A Corpus-based Study on High-level Vocabulary in TEPS Vocabulary Section in Terms of Frequency Values

Joohee Shin

(Seoul National University)

Joohee, Shin. 2012. A Corpus-based Study on the High-level Vocabulary in TEPS Vocabulary Section in Terms of Frequency Value. *SNU Working Papers in English Linguistics and Language 11, 126-141.* This study aims to investigate whether TEPS vocabularies in questions number 46 through 50 in vocabulary part are frequently used by native speakers in daily life in terms of frequency value based on Corpus of Contemporary American English. The five questions have been criticized for being too difficult or too outdated to be used in daily life of native speakers. The result was that correct answers tended to have higher frequency value than that of distractors. Accordingly, these questions in controversy were proved to be valid for they satisfied the conditions of asking the answers of frequent words. The result of this study would contribute to settle the controversy over the validity of high-level vocabulary in TEPS. **(Seoul National University)**

1. Introduction

One of the most widely used tests in Korea is TEPS[1] (Test of English Proficiency developed by Seoul National University). Despite its wide use, there has been controversy about whether its vocabulary part is valid for assessing English proficiency. One argument is that vocabulary part is not a valid construct for assessing English proficiency. It is because a vocabulary ability is the one not to be assessed independently but to be a fundamental element of a reading

---

[1] TEPS is a test of English proficiency developed by Seoul National University, and it measures the ability to use English in real-life situations. It consists of four parts: listening comprehension, grammar, vocabulary, and reading comprehension.

skill. The argument is supported by the fact that TOEIC, TOEFL, and other English proficiency tests do not have a vocabulary part independently (Cho, 2000).

A controversy is also caused by the questions 46 through 50 in a vocabulary part which are considered the most difficult items. A vocabulary part consists of two sections; section 1 (questions 1 through 25) is to fill in the blanks in a conversation, and section 2 (questions 26 through 50) is to fill in the blanks in a statement. The controversial questions belong to section 2. It has been argued by TEPS instructors and examinees that these five items are not appropriate for assessing English proficiency because the options are too outdated or difficult to be used in general.

The present study attempts to investigate the second issue, which has not been addressed before: validity of the questions 46 through 50 in vocabulary section in terms of corpus data, assuming that more frequently used words have higher frequency on corpus data. Based on the research findings, this study also aims to figure out if there is any relation between correct answers and distractors.

2. Literature Review

It is firstly needed to review studies that treat validity of TEPS. It has been 15 years since TEPS was first taken. For a relatively short history, it has not been widely studied as much as TOEIC and TOEFL. It was Park et al. (1998[2]) who first argued validity of TEPS. He argued that dividing communicative language ability into four parts seems to be outdated. He recognized that it is not ideal, but he said it is most realistically appropriate because ETS and UCLES still do so. In other

---

[2] Park et al. (1998) was published one year before TEPS was first implemented.

words, he justified the TEPS construction of four because renowned institutions do so.

Concerning TEPS vocabulary section, Cho (2000) refuted Park et al. (1998) and argued that TOEFL does not assess vocabulary as an independent part, and this fact reflects that a vocabulary skill is not the one to assess independently but the one which is a fundamental element of a reading skill. Moreover, it is against the current trend of assessment that moves to an integrative test.

There is one study that mentioned the second part of TEPS vocabulary section. Choi (2008) stated that the second part of the vocabulary is more difficult yet non-technical words largely used in academic situations. This study does not provide any supporting evidence, so it is required to have evidence concerned.

Two other studies revealed the correlations between the TEPS score and GPA score of practical English subject to see whether TEPS score is a good predictor for determining success in college work as well as a measure of overall English proficiency (H. Lee & S. Lee, 2003). The participants were in the top 5 to 10 percent of CSAT (College Scholastic Ability Test) in Korea. Correlation between TEPS vocabulary score and GPA score was $r = .203$ which was lower among 12 correlations from the three tests. These figures showed that TEPS vocabulary part does not have a good predictive validity for the college work and is the least related to the practical English capacity.

The other study investigated the correlations between university students' test scores in TEPS, TOEIC, and TOEFL and their class performances in English conversation and composition courses (H. Lee & S. Lee, 2004). Their scores in English conversation and English composition courses were evaluated by native speakers. The result of the first sample study was analyzed in terms of skewness[3]. In their

---

[3] Skewness is a measure of the extent to which a probability distribution of a real-valued random variables "leans" to one side of the mean (Wikipedia). Minus skewness value represents negative distribution, and it implies that mean value is higher than that

sample study, skewness of total TEPS score was -11, and skewness of TEPS vocabulary part was +35. TEPS vocabulary part had the highest skewness value among all subparts of the three tests. It means that this part was the most difficult one. In the second sample study of it, correlation between TEPS vocabulary score and English conversation score was $r = .13$ which is low, and correlation between TEPS vocabulary score and English composition score was $r = .10$ which is low. These two figures were the second least and the forth least ones among 91 correlations. Through the last sample study, it was revealed that TEPS vocabulary score is hardly relevant to English proficiency in real-life.

Most of the previous researches raised a question of the assessing English proficiency of TEPS. The present paper specifically focused on whether high-level vocabulary in vocabulary part question number 46 through 50 is used in daily life in terms of frequency value to see the test validity. In addition, the assumption is that the higher frequency value is, the more a word is used in daily life. Taking an example, a word 'pneumoconiosis' is hardly used, and we cannot take it as a standard to decide one's English proficiency. Its frequency value is 33 according to COCA, so this paper uses the notion of a frequency value. Moreover, if there is any consistent tendency for the options, it is worth investigating any relation between correct answers and distractors.

So, the research questions can be the two;

1. Are so called 'difficult questions' of 46 through 50 in TEPS vocabulary section really used frequently in daily life among English native speakers?

2. Is there any relation between correct answers and distractors?

of standard normal distribution. Plus skewness value means positive distribution, and it implies that mean value is lower than that of standard normal distribution.

3. Method
3.1 Materials

Three sets of the latest sample questions from TEPS published by the TEPS Council of Seoul National University (서울대 텝스 관리위원회 제공 최신기출 시크릿, 2010) were used as sample tests. In addition, Corpus of Contemporary American English (COCA) was used to get frequency value of the words. This corpus contains 450 million words used from 1990 to 2012, so it may reflect recent trend in use of words.

3.2 Procedure

A preliminary study was fulfilled to confirm that item difficulty is low in questions 46 through 50 and to get a tentative value for the following studies. For this, 8 items were randomly selected in one test set. The median values of the options in 8 items were estimated . The present study did not employ a mean value but a median value, for there was much deviation among options.

   The first study was set to investigate frequency values of options in questions 46 through 50 in three sets. The purpose was to figure out whether the items consists of frequently used words of options. In other words, it aimed to confirm validity of TEPS as a communicative language test. Five questions from three sets of sample questions were investigated to figure out median values by means of COCA (total 15 questions and 60 choices). It was assumed that the higher a COCA value is, the more frequently a word is used in daily life. The results were compared to that of the preliminary study.

   The second study was to figure out any relation between correct answers and distractors. Four results could be predicted among. The first one is high frequency values in both correct answers and distractors. The second one is high frequency values in correct answers

and low frequency values in distractors. What examiners want to test is for examinees to know the target correct answer options, so it would be the most appropriate for the correct answer options to have the highest frequency value when assessing English proficiency. The third one is low frequency values in correct answers and high frequency values in distractors. The final one is low frequency values in both correct answers and distractors. It may result in that the last five questions estimate really high-level lexical competence, so it could be thought to hinder validity as a proficiency test.

The final study was to figure out any tendency of correct answers according to frequency values. If correct answers tend to belong to higher frequency value categories among four options, then it may imply that more frequently-used words tend to be correct answers. It means that within difficult questions though, examiners want examinees to know more frequently-used words among four options, so we cannot say that those questions are not valid.

4. Result

In the preliminary study, the results confirmed that the options from 2 randomly selected questions from 46 to 50 rated lower frequency values compared to 8 other question options from questions 1 to 45 based on the COCA data. Up to the question number 45, an average median value seems to be high. However, T1 and T2 shows far lower median values (see Table 1). Based on the result, the median value of R1 to R8 was approximately 6000. This figure was calculated to be used as a tentative value which determines whether the following words can be considered to be frequently used ones.

Table 1
Preliminary Research Result

Question type	Median value	Question Type	Median value
R1	27720	R6	40565
R2	10950	R7	1798
R3	1250	R8	4286
R4	2886	T1	507
R5	9241	T2	729

Note: R=randomly selected questions, T=target questions randomly selected from 46 to 50

In the first main study, frequency values of options of questions 46 through 50 in three sample tests were figured out (see Table 2). The obtained values of each choice were 20 in each set (an item consists of 4 choices). In contrast with the preliminary study in which median value was gained to get a brief figure, the main studies used exact frequency values. Compared to the tentative figure of 6000 in the preliminary study, set 1 showed far lower values. The bottom four frequency values were approximately below 300 (frequency values of 98, 248, 280, and 298). The most frequently used words of two were the value of 6735 and 5583 which imply that even in the difficult questions, relatively frequent words are included.

Table 2

Frequency Values of Options in Questions 46 through 50 in Set 1

Question	Option	Frequency Value
46	predisposition	630
	pervasive	298
	preeminence	280
	percolation	98
47	Saturated	5583

	Mired	1082
	Tackled	1072
	enamored	728
48	Relentless	3119
	salient	2045
	immaculate	1095
	Pernicious	658
49	alienate	780
	scrutinize	714
	deplore	301
	exonerate	248
50	appreciation	6735
	enlargement	842
	compound	616
	extravagance	508

Note: The shaded sections indicate correct answers.

Set 2 showed much deviation than set 1 (see Table 3). The interesting point was that three values of them(10199, 10100, 8208,) rated over 6000 that were very high. However, especially two of them rated very low frequencies of 13 and 39 which can be considered that these are the words hardly used in daily life. Looking at all the frequency values, it can be concluded that even though three of them had high frequency, most of them had far less frequency compared to the preliminary study result.

Table 3
Frequency Values of Options in Questions 46 through 50 in Set 2

Question	Option	Frequency Value
46	Grabbed	10100
	Committed	8208

	Opposed	5807
	Enclosed	470
47	Terrestrial	2006
	existential	1408
	Foundry	460
	Sublunary	13
48	Consummated	949
	Unequivocal	687
	Abreast	628
	errant	849
49	Decayed	2963
	Pivoted	971
	Plummeted	632
	Stratified	39
50	Discourses	10199
	Conundrums	721
	Quandaries	613
	Tantrums	596

Note: The shaded sections indicate correct answers.

Table 4 showed a similar trend to set 2. Four choices had more than 6000 frequency value, and others had relatively low frequency values compared to the result from the preliminary study. It showed the greatest deviation among the three sample tests. The interesting point was that choice "c" in question 47 had 0 frequency value, which means native speakers didn't use the word at all based on the corpus data. The words was 'antiquate.' The five words used the least frequently were frequency values of 0, 12, 91, 136, and 157.

Table 4
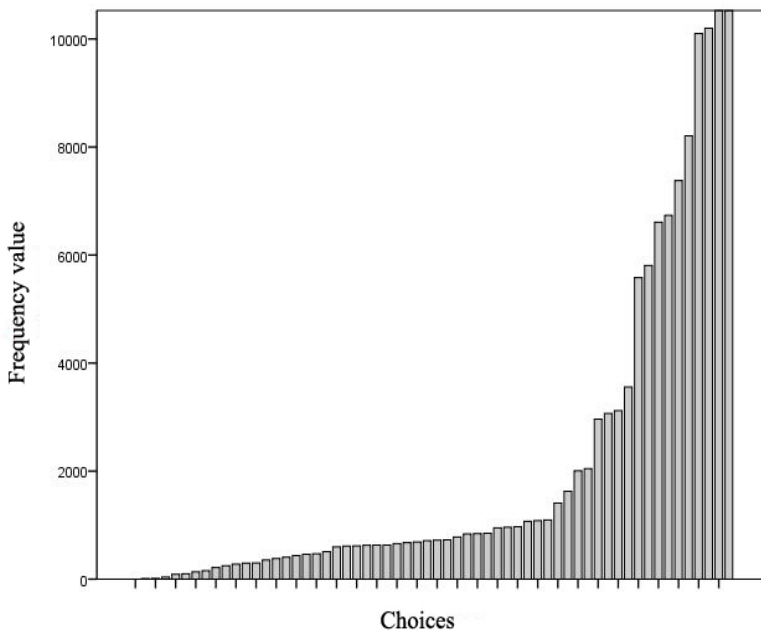Frequency Values of Options in Questions 46 through 50 in Set 3

Question	Option	Frequency Value
46	False	14999
	Blank	7380
	mock	3068
	invalid	962
47	Update	6608
	Aggravate	407
	Understate	215
	Antiquate	0
48	Proliferation	3556
	Annihilation	675
	Desecration	353
	Edification	157
49	Charged	35028
	Tallied	1626
	Amassed	381
	Itemized	136
50	augment	838
	Assuage	437
	tantalize	91
	Temporize	12

Note: The shaded sections indicate correct answers.

Figure 1 arranged frequency values of the total sixty choices from the least to the most. In order to clearly identify the frequency values under 2000, the highest frequency values of 30528 and 14999 were not expressed. It showed that most of them were below 2000. Compared to the tentative median value of 6000 in the preliminary study, about 87% was below 6000. This tentative value is the frequency value of frequently used word in general based on these sample sets and COCA as stated above. Below 87% implies that most of them are infrequently used among native speakers in real-life. In addition, 40 choices were

below 1000 which were way too low from the tentative value. 66% were below 1000. Through the figure 1, it was found that most of the choices from question number 46 to 50 had low frequency values based on the result of the preliminary research.
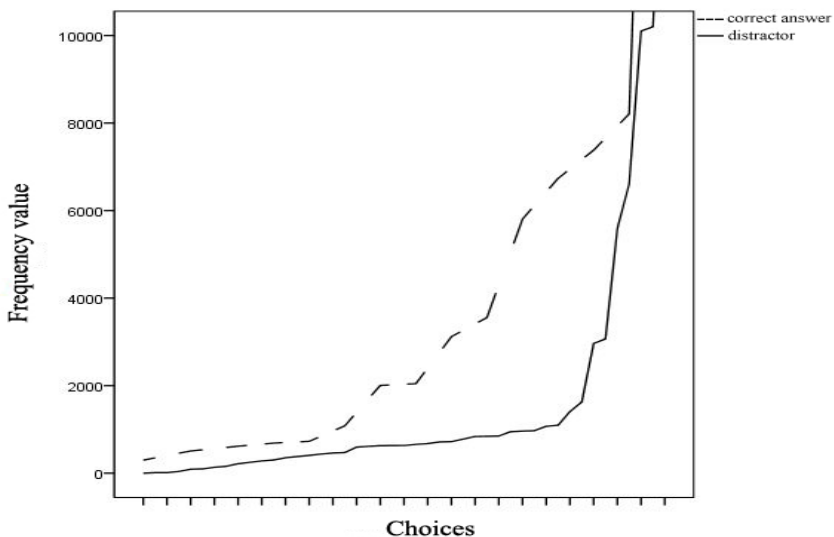
Figure 1
Arrangement of Frequency Values of 60 Choices

So far, the previous study examined the frequency of 60 options. After examining the frequency values, it was needed to look for the relation between correct answers and distractors. It aimed to see whether correct answers have higher frequency values even though the preceding study revealed overall low frequency values. If correct answers tend to have higher frequency values, it can be more adequate for the purpose of assessing English proficiency by asking more frequent words (than distractors) even within overall low frequently-used words.
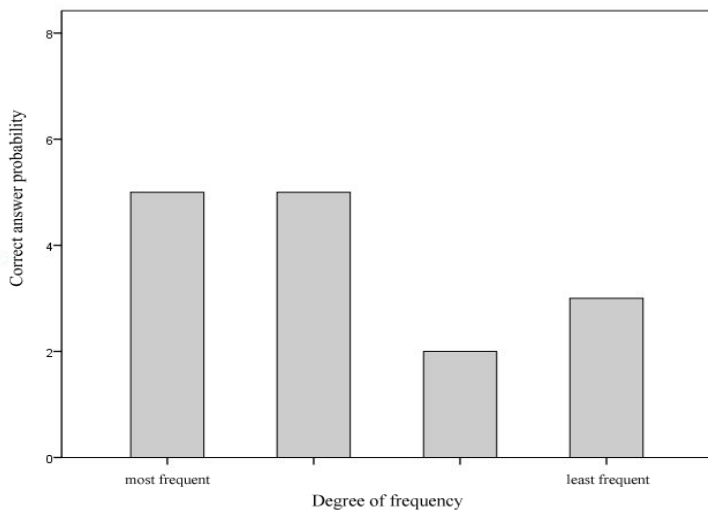
Like Figure 1, the frequency values of 30528 and 14999 were too high to include in Figure 2 to make frequency values under 2000 more clearly identified. Figure 2 compares the frequency values of correct answers and those of distractors. It is not cumulative but arranged from the least frequency value to the most frequency value. The dotted line represents the frequency values of correct answers, and the solid line is those of distractors. It should be noted that the number of correct answers is 15, and that of distractors is 45. According to figure 2, it was found that the frequency values of the correct answers tend to be higher than those of the distractors. Even though the first main study result showed the tendency of low frequency values in all options, figure 2 raised possibility that correct answers tend to have higher frequency values, and lower frequency values are of distractors. The following study specified this possibility.

Figure 2
Frequency Value Comparison between Correct Answers and Distractors

In this section, it is to look into the probability of correct answers in each question according to the frequency value (see Figure 3). There are four bars in Figure 3, and they are named, *'the most frequent'*, *'more frequent'*, *'less frequent'*, and *'the least frequent'* respectively. In one item, the frequency value of correct answers can be the most frequent, the second most frequent, the least frequent, or the second least frequent. All the correct answers were distributed in terms of their relative ranking in one item they belong to. If correct answers belong to the frequent categories of two, it is probable that TEPS vocabulary questions of 46 to 50 are valid ones because the ultimate goal of the test is for examinees to know the correct answer vocabulary. Moreover, being frequent words means that they are more likely to be used in real-life by native speakers, so validity can be raised. The total number of questions was 15, so there were 15 answers among four choices of each question.

Figure 3

Distribution of Correct Answers in terms of Frequency Value Degree

The percentage of the correct answers which belongs to the most frequent ones among four choices was 33%, and there was the same percentage for the second most frequent ones. Among four choices, the correct answer probability of the most and the second most frequent ones was 66%. According to the figure, TEPS vocabulary questions of 46 through 50 can be said to be valid because the questions tend to require examinees to know more frequent words among four options.

5. Conclusion

In a study of frequency values, 87% of choices were below the tentative value from the preliminary research. In the following studies on the frequency of correct answers and distractors, it was revealed that correct answers tend to have more frequency values than distractors. In the final study, it was more clearly shown that correct answers tend to belong to the most frequent or the second most frequent choices.

Options of TEPS vocabulary question number 46 to 50 tend to have less frequency values, but it was also found that less frequency value is of distractors, and correct answers actually tend to have more frequency values. Based on the results, this paper concludes that TEPS vocabulary questions from 46 to 50 can be said to be valid because its goal of requiring examinees to know the words of daily life which are frequently used is achieved through them.

6. Limitations

There are some limitations for generalization of the result of the present study. First, it is from the use of median value only. To get an approximate frequency value to figure out item difficulty, there were three options of mean value, median value, and mode. It is more

common to use mean value to narrow a margin of error, but there was no choice but to use median value because of a wide gap among frequency values of four options.

Secondly, it is owing to the use of corpus data only instead of using native speakers' intuition. The best way is to get information about whether each word is frequently used by native speakers. However, it was hard to design the experiment, collect data from a considerable number of native speakers, and get a meaningful result within a short time, so only corpus was possible to use. This factor made the test result limited to generalize.

The third reason is in consideration of various parts of speech. In one word, it has various forms. A verb has its past tense, past participle, and present participle. For a noun, it has its singular form and a plural form. Each form has different frequency values. For example, a verb *introduce* has 8686 frequency value, but its noun *introduction* has 12009 frequency value. As can be seen, there is more or less frequency value gap among various parts of speech. In this paper, this variation was not considered, instead, the exact word forms from the options were used. It may hinder generalization of this paper.

7. Further Study

For more reliable study, it needs mean values additionally because using only the median value is not considered to be reliable in general. The present study treated the corpus data only, but in determining whether a word is appropriate for testing English proficiency or not, it is better to depend on native speakers' intuition in addition. It would be beneficial to take into account of native speakers' intuition in deciding validity of 46 to 50 questions in TEPS vocabulary part. Finally, it was hard for a student to get authentic TEPS materials. In this paper, sample sets of TEPS by TEPS Committee were used, but the book to which

these sample sets belong is considered to be relatively easy. For a more reliable result, it needs to use not sample sets but authentic test materials to search for the validity of a vocabulary part.

References

Choi, I. (2008). The impact of EFL testing in EFL education in Korea. *Language Testing, 25*(1), 39-62.

Cho, D. (2000). A study on the validity of TEPS. *Journal of the Applied Linguistics Association of Korea, 16*(1), 197-213.

Lee, H., & Lee, S. (2003). A study on the relationship between the scores of TOEFL, TOEIC, and TEPS and college academic performance. *English Language & Literature Teaching, 9*(1), 153-171.

Lee, H., & Lee, S. (2004). Validity of TOEFL, TOEIC, and TEPS as a measure of communicative competence. *English Language & Literature Teaching, 10*(3), 191-210.

Park, N., Cho, C., Kim, C., Kim, I., & Choi, I. (1998), Certification of practical English language proficiency. *English Teaching, 53*(2), 179-205.

Wikipedia. *Skewness*. Retrieved December 17, 2012, from http://en.wikipedia.org/wiki/Skewness

서울대 텝스 관리위원회. (2010). **서울대 텝스 관리위원회 제공 최신기출 시크릿**. 서울: (주)넥서스.

Joohee Shin
jhshinii@snu.ac.kr