A comparison of test scores between monolingual and bilingual versions of the *Vocabulary Size Test*: A pilot study

Jeffrey Stewart
Kyushu Sangyo University

Reference data:

Stewart, J. (2009). A comparison of test scores between monolingual and bilingual versions of the Vocabulary Size Test: A pilot study. In A. M. Stoke (Ed.), *JALT2008 conference proceedings*. Tokyo: JALT.

There are high correlations between vocabulary knowledge and language proficiency (Zimmerman, 2004), and a need for tests that can assess learners' vocabulary size. The Vocabulary Size Test (Beglar & Nation, 2007) is a monolingual 40-question test used to determine vocabulary size. To evaluate the effectiveness of the test, the author tested subjects with the original monolingual version of the test and an otherwise identical bilingual version of the test. Nine intact classes of first and second year Japanese university students were tested (N = 154). Participants were sorted into three proficiency groups (low, mid, and high) based on TOEIC Bridge Scores. Data derived from the study thus far provide grounds for the belief that bilingual tests could produce more accurate scores than monolingual tests for lower-level learners.

語彙の知識と言語能力の間には高い相関関係があり(Zimmerman, 2004)、学習者の語彙サイズを評価できるテストが必要である。ボキャブラリーサイズテスト(Beglar & Nation, 2007) は、語彙サイズを判断するた めに使用される40題の単一言語テストである。単一言語テストの効果を判断するために、筆者は単一言語テストとそれと全く同一内容のバイリンガル版テストで問題を検証している。テストを行っていない日本の大学の1, 2年生、9クラスの学生にテストを行った(N = 154)。学生はTOEIC Bridgeのスコアによって3つのグループ（初級、中級、上級）に分けた。この研究により解明されたデータはこれまで、バイリンガル版テストが初級レベルの学習者のための単一言語テストよりもより正確なスコアを得ることができるという確信の理由となっている。

Despite vocabulary receiving relatively little focus in ESL programs (Folse, 2004), there are high correlations between vocabulary knowledge and language proficiency (Zimmerman, 2004), and a need for tests that can assess learners' vocabulary size. This paper details the results of pilot research done to evaluate differences in scores between the original, monolingual version of the Vocabulary Size Test (Beglar & Nation, 2007), and a bilingual version, which provides answers and distractors in the learner's first language (L1), under the hypothesis that this alternative test version would provide a more accurate description of test-takers' vocabulary sizes, particularly in the case of low-level learners.

In 2007, Nation and Beglar released the Vocabulary Size Test to "provide a reliable, accurate, and comprehensive measure of a learner's vocabulary size from the 1st 1,000 to the 14th 1,000 word families in English" (Beglar & Nation, 2007, p. 9). The test is not a diagnostic measure such as the Vocabulary Levels Test (Nation, 1990), and does not attempt to measure which wordlists learners should focus their efforts on. Rather, it is a proficiency measure used to determine a learner's overall receptive vocabulary size (Beglar & Nation, 2007). The test draws items from the 14,000 most common word families in English, with ten multiple choice questions from each word level, drawn from the General Service Word List (West, 1953; Beglar & Nation, 2007). Each word is accompanied by a simple non-defining sentence followed by four English definitions as possible answers.

Defining common vocabulary

It has been noted that it is difficult to create definitions for the first 1,000 most common words in English, because there are no simpler, more common words with which to describe them (Read, 2000; Beglar & Nation, 2007). For the first and second most common word levels of the Beglar-Nation Vocabulary Size Test, only words from the first 1,000 of the General Service List (West, 1953) were used. Producing a test of English vocabulary using English definitions of the same words has traditionally been difficult for the most common words in English (Read, 2000). For example, many words such as "now" cannot be defined using lower-frequency words. Therefore, these words can be the most difficult to define for testing purposes.

Effects of a bilingual Vocabulary Size Test

The central question of this study is whether results would differ if students were given another version of the test with L1 words as possible answers. The results could be useful for a variety of reasons.

Determining effectiveness of the monolingual test on the first two 1,000 word families

If the results do not differ significantly between the two versions of the test, it could serve as an indicator that the monolingual version is suitable for use even with low-level learners.

Determining how scores differ by word level

It is widely believed that monolingual definitions of the most common words in English present difficulties for learners, but that these differences become progressively less significant for lower frequency words. One goal of this experiment was to attempt to discover to what degree this is accurate, and at what word levels. Determining at what levels the differences are greatest between proficiency levels could be useful for future vocabulary test writers.

Predicting differences in scores

Even if the results differ significantly, the monolingual test could still be an effective diagnostic tool. If the gap between test scores is found to be consistent, then the point increase can simply be calculated into the score of the monolingual test.

Determining differences between low-level students

Average score increases could vary depending on student level. For example, high level students could experience little difference between monolingual and bilingual versions, but lower level students could experience greater, and less predictable, increases. If this proved to be the case, there is a possibility that a bilingual version of the test could be a more accurate diagnostic tool for lower-level students.

The Vocabulary Size Test's purpose is to measure receptive vocabulary size, yet identifying an all-English definition of a word, arguably requires more vocabulary power and proficiency than identifying an L1 definition (Laufer & Goldstein, 2004). However, studies have shown that correlations between all L1 vocabulary tests and passive L1 to L2 vocabulary tests and academic achievement are reasonably close, and that passive knowledge does have value, with a correlation of 0.63 between grades and scores on passive recall, and a correlation of 0.49 for passive recognition (Laufer & Goldstein, 2004). In the case of low-level learners, then, there could be a possibility that passive recognition could be the most revealing way for students to display L2 vocabulary knowledge. A bilingual version of the Vocabulary Size Test using L1 definitions would be a measure of this.

Methodology

For the bilingual version, monolingual definitions were translated into single L1 words, or the most basic units of expression. An important point is that the distractors were translated as the words they described, rather than as translations of the descriptions themselves. For example, the Vocabulary Size Test lists the following definitions for the word "where": "at what time", "for what reason, "to what place" and "in what way" (Beglar & Nation, 2007). For the bilingual version, these definitions were simply replaced with the direct Japanese equivalents for "when", "why", "where" and "how", rather than more complex, multiple-word translation of the definition.

The reasoning for this was that while understanding of the definitions could serve as an indirect indicator of vocabulary power, L2 definitions of L2 words are essentially a necessity of monolingual testing, not a goal, and while they could have indirect benefits, ambiguities they present are seen as a disadvantage and challenge for the testing format rather than an essential feature. A possible advantage of bilingual testing is that it eliminates this necessity.

Initial translations were done by the author, and then checked by a native Japanese speaker with a professional background in English-Japanese proofreading and translation.

Selection of test items

The full Vocabulary Size Test has 140 questions, with ten items from each 1,000 word family. Since the challenges of monolingual vocabulary testing are generally considered to occur most often with the most frequent word families, it was believed that examination of these levels would give the most important results. Therefore, the first 1,000 word family was split into two 10-question sections, the first 500 and second 500. Two other 10-question sections were drawn from the second and third 1,000 word families. To ensure an adequate number of test items, questions were selected from three versions of the test, provided by the authors.

Two versions of the test were prepared, one in its original monolingual form, and an identical test with L1 translations of each distractor included. To ensure that the tests versions were as close as possible, the original English definitions from the original version were left unchanged.

Student levels

An important variable in the study was student level, which was appraised by scores on the TOEIC Bridge Test. Designed by Educational Testing Services (ETS), the test is a measure of emerging English language competencies. It is a 60-minute paper and pencil test with two main sections, listening and reading, and features questions that are significantly easier than questions on ETS's TOEIC Test**.** Essentially, it is used to give a finer measure of ability for students that would score so low on the TOEIC Test that the information would be of little use for evaluations.

Since it has been demonstrated that vocabulary size is a strong predictor of English proficiency (Zimmerman, 2004), differences between scores on monolingual and bilingual versions of the test were tracked by the group's TOEIC Bridge scores, under the assumption correlations between these scores and scores on the monolingual and bilingual versions of the test could demonstrate which version is a more appropriate measure of proficiency.

One question of concern was that even if scores on the bilingual version of the Vocabulary Size Test correlated with the TOEIC Bridge, this may simply demonstrate that success on the bilingual size test simply correlates with another test. However, the TOEIC Bridge Test and the Vocabulary Size Test are markedly different measures of English proficiency. The TOEIC Bridge Test is entirely in English, and therefore does not measure passive L2-L1 vocabulary knowledge.

Participants came from nine intact classes (N = 154), and sorted by TOEIC Bridge scores as described in Table 1. The sorting reflects class levels at the participants' university, and is essentially a convenience sample for the purposes of this paper.

Table 1. Participants by proficiency level

Test score	Proficiency level	Number
TOEIC Bridge 90 - 98	Low	76
TOEIC Bridge 108 - 118	Mid	53
TOEIC Bridge 140+	High	25
	Total	154

It should be noted that the final group is comprised of students that scored over 140 on the TOEIC Bridge Test, and therefore a significant gap may exist between them and the group below. In the cases of students that scored near 160, it could be said that the TOEIC Bridge Test is no longer an accurate measure of their more advanced ability, and that the original TOEIC Test would be a better measurement of their ability. According to ETS, students with scores ranging from 140-160 on the TOEIC Bridge can be estimated to have TOEIC scores of roughly 395-570.

Procedure

Each class was given the monolingual version of the test first. After completion, all students were then given the same test with L1 translations. Students were given as much time as needed to complete the tests, with an average of approximately fifty minutes. One concern was that if the same student wrote both versions of the test, completion of the first version could have an effect on the score on the second. If either version of the test included problems that could take time to solve, this could be a consideration, because the first test could "prime" students, who may solve an item answer only after the first test is completed, and they see the same question again. However, the Vocabulary Size Test is intended as a measure of vocabulary knowledge, not problem-solving ability, and therefore if students do not have an opportunity to learn the words between writing the tests, there is arguably little advantage in facing the monolingual version prior to writing the bilingual version. There is a possibility that the intended meaning of an English definition on the monolingual version may not strike a student until

Table 2. Comparison of monolingual and bilingual versions of the Vocabulary Size Test

Student Proficiency									
		Low			Mid			High	
Word level	Monolingual	Bilingual	Change	Monolingual	Bilingual	Change	Monolingual	Bilingual	Change
0-500	6.73	8.69	1.96	7.33	9.2	1.87	9.08	9.75	0.68
501-1,000	5.73	7.59	1.86	7.09	8.4	1.31	9.04	9.24	0.2
1001-2000	3.09	4.73	1.64	3.38	5.42	2.04	7.2	7.44	0.24
2001-3000	3.34	4.44	1.10	3.64	4.87	1.23	5.96	4.68	-1.28
Total Score	18.88	25.43	6.55	21.41	27.75	6.34	31.28	31.12	-0.16
Note: k = 40									

after the first test is completed, since reading and interpreting an L2 definition arguably requires a more complex level of L2 proficiency. However, the purpose of the bilingual version was to eliminate the need for these definitions, and simply give a measure of passive receptive L2-L1 vocabulary knowledge.

Results and discussion

Paired-samples T-tests were calculated to compare students' scores on the monolingual tests and the bilingual versions. The results indicated that for the low group, with TOEIC Bridge scores of 90-98, the mean for the bilingual tests (M = 25.4342, SD = 3.6454) was significantly greater than the mean for monolingual versions of the test (M = 18.8816, SD = 4.40293), -14.147(75) = 0.510, p < .000. The mid group saw similar results, with the mean for the bilingual tests (M = 27.75, SD = 3.24) significantly greater than the mean for the monolingual version (M = 21.41, SD = 4.46), -10.421(52) = .373, p < .006. Only the high group did not see a gain between the monolingual test (M = 31.2800, SD = 5.42771) and the bilingual test (M = 31.12, SD = 3.05941), .155(24) = .369, p < 0.69.

Interestingly, the lower the level of the group, the greater the gains in test score. The low group saw an average increase of approximately 6.55 points. The mid group saw a similar, slightly lower gain of 6.33 points. The high group, however, saw no significant improvement at all, and in fact actually saw slightly lower scores (by 0.16 points). Unlike the lower-level groups, due to the smaller sample size, it is unconfirmed if this difference is statistically significant.

In all cases, even with the highest-level group that saw no significant improvement to mean total score, the standard deviations and standard error means were smaller for the bilingual version than for the monolingual version, which suggests that the bilingual version may be more reliable.

Differences by word level

Total differences in scores do not tell the whole story, however. To understand the differences between the test versions, we must also look at differences by test section and word level. Table 2 describes the mean scores for the monolingual and bilingual test versions, sorted by student level and word level. Individual sections are scored out of ten; the total is out of forty.

The same information divided into monolingual test scores and bilingual test scores can be seen diagrammatically represented in Figure 1 and Figure 2 respectively.

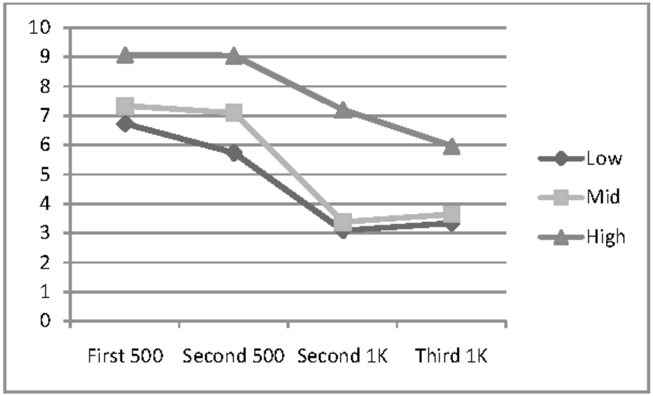

Figure 1. Monolingual test scores

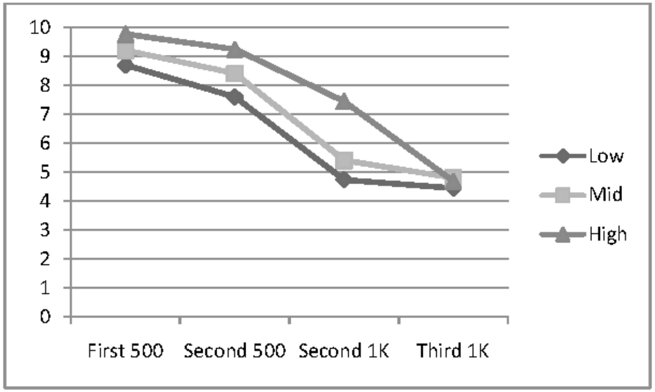

Figure 2. Bilingual test scores

Differences in total score

In regard to total score, the lower the TOEIC Bridge score, the greater the increase on the bilingual test. However, improvements in scores on individual test sections (the first 500 most common words, second 500 most common, second 1,000 most common, and third 1,000 most common respectively) differed by student level. The greatest gains for any group occurred not for the first 1,000, but for the second 1,000 words.

For the first 1,000 words, the low-level group saw the largest increases. For the second thousand, the mid-level group saw the largest increase, both in terms of raw point increase and increase as a percentage of their original score. For the third 1,000 words, the two groups were very similar, with a difference of only 0.08. The bilingual test had only slight effect on the high-level group.

Conclusion

As hypothesized, the bilingual version had a substantial and statistically significant effect on test scores. Perhaps the most striking difference in the data, was the gap between the score differences of the high group and the low and mid groups; while the two lower-level groups saw sizeable improvements in scores, the highest level group saw almost no difference in scores for the first three sections, and even saw a decrease in scores for the third 1,000 words section.

This indicates that there could be a threshold of proficiency at which L1 translations no longer significantly affect students' results. Conversely, for lower-level students, it appears that the translations have notable consequences on scores.

Data derived from the 154 students that have participated in the study thus far provide ample grounds for the suggestion that bilingual tests could give more accurate scores for lower level learners than monolingual tests. While the bilingual version of the Vocabulary Size Test yields only small differences with higher proficiency learners, lower proficiency learners see differences that are both sizeable and statistically significant.

Future research will focus on attaining greater sample sizes of learners with TOEIC Bridge scores higher than 140, to confirm the score differences of the current sample. Larger sample sizes of students with identical scores could confirm that the bilingual tests are in fact more useful for differentiating between low-level students. Currently the lowest-level learners tested have had TOEIC Bridge scores of 90. Future research will examine score differences of lower level students, and determine whether or not gaps in scores vary with level. On the upper threshold, students with incrementally higher TOEIC Bridge scores will be tested, to determine at precisely what level bilingual tests cease to make a statistically significant difference in test scores, and at what level monolingual tests cease to have differences that could be problematic for educators.

Jeffrey Stewart is a lecturer at Kyushu Sangyo University in Fukuoka.

References

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

Educational Testing Services (ETS), TOEIC Bridge and TOEIC score comparisons. Retrieved January 13, 2008, from <http://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC-TOEIC_Bridge_Score_Comparisons.pdf>.

Folse, K. (2004). *Vocabulary myths*. Ann Arbor: The University of Michigan Press.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399-436.

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

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

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