Vocabulary size, TOEIC scores, and *testwiseness*

Masaya Kanzaki
Kanda Institute of Foreign Languages

Reference data:
Kanzaki, M. (2010). Vocabulary size, TOEIC scores, and *testwiseness*. In A. M. Stoke (Ed.), *JALT2009 Conference Proceedings*. Tokyo: JALT.

The relationship between vocabulary and English proficiency tests was investigated among 31 learners. They were given the Vocabulary Levels Test (120 questions) and a TOEIC practice test (200 questions), and their scores were compared. The correlation between the two tests was .64. However, when the subjects were divided into two groups based on the number of times they had taken the TOEIC, the figure for the expert group (***n*** = 15, exam taken = 8 times or more) was .76 and that for the novice group (***n*** = 16, exam taken = 6 times or fewer) was .49. This paper examines the difference between the two groups in relation to ***testwiseness***, a respondent's capacity to use knowledge about a test itself, in spite of what the test is supposed to measure, in order to gain a high score.

31人の学習者の語彙テストと英語能力試験の関係を調査した。彼らに語彙レベルテスト（120問）とTOEIC練習テスト（200問）を与え、点数を比較した。2種類のテスト間の相関係数は0.64であった。しかし、被験者をTOEICの受験回数を基準に2つのグループに分けると、熟練者グループ（被験者数15、受験回数8回以上）における数値は0.76、初心者グループ（被験者数16、受験回数6回以下）では、0.49となった。本研究は2グループ間の差異を「受験技術力」（テストが測定しようとしている能力に関係のない、高得点を取るためテスト自体に関する知識を利用する能力）との関係において考察する。

Given the huge amount of English vocabulary, EFL learners have to learn a lot of new words when acquiring proficiency in the language. The development of a learner's overall English ability often correlates with growth of vocabulary knowledge, and vocabulary size is reflected in a learner's English proficiency test performance. Nation and Meara (2002) observed that "there is a relatively close relationship between how many words you know, as measured on the standard vocabulary tests, and how well you perform on reading tests, listening tests and other formal tests of your English ability" (p. 50).

Several studies have explored this close relationship. Qian (1999) compared the Vocabulary Levels Test scores with the TOEFL reading comprehension section scores of 74 ESL students in Canada and found that the two tests correlated well at .78. Similarly, Beglar and Hunt (1999) compared the 2000 Word Level Test scores with the TOEFL scores of 496 EFL students in Japan, and the correlation between the two tests was .71.

These results agree with what is often observed in the classroom. When students cannot answer practice test questions correctly, it is often a lack of vocabulary hindering their success.

This is particularly noticeable among those preparing for the Test of English for International Communication (TOEIC). The purpose of this study is to investigate how closely vocabulary size and TOEIC scores are related.

In addition, this study interpreted the results in relation to *testwiseness*, which Millman, Bishop, and Ebel (1965) defined as:

A subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score. Test-wiseness is logically independent of the examinee's knowledge of the subject matter for which the items are supposedly measures. (p. 707)

Since this study investigated *testwiseness* in regard to the TOEIC, the term used in this paper refers to the capacity to use any ability, knowledge, or skills to improve a TOEIC score that do not include English ability, which the exam is designed to measure. This includes how familiar test takers are with the test format and question types, how well they manage time, and how well they can guess answers, all of which improve with test taking experience. To investigate the effect of *testwiseness*, vocabulary and TOEIC scores were examined in relation to the numbers of times the participants had taken the TOEIC, on the assumption that those who have taken the test many times have higher levels of *testwiseness* than those with less test taking experience.

Table 1. TOEIC score distribution

TOEIC Score	Under 500	500–545	550–595	600–645	650–695	700–745	750–795	800–845	850–895	900–990
N	1	3	0	5	4	8	2	2	4	1

Note: Based on highest scores. One of the participants had never taken the TOEIC.

Table 2. Numbers of TOEIC tests participants had taken previously

TOEIC taken	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16+
N	1	2	4	0	2	5	2	0	2	1	7	2	0	0	0	1	2

Table 3. Age distribution

Age group	20–29	30–39	40–49	50–59	60–69	70+
N	5	12	7	3	0	1

Note: Three of the participants did not disclose their ages.

Method

A vocabulary test and a TOEIC practice test were given to 31 participants, and the scores of the two tests were compared.

Participants

Participants were 31 adult learners who took part in a four-day intensive TOEIC preparatory course at a private English school in Tokyo in August 2007. Of the participants, 20 were male, 11 were female, and all of them were Japanese speakers. The group was widely varied in terms of TOEIC scores (Table 1), numbers of TOEIC tests taken (Table 2), and age (Table 3).

Materials

As a measure of vocabulary size, four levels of the Vocabulary Levels Test were used (30 questions per level, 120 questions in total). As a substitute for the actual TOEIC, 200 questions from a practice test book were used.

Vocabulary Levels Test

In this study, Schmitt's (2000) version of the Vocabulary Levels Test Version 1 was used, which is included in Paul Nation's vocabulary resource booklet available on his website.

The Vocabulary Levels Test consists of five levels, namely the 2000, 3000, 5000, 10000, and academic word levels. The levels are based on frequency counts, except the academic word level, which is based on Coxhead's (2000) Academic Word List. In this study, the 10000 level was excluded because the words tested in that level were considered too difficult for the participants.

Each level of the test has 10 clusters of 6 words with 3 definitions, which makes a total of 30 questions per level. An example of a cluster is shown below.

1. business
2. clock _____ part of a house
3. horse _____ animal with four legs
4. pencil _____ something used for writing
5. shoe
6. wall

Test takers are instructed to match the words on the left with the definitions on the right. Schmitt, Schmitt, and Clapham (2001) reported the Cronbach's alpha reliability values for the 2000, 3000, 5000, and academic word levels of Version 1 as .920, .929, .927, and .958, respectively. Therefore, this vocabulary test is reliable.

TOEIC practice test

Questions from *TOEIC Shin Koshiki Mondaishu Vol. 2* (2006) [the New Official Practice Tests for the TOEIC Vol. 2] were used in this study. As the Educational Testing Service, the creator of the actual TOEIC, provided the materials, the two practice tests in the book are very close to the actual TOEIC in terms of vocabulary, question types, topic types, length, level of difficulty, and formatting of questions in the test book.

The TOEIC consists of 200 multiple-choice questions, half of which are in the listening section with the other half in the reading section. In the actual TOEIC, raw scores (0–200) are converted to scaled scores (10–990). In this study, only raw scores are used for calculations and analyses.

The TOEIC has seven parts; the first four parts are in the listening section and the last three parts are in the reading section. Details of each part are shown in Table 4.

Table 4. Seven parts of TOEIC

Section	Part	Task	# of Qs
Listening	1	For each question with photo, listen to four sentences and choose the one that best describes the image.	10
	2	Listen to a question or statement followed by three responses and choose the most appropriate response.	30
	3	Listen to a conversation and answer three questions.	30
	4	Listen to a short talk and answer three questions.	30
Reading	5	Choose a word or phrase to fill in a blank in a sentence.	40
	6	Choose a word or phrase to fill in three blanks in a passage.	12
	7	Read a passage or a set of two passages and answer two to five questions.	48

Procedures

Two practice tests with a total of 400 questions in the practice test book were used during the four-day intensive course. To provide the same number of questions on each day of the course, the two tests were divided into four half tests and the participants answered 100 questions each day (i.e., the first half of Test One on the first day, the second half of Test One on the second day, the first half of Test Two on the third day, and the second half of Test Two on the fourth day). In addition, the participants were given the Vocabulary Levels Test; the 2000 and 3000 levels were administered on the second day, and the 5000 and academic word levels were administered on the third day. Only those who attended both the second and the third days were included in this study. The results of the practice tests administered on the first and fourth days (i.e., the first half of Test One and the second half of Test Two) were excluded from this study because some of the participants who attended both the second and the third days missed either the first or the fourth day. The data collected on the second and the third days were compiled and descriptive statistics and Pearson product moment correlation coefficients were computed.

Results

Descriptive statistics

Table 5 shows the descriptive statistics for the Vocabulary Levels Test. The mean of the academic word level is between those of the 2000 and 3000 levels, which agrees with the results reported by Schmitt et al. (2001).

Table 5. Descriptive statistics for Vocabulary Levels Test (*N* = 31)

Level	2000	3000	5000	Academic	Total
K	30	30	30	30	120
M	27.58	24.42	21.29	25.48	98.77
Range	16	17	20	20	67
SD	3.29	4.24	5.20	3.71	14.57

Note: Total = all four levels combined

The Kuder-Richardson 21 reliability values for the listening section, the reading section, and the whole of the TOEIC practice test are .87, .87, and .92, respectively. The entire test shows high reliability, and the listening and reading sections show moderately high reliability. The descriptive statistics for these three sections are shown in Table 6.

Table 6. Descriptive statistics for TOEIC practice test (*N* = 31)

Section	Listening	Reading	Total
K	100	100	200
M	70.39	74.71	145.10
Range	49	48	83
SD	12.25	11.51	21.17
K-R21	.87	.87	.92
SEM	3.97	3.77	5.51

Note: Total = entire practice test

Correlation

Table 7 shows the correlation coefficients between the two tests. The reading section of the practice test correlated well with the vocabulary test at .72 with the 2000 level, .71 with the 3000 level, .67 with the 5000 level, .62 with the academic word level, and .76 with the total of the four levels. The figures for the listening section, however, were much lower at .28, .36, .42, .27, and .39, respectively. The correlation coefficient between the totals of the two tests was .64, which is statistically significant ($p < .01$).

Table 7. Correlations between TOEIC practice and vocabulary tests (*N* = 31)

Level	Listening	Reading	TOEIC-Total
2000	.28	.72	.55
3000	.36	.71	.61
5000	.42	.67	.61
Academic	.27	.62	.49
VLT-Total	.39	.76	.64

Note: TOEIC-Total = total of TOEIC practice test. VLT-Total = total of Vocabulary Levels Test.

Dividing participants into two groups

To investigate the effect of *testwiseness*, the participants were divided into two groups based on how many times they had previously taken the TOEIC. This was done with the assumption that the level of *testwiseness* increases with test taking experience. As shown in Table 2, 16 out of the 31 participants had taken the test 6 times or fewer and the other 15 had taken it 8 times or more. The data was divided into two categories, namely the novice group (TOEIC taken = 6 times or fewer) and the expert group (TOEIC taken = 8 times or more), and recalculated.

Table 8 shows a comparison of the vocabulary test results of the two groups. The means of the novice group are slightly higher than those of the expert group, except the 5000 level. The difference in the total scores is 1.25 out of 120.

Table 8. Comparison of vocabulary test results between two groups

Level	2000		3000		5000		Academic		Total	
	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert	Novice	Expert
M	27.75	27.40	25.00	23.80	20.94	21.69	25.69	25.27	99.38	98.13
Range	12	16	13	17	18	20	13	19	43	67
SD	2.73	3.79	3.95	4.45	5.24	5.13	3.23	4.14	12.91	16.13

Note: Novice = novice group (n = 16). Expert = expert group (n = 15).

Table 9. Comparison of TOEIC practice test results between two groups

Section	Listening		Reading		Total	
	Novice	Expert	Novice	Expert	Novice	Expert
M	68.88	72.00	75.56	73.80	144.44	145.80
Range	45	44	37	46	70	83
SD	12.45	11.82	10.73	12.23	20.07	22.27

Table 10. Comparison of correlations of two tests between two groups

Level	Listening		Reading		TOEIC-Total	
	Novice	Expert	Novice	Expert	Novice	Expert
2000	.00	.54	.48	.88	.25	.77
3000	.23	.53	.55	.84	.44	.74
5000	.37	.47	.64	.72	.57	.65
Academic	.01	.53	.50	.71	.27	.67
VLT-Total	.22	.56	.65	.85	.49	.76

Table 9 shows a comparison of the TOEIC practice test results of the two groups. Compared with the novice group, the expert group performed better in the listening section but worse in the reading section. The average total score of the expert group is 1.36 higher than that of the novice group.

Table 10 shows a comparison of the correlation coefficients of the two tests between the two groups. The correlations are distinctively higher among the expert group than the novice group with all the 15 pairs compared.

Discussion

Possible reasons for weak correlations

As shown in Table 7, the reading section of the practice test correlated well with the vocabulary test while the correlation between the listening section and the vocabulary test was weak. One possible reason for this is the Vocabulary Level Test measures vocabulary knowledge only in written form. Since the listening section requires test takers to recognize words in spoken form, an aspect of vocabulary knowledge not measured in the Vocabulary Levels Test is relevant in the listening section of the TOEIC.

Another possibility is that vocabulary size is less important in the listening section than in the reading section because the vocabulary load is lighter in the listening section. The vocabulary profile of the TOEIC practice test used in this study indicates that the coverage of high frequency words is higher in the listening section than in the reading section. Table 11 shows the cumulative coverage of the first five frequency levels of 1000 word bands.

Table 11. Cumulative coverage for listening and reading sections of TOEIC practice test

Word list	Listening	Reading
1000	86.56	74.40
2000	94.07	84.93
3000	95.89	88.63
4000	96.80	91.36
5000	97.14	92.60

Note: Figures are percentages. For vocabulary analysis, RANGE (Heatley, Nation, & Coxhead, 2002) was used with British National Corpus base lists.

In addition, "processing time to use conscious knowledge about vocabulary" (Beglar & Hunt, 1999, p. 149) may have influenced the correlation. Listening allows less processing time than reading, and therefore the participants may not have had enough time to fully utilize their vocabulary knowledge in the listening section.

Differences between novice and expert groups

Table 8 shows that the average vocabulary test score of the novice group is 1.25 higher than that of the expert group, and Table 9 shows that the average TOEIC practice test score of the expert group is 1.36 higher than that of the novice group. Despite the lower vocabulary test scores, the expert group had higher TOEIC practice test scores than the novice group, which can be attributed to the higher level of *testwiseness* the expert group attained through test taking experience.

The most notable difference is in the correlations between the vocabulary and TOEIC practice tests. As shown in Table 10, the correlations between the two tests are significantly higher among the expert group than the novice group. The high correlations among the expert group suggest that vocabulary size was an important factor affecting their TOEIC scores, whereas the low correlations among the novice group suggest that vocabulary size was not a determining factor for TOEIC scores among the group. This can be explained in relation to the levels of *testwiseness* among each group: the level of *testwiseness* was even among the expert group and varied widely among the novice group.

There are three possible reasons for the varied levels of *testwiseness* among the novice group. First, *testwiseness* increases not only through taking actual tests, but also through taking mock tests. This study only looked at how many times the participants had taken the TOEIC and did not consider how much they had studied in preparation for the test. Those who have done many practice tests are likely to be more *testwise* than those who have done only a few, even when the number of the TOEIC tests taken is the same. Second, *testwiseness* gained through experience taking other tests can be applied when taking the TOEIC. This study only looked at test taking experience of the TOEIC; therefore, it is possible that those who had gained a high degree of *testwiseness* from other tests were included in the novice group. Last, the rate of progress in regard to *testwiseness* may vary. *Testwiseness* is a type of ability; therefore, the rate at which people acquire it may differ from person to person. For example, some learners may need to take six tests to gain a certain degree of *testwiseness* while others only need to take three tests to reach the same level.

Similar varieties likely existed among the expert group. In other words, some had taken more practice tests than others, some had more test taking experience from having taken other tests, and some were more adept at acquiring *testwiseness*. Nevertheless, the level of *testwiseness* appears to be even despite the differences mentioned above. This suggests that there is a ceiling to how much learners can improve their *testwiseness*; those in the expert group had the same level of *testwiseness* because they had reached this *testwiseness* ceiling by the time they had taken the TOEIC eight times.

Testwiseness threshold

The *testwiseness* ceiling is a threshold for fair score comparison. Until learners have come to this ceiling, their TOEIC scores cannot be compared fairly. This is because those with low levels of *testwiseness* have an unfair disadvantage; their scores are likely to be lower than those with higher levels of *testwiseness*, even when English proficiency levels are the same. In other words, when learners have reached the *testwiseness* threshold, their TOEIC score begins to work as an indicator of their English proficiency.

Determining how many times learners need to take the TOEIC to reach the *testwiseness* threshold is not a simple task. Level of *testwiseness* is affected not only by how many times the TOEIC has been taken, but by the factors mentioned above, such as how much test preparation has been done, how much test taking experience through other tests has been acquired, and how capable the test taker is at gaining *testwiseness*. The results of this study show that those who have taken the TOEIC eight times or more have reached the threshold. However, this study does not take the other factors into consideration, and therefore it is still open to speculation whether taking the TOEIC eight times alone is enough to reach the threshold.

Conclusion

This study revealed that vocabulary size and TOEIC scores correlate to a moderate degree, as the correlation coefficient

between the Vocabulary Levels Test and the TOEIC practice test was .64. The reading section of the practice test correlated well with the vocabulary test at .76, and the listening section correlated poorly at .39.

However, higher correlations were obtained among those who had taken the TOEIC eight times or more. This suggests that they had reached a uniformly high level of *testwiseness* through test taking experience. The findings in this study also suggest that levels of *testwiseness* vary widely among those who do not have much test taking experience. There seems to be a *testwiseness* threshold after which learners can achieve TOEIC scores that accurately reflect their English proficiency. Those who have not reached that threshold are likely to get lower scores than their more *testwise* peers even when they have the same level of English proficiency.

The implication is that TOEIC scores of those who have not reached the *testwiseness* threshold cannot be compared fairly because the difference in *testwiseness* may distort the relationship between their TOEIC scores and English ability. To compare learners' English ability using the TOEIC, the effect of *testwiseness* should thus be considered.

Bio data

Masaya Kanzaki recently completed his Master's degree in TESOL at Temple University and is currently teaching at Kanda Institute of Foreign Languages. His research interests include vocabulary acquisition and language testing. <msyknzki@poplar.ocn.ne.jp>

References

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, *16*(2), 131-162.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). RANGE (version 32) [Computer program]. Retrieved January 12, 2007, from http://www.vuw.ac.nz/lals/staff/Paul_Nation

Institute for International Business Communication (2006). *TOEIC Shin Koshiki Mondaishu Vol. 2* [The new official practice tests for the TOEIC Vol. 2]. Tokyo: IIBC.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, *23*(3), 707-726.

Nation, I. S. P. (2005) Vocabulary resource booklet [Collection of files]. Retrieved February 14, 2007 from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx.

Nation, I. S. P., & Meara, P. (2002). Vocabulary. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 35-54). London: Hodder Arnold.

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*(2), 282-307.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge University Press.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing*, *18*(1), 55–88.