoped to eliminate the problems encountered by Schmitt et al. (2001) in equating test forms for speakers from multiple L1 backgrounds.

Such an instrument could serve several valuable purposes. First, it could provide learners with diagnostic feedback on gaps in knowledge of the core vocabulary needed in academic settings. Second, it could help teachers choose texts of appropriate lexical difficulty. Third, it could assist English programs in setting suitable vocabulary learning objectives and determining whether those objectives are being met. Finally, it could provide researchers with a tool for longitudinal studies of vocabulary development where repeated measurement is required.

The following sections will describe the test and its development and report the results of field-testing in terms of item quality, test reliability, and equivalency of test forms.

Instrument Development

Item Development

Test items were designed to assess written receptive knowledge of the GSL1, GSL2, and the AWL. Items were written for 80 target words randomly selected from each of these bands, creating a bank of 240 items.

Test items share many of the same specifications as those in the VST (see Beglar, 2010; Nation & Beglar, 2007). A multiple-choice format was used because of its universal familiarity and because unambiguous results can be quickly obtained. The stem of each item includes the target word in bold typeface followed by a short sentence that uses the word in a natural, nondefining context. This contextualized format has been found to help examinees clarify word meaning (Henning, 1991) and can lead to beneficial washback when compared to discrete point vocabulary measures (Qian, 2008). For the stem of each item, the Corpus of Contemporary American English (http://corpus.byu.edu/coca/) was consulted to confirm that one of the most frequently occurring members of the target word family and its common collocates were used in the example sentence. As in the VST, the stem is followed by answer choices that include the definition of the target word and three distractors.

To avoid construct-irrelevant difficulty (Messick, 1995), test items were written with simplified language. Specifically, items targeting knowledge of the GSL were written with the most frequent 1,000 words of the GSL, and items targeting knowledge of the AWL were written with words from the GSL. A small number of items did not conform to these guidelines, but in each of these cases the words used were among the most frequent 1,000 of either the British National Corpus (accessed at http://www.lextutor.ca/vp/bnc/) or the JACET 8000 list (Aizawa, Ishikawa, & Murata, 2005) (e.g., *conversation*, *rain*), or they were English loanwords in the Japanese language (e.g., *coffee*, *computer*). None of these exceptions was judged to be overly difficult for the target population of examinees.

Though several item features are shared with the VST, a distinct difference is that, for some GSL items (e.g., *metal*, *curve*, *pull*), the four answer choices are in the form of pictures rather than words. It was felt that in cases such as this, pictures would better assess knowledge of the target word than written choices which require less frequently occurring words than the target word itself. This was the approach taken by Nation (2001) in the 1,000-word level version of the VLT.

Expert Review and Piloting

Each test item underwent expert review and was then piloted with learners of English in one Japanese university. The information collected during piloting was utilized to identify items in further need of revision and to estimate item difficulties. It also led to the following two changes in item characteristics. First, in addition to the four choices of word meaning for each test item, a fifth option was added which reads, "I DON'T KNOW THIS WORD" (hereinafter choice E). In addition, the threat of a penalty for wrongly answered items was specified in the test instructions. (Example items are provided in Figures 1 and 2.)

bias: Be careful of **bias** in your writing.

a. grammar mistakes
b. language that is not exact
c. unfair opinions
d. informal language

e. I DON'T KNOW THIS WORD.

Figure 1. Example Text Item

Figure 2. Example Picture Item

Nation (2012) has stated a preference for not using penalties or the *I don't know* option but notes they may be justified when vocabulary tests are used for "proficiency-related decision-making" (p. 13). We introduced these conventions to address the likelihood that scores were being inflated by guessing. Even with explicit directions to skip unknown words, most examinees had far more wrongly answered than skipped items, which suggested that they were guessing. A comparison of data collected before and after these changes revealed a significantly reduced ratio of wrongly answered to skipped items (Bennett & Stoeckel, 2012) and an improvement in Rasch person reliability from .86 to .92. These results are indicative of more accurate estimates of vocabulary knowledge.

Test Form Development

The initial item difficulty estimates obtained during piloting were the basis for distributing the 240 items across four test forms of equal length. Because these estimates came from a small sample, an effort was also made to balance the four forms for parts of speech and for English loanword status in the Japanese language, two variables associated with word difficulty (Daulton, 2008; Milton, 2009). This resulted in test forms A, B, C and D, each of which consists of three 20-item sections to assess knowledge of the GSL1, GSL2, and the AWL. For the purposes of item calibration and test form equating, these test forms were revised by taking some items from their original form and sharing them across the other forms to act as anchors. The end result was four 90-item forms with 30 items at each level.

Field-Testing

The four versions of the instrument were then field-tested and assessed for item quality, test reliability, and test form equivalence under the Rasch measurement model.

Method

A convenience sample of 334 native speakers of Japanese from 21 intact classes at two universities in Japan (university A: n = 205 [137 women, 68 men; TOEIC data unavailable], university B: n = 129 [77 women, 52 men; TOEIC mean = 408.7, SD = 130.5]) participated in this phase of test development. The four 90-item test forms were spiraled in each class section. The data was analyzed with Winsteps software (version 3.72.2). The quality of the links within and between each test form was assessed and found to be satisfactory. Items were then simultaneously calibrated using the Rasch dichotomous model. These item calibrations were used in four separate analyses for converting raw scores to Rasch person measures for each of the test forms.

Results

A preliminary examination of the data revealed satisfactory person fit, item fit, and dimensionality. Item quality was assessed by inspecting point measure correlations and Rasch item fit indices. Four items were flagged as misfitting the Rasch model: *GSL1 include*, *GSL1 offer*, *GSL2 pale*, and *AWL transform*. Inspection of these items revealed ambiguity or grammatical complexity in the wording of the questions. The original item for *AWL transform* is given in Figure 3 as an example.

transform: The **transformation** of the town has had a big effect.

a. terrible damage
b. money that has been spent
c. people arriving from other countries
d. a complete change

e. I DON'T KNOW THIS WORD.

Figure 3. Original Test Item for *AWL transform*

The sentence stem and the four answer choices all contain modified noun phrases, which may have added unnecessarily to item difficulty. Another possibility is that the use of the indefinite article *a* in choice d confused respondents because the definite article *the* is already in the item stem. In addition, all four of the answer choices could constitute examples of transformation. As a consequence, this item was revised as shown in Figure 4. Here, less complex language has been used, and the distractors, while plausible replacements for *transformation* in the sentence stem, are not themselves examples of transformation. The other misfitting items have also been revised and all of these items will be monitored in future test administrations. The remaining 236 items appear to have good technical quality.

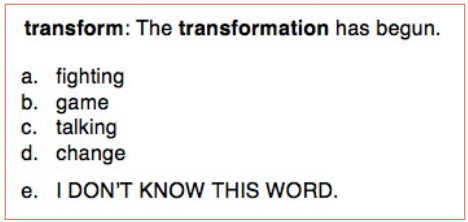

Figure 4. Revised Test Item for *AWL Transform*

Test reliability was assessed by inspecting Rasch person reliability estimates for each test form. Person reliability is an indication of person measure-order reproducibility and is similar conceptually to Cronbach's alpha. Reliability estimates ranged from .92 to .95 for the four 90-item forms and from .87 to .93 with the anchor items removed, indicating that all test versions had acceptable internal consistency (see Table 1).

To assess the relative difficulty of the 60-item forms, Rasch person measures for each possible raw score were compared across the four tests. Partial results are shown in Table 2. At any given raw score, Rasch person measures are within one standard error (*SE*) of each other. However, it is clear that, whereas Forms A and C are nearly identical, Form B is somewhat more difficult (indicated by lower person measures), and Form D somewhat easier. When comparing any person measure from Form B with its closest equivalent on Form D, the difference is about 3 points.

Table 1. Rasch Person Reliability Estimates

Test form	90-item version			60-item version (no anchors)		
	Person reliability	Mean	*SD*	Person reliability	Mean	*SD*
A	.93	57.1	13.6	.89	38.7	8.9
B	.92	60.2	12.8	.87	41.2	8.2
C	.92	61.4	12.5	.88	41.3	8.2
D	.95	57.0	16.7	.93	38.3	10.4

Table 2. Comparison of Raw Scores With Person Measures Across Four Forms

Raw score	Rasch person measure (*SE*)			
	Test form			
	A	B	C	D
37	.82 (.33)	.58 (.34)	.78 (.34)	.90 (.34)
38	.94 (.34)	.69 (.34)	.89 (.34)	1.01 (.34)
39	1.05 (.34)	.81 (.34)	1.01 (.34)	1.13 (.35)
40	1.16 (.34)	.93 (.35)	1.12 (.34)	1.25 (.35)
41	1.28 (.34)	1.05 (.35)	1.24 (.34)	1.37 (.35)
42	1.40 (.35)	1.17 (.35)	1.35 (.34)	1.50 (.36)

Discussion

A primary goal of this project was to develop equivalent test forms so as to avoid the possibility of a testing effect when repeatedly assessing vocabulary growth in a program of study. The instrument displays good item quality and overall reliability, but there are several issues in need of further review.

First, in light of the differences in test form difficulties and our preference for reporting raw scores, a redistribution of items among test forms is necessary to more closely approximate test equivalency. Because Rasch analysis provides difficulty estimates for each item, this is a relatively uncomplicated procedure. However, the stability and precision of item calibrations should first be explored with a larger, more representative sample.

Second, guessing in multiple-choice test formats can inflate estimates of vocabulary knowledge (Milton, 2009; Schmitt, 2010; Stewart & White, 2011). When a vocabulary test is appropriately designed for its intended population, examinees will encounter unknown words, and if these are not accounted for in the scoring rubric, estimates of vocabulary knowledge will be inaccurate. Even with the addition of choice E in our test, some examinees continued to have a rather high proportion of wrongly answered to skipped items (Bennett & Stoeckel, 2012), implying that some correct items were unknown but answered correctly by chance. Because this directly relates to a large body of research on the vocabulary sizes required to accomplish certain tasks in a foreign language (e.g., Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006), further studies are required in this area.

A third issue is the functionality of picture items. The rationale for such items was that some target words could not easily be defined with a limited vocabulary. Though analysis has not flagged any of these items as misfitting the Rasch model, the mental processes involved in answering the two formats are likely to be different, and, as such, there is merit in investigating in greater detail the use of picture items.

A final concern is that items in the GSL1 word band are defined with words from the same frequency band in the current test format. It was expected that the target population would know the vast majority of these words, but results of field-testing demonstrated that this was not the case. Bilingual tests may be a solution to this problem because they would eliminate the difficulty of respondents having to read the answer choices in the L2. However, care must be taken in score interpretation because research has shown that examinees score higher on bilingual tests (Ruegg, 2007).

Although these questions should be addressed, the test in its current format appears to be a useful tool for assessment of threshold vocabulary in Japanese academic contexts. For repeated testing, forms A and C were found to be of approximately equivalent difficulty for this sample, and could be treated as such for low-stakes purposes. This test adds to the instruments currently available to language instructors in that it allows for repeated testing without the risk of a testing effect and enables informed, reliable feedback on each of the word bands that are essential for learners in academic settings. The four test forms are available from either of the authors.

Acknowledgements

The authors wish to express their gratitude to Jeffrey Stewart for his thoughtful feedback on the manuscript.

Bio Data

Phil Bennett is a lecturer at Miyazaki International College. He is interested in all aspects of lexical development, with a current focus on acquisition of metaphorical language. <pbennett@sky.miyazaki-mic.ac.jp>

Tim Stoeckel teaches at Miyazaki International College. His interests include vocabulary teaching and learning and language testing. <tstoecke@sky.miyazaki-mic.ac.jp>

References

Aizawa, K. (2006). Rethinking frequency markers for English-Japanese dictionaries. In M. Murata, K. Minamide, Y. Tono, & S. Ishikawa (Eds.), *English lexicography in Japan* (pp. 108-119). Tokyo: Taishukan-shoten.

Aizawa, K., Ishikawa, S., & Murata, T. (2005). JACET 8000 eitango [JACET 8000 Word List]. Tokyo: Kirihara-shoten.

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*, 253-279. doi:10.1093/ijl/6.4.253

Bauman, J., & Culligan, B. (1995). *About the General Service List*. Retrieved from http://jbauman.com/aboutgsl.html

Bennett, P., & Stoeckel, T. (2012). Variations in format and willingness to skip items in a multiple-choice vocabulary test. *Vocabulary Education and Research Bulletin, 1*(2), 2-3.

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

Brown, D. (2012). The frequency model of vocabulary learning and Japanese learners. *Vocabulary Learning and Instruction, 1*(1), 20-28.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213-238. doi:10.2307/3587951

Daulton, F. E. (2008). *Japan's built-in lexicon of English-based loanwords*. Clevedon, UK: Multilingual Matters.

Dörnyei, Z. (2001). *Teaching and researching motivation*. Harlow, UK: Pearson.

Hancioğlu, N., Neufeld, S., & Eldridge, J. (2008). Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes, 27*, 459-479. doi:10.1016/j.esp.2008.08.001

Henning, G. (1991). *A study of the effects of contextualization and familiarization on responses to the TOEFL vocabulary test items*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*, 15-30. Retrieved from http://nflrc.hawaii.edu/rfl/

Meara, P. (1992). *EFL vocabulary tests* (1st ed.). Swansea, UK: Centre for Applied Language Studies, University College Swansea.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*, 142-154. doi:10.1177/026553228700400202

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi:10.1037//0003-066X.50.9.741

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

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners. *Canadian Modern Language Review, 63*, 127-147. doi:10.1353/cml.2006.0048

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacon-Beltran, C. Abello-Contesse, & M. Torreblanca-Lopez (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83-98). Bristol, UK: Multilingual Matters.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines, 5*(1) 12-25.

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

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*, 59-82. doi:10.3138/cmlr.63.1.59

Nation, I. S. P. (2012, August). *Measuring vocabulary size in an uncommonly taught language*. Paper presented at the International Conference on Language Proficiency Testing in the Less Commonly Taught Languages, Bangkok, Thailand. Retrieved from http://www.sti.chula.ac.th/files/conference%20file/doc/paul%20nation.pdf

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.

Petersen, N. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales: Statistics for social and behavioral sciences* (pp. 59-72). New York: Springer New York.

Qian, D. D. (2008). From single words to passages: Contextual effects on predictive power of vocabulary measures for assessing reading performance. *Language Assessment Quarterly, 5*, 1-19.

Ruegg, R. (2007). The English vocabulary level of Japanese junior high school students. In K. Bradford Watts, T. Muller, & M. Swanson (Eds.), *JALT2007 Conference Proceedings*, 103-109. Tokyo: JALT.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke, UK: Palgrave Macmillan. doi:10.1057/9780230293977

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*, 55-88.

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

Stewart, J., & White, D. A. (2011). Estimating guessing effects on the vocabulary levels test for differing degrees of word knowledge. *TESOL Quarterly*, *45*, 370-380.

West, M. (1953). *A general service list of English words*. London: Longman.