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

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

Stuart McLean[a], Nicholas Hogg[b] and Brandon Kramer[c]
[a]*Temple University, Japan;* [b]*Osaka Yuhigaoka Gakuen High School;*
[c]*Momoyama Gakuin University*
doi: http://dx.doi.org/10.7820/vli.v03.2.mclean.et.al

Abstract

Measuring students' lexica is time-consuming, as one sitting of the Vocabulary Size Test (VST) usually takes 40–60 minutes. As a result, teachers would benefit from being able to make reasonable estimates from commonly available information. This paper aims to investigate: (1) What are the mean vocabulary sizes of students at Japanese universities as a whole, and by university department (*hensachi*)? and (2) Are a university's department standardized rank scores (*hensachi*) a useful proxy for English vocabulary size? This study used a cross-sectional design where 3,449 Japanese university students were tested using Nation and Beglar's VST. The results showed an average score of 3,715.20 word families and that VST scores were significantly higher for students in higher department *hensachi* programs. This current department *hensachi* was also found to have a stronger correlation with VST scores than with other covariates when the entire sample was considered. Lastly, there appears to be a lack of consistent knowledge of the most frequent words of English, suggesting that curriculum designers at Japanese universities should focus on teaching high-frequency English words. Although the findings support the use of the VST for comparing receptive written vocabulary knowledge between learners, they perhaps do not support its use in establishing a vocabulary size to decide lexically appropriate materials.

1 Background and Aim

Measuring students' lexica is time-consuming, as one sitting of the Vocabulary Size Test (VST) (Nation & Beglar, 2007) usually takes 40–60 minutes. As a result, teachers would benefit from being able to make reasonable estimates from commonly available information.

Previous research on the vocabulary size of Japanese students is limited. Shillaw (1995) and Barrow, Nakanishi, and Nishino (1999) suggested that the vocabulary size of non-English-major Japanese university students is between 2,000 and 2,300 word families. In these studies, vocabulary knowledge was assessed over 3,000 word families, with students completing self-checking familiarity surveys. However, this approach may have simply measured the word forms students recalled being exposed to rather than measuring receptive reading vocabulary knowledge. This is important as Waring and Takaki (2003) demonstrated that the knowledge thresholds for recognizing a previously encountered word and correctly

Please cite this article as: McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' english vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction*, *3* (2), 47–55. doi: 10.7820/vli.v03.2.mclean.et.al

answering a receptive vocabulary knowledge item differ, and instructors are probably more interested in the latter type.

Based on the need for more recent and thorough research, the following research questions were developed:

(1) What are the mean vocabulary sizes of students at Japanese universities as a whole, and by university department rank (*hensachi*)?
(2) Are a university's department standardized rank scores (*hensachi*) a useful proxy for English vocabulary size?

2 Methodology

2.1 Instruments

To estimate students' receptive vocabulary knowledge, Nation and Beglar's (2007) VST was used. The test is based on the spoken portion of the British National Corpus (BNC), ordered by word frequency and grouped into 1,000 word bands up to the 20,000 word level. For this research, however, only the first 8,000 word families were tested, measured with 80 items, each representing knowledge of 100 words. Tests were completed in one of the two formats: traditional paper tests and online tests administered through Survey Monkey <surveymonkey.com>. Both tests were the same version of the VST and were presented with the same instructions according to the test authors (Nation & Beglar, 2007) in order to minimize variance between the two mediums. The online test was advantageous administratively in that it allowed for easy marking and statistical analysis. It was also noticed by the researchers that students who conducted the paper tests were more inclined to not answer some questions, particularly as the difficulty increased, representing a weakness in the study.

In addition to the VST, a questionnaire was given to participants to collect personal information such as year, major, study abroad experience, native language, available Test of English for International Communication (TOEIC) and Test of English as a Foreign Language (TOEFL) scores, and previous *hensachi* data. A *hensachi* is a score assigned to individual students or school departments based on student performance in a national test standardized to the national mean. A *hensachi* of 50 represents the mean, where one standard deviation above or below is represented by 60 or 40, respectively. The scores can range from 20 to 80, but 95.4% of all university departments fall between 30 and 70 (Newfields, 2006). The *hensachi* data collected in the questionnaire refer to standardized tests the students took toward the end of high school, while the department *hensachi* scores used to address the research questions were found online through the website of Benesse, a large testing company in Japan <http://shinken.zemi.ne.jp/hensachi>. In this study, a university's department *hensachi* is the mean *hensachi* of the first-year students in a given department who took the university entrance exam and Benesse's proficiency test. It does not include those who entered the university through other means, such as recommendations. Unfortunately it was not feasible to similarly verify the sources of the students' previous *hensachi* scores obtained through the questionnaire, as each specific score varies based on the particular

exam board where it was calculated. Lastly, the TOEFL test exists in a number of formats: the internet-based test (iBT), computer-based test (CBT), and paper-based test (PBT). As a result, it was necessary to extrapolate the equivalent PBT scores from self-reported iBT and CBT scores using a conversion table provided by the makers of the TOEFL test (the Educational Testing Service, 2005).

2.2 Participants and Research Design

This study used a cross-sectional design where data from 3,449 Japanese university students were collected through a snowball sampling approach, thanks to the assistance of many university instructors who kindly agreed to administer the test in their classrooms after requests in regional newsletters, at local events, and through personal correspondences. The tests were generally completed at the beginning of term during April and May of 2012 and 2013; however, a small number of tests were completed at the start of the autumn term in September and October of 2013.

The VST scores of 22 participants were identified as outlying the sample based on the calculation of their Rasch person ability estimates using the WinSteps statistical program. Participants with scores at or beyond 3.29 standard deviations from the mean based on their ability estimates were removed from the study, in line with the recommendation of Field (2009). Closer inspection revealed that these outliers were mostly the result of participants not answering a large number of the items when taking the test. Table 1 shows a summary of the participants divided by year, major, and department *hensachi*.

Table 2 shows that the mean written receptive vocabulary size (VST score $\times$ 100) was found to be 3,715.20 word families for the remaining participants (n = 3,427). A visual inspection of the histogram of the VST scores, shown in Figure 1, indicates that the data are normally distributed. For the purpose of comparison, participants were separated into three *hensachi* groups: $\geq$61, 51–60, and $\leq$50. The correlation coefficients were computed between the participants' VST scores and available TOEIC, TOEFL, individual *hensachi*, and department *hensachi* scores.

Table 1. Summary of Participants

	Hensachi (Department)	1st year	2nd year	3rd year	4th year
English Majors	≥61	n = 1			
	51–60	n = 76	n = 46	n = 3	
	≤50	n = 101			
Science Majors	≥61	n = 320	n = 3	n = 1	n = 2
	51–60	n = 224	n = 228	n = 98	
	≤50	*n = 184*	*n = 193*	*n = 14*	
Arts Majors	≥61	n = 253	n = 23		
	51–60	n = 519	n = 337	n = 22	n = 5
	≤50	*n = 478*	*n = 210*	*n = 50*	*n = 36*

Table 2. Descriptive Statistics VST Scores

N	Min	Max	*M*	*SEM*	*SD*	Skew	*SES*	Kurt	*SEK*
3,427	500	7,400	3,715.20	21.66	1,268.15	−0.20	0.04	−0.16	0.08

SD = Standard deviation.

3 Results

Before considering the research questions, the internal consistency reliability of the VST data collected in this study was examined using the Kuder–Richardson Formula 20 test. The coefficient of reliability (KR-20 $\alpha = .91$) was high, indicating that it is highly probable that all the items measured the same construct.

As shown in Table 3, the three *hensachi* groups achieved different mean VST scores. A one-way analysis of variance (ANOVA) was conducted to evaluate the hypothesis that the participants' VST scores would significantly vary between department *hensachi* groups. The independent variable was department *hensachi* group with three levels: ≥ 61, 51–60, and ≤ 50. The dependent variable was VST scores.

The ANOVA showed a significant effect of department *Hensachi* group on VST scores, $F(2, 3424) = 1{,}383.14$, $p < .001$, $\eta^2 = 45$. The strength of the relationship between *hensachi* groups and VST scores was very strong, accounting for 45% of the variation of the dependent variable. Because the overall F test was significant, follow-up tests were conducted to evaluate pairwise differences among the means. As the variances of the three groups were heterogeneous, the Games

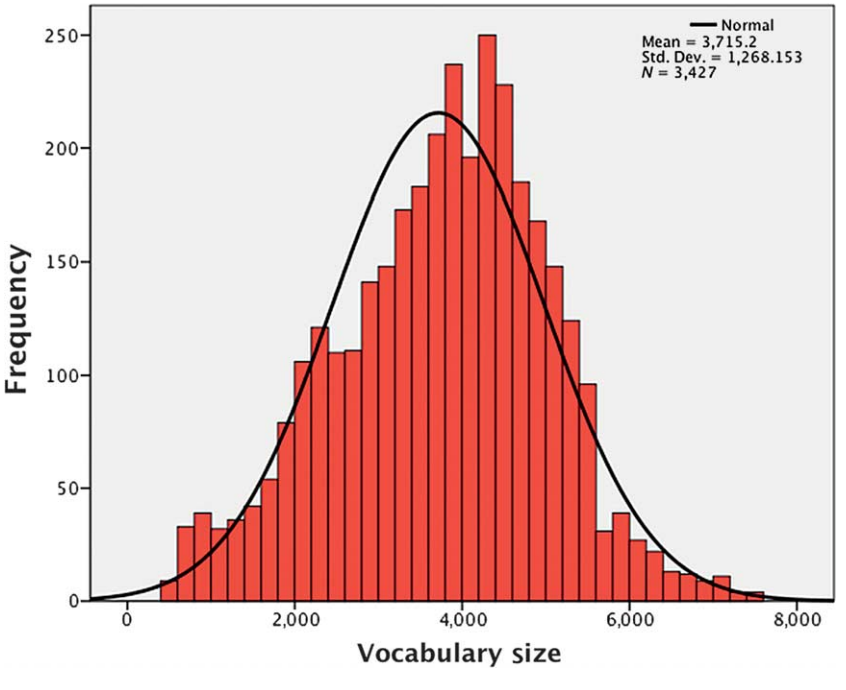

Figure 1. Histogram of grouped distribution of VST scores showing the normal distribution curve.

Table 3. Descriptive Statistics VST Score by Hensachi Group

Hensachi	*N*	Min	Max	*M*	*SEM*	*SD*	Skew	*SES*	Kurt	*SEK*
≥61	603	1,900	7,400	4,903.15	29.44	723.00	−0.19	0.10	0.77	0.20
51–60	1,558	900	7,400	4,102.89	23.18	915.15	0.31	0.06	1.39	0.12
≤50	1,266	500	7,100	2,672.27	29.87	1,062.82	0.29	0.07	0.33	0.14

SEM = Standard Error of the Mean; SES = Standard Error of Skewness; SEK = Standard Error of Kurtosis.

Howell test, which does not assume equal variance among groups, was used. To control for Type I errors, the Bonferroni approach was used, thus a p value of less than .017 (.05/3 = .017) was required for significance. The ≥61 *hensachi* group had significantly higher VST scores than the other two groups (p = <.001 for both comparisons), and that the 51–60 *hensachi* group had significantly higher VST scores than the ≤50 *hensachi* group (p = <.001).

Spearman's rho correlation coefficients were calculated between six different student scores and VST scores. To control for Type I errors, the Bonferroni approach was used, thus a p value of less than .008 (.05/6 = .008) was required for significance. As displayed in Table 4, while all six scores correlated significantly with VST scores, the highest correlation was found between department *hensachi* and VST scores at .73 ($p < .001$), indicating that students studying in departments with higher *hensachi* scores tend to score higher on the VST. Examining the correlational strength of department *hensachi* scores for only students whose TOEFL scores were available ($n = 259$), TOEFL had a higher correlation with VST scores ($r_s = .58$, $p < .001$) than department *hensachi* ($r_s = .34$, $p < .001$). Likewise, the correlation between TOEIC and VST scores ($n = 412$) was higher ($r_s = .57$, $p < .001$) than department *hensachi* ($r_s = .32$, $p < .001$) for the same students.

The mean item accuracy on the VST was calculated for the three *hensachi* groups with items grouped into 1,000-word frequency levels, as shown in Table 5. The accuracy ranged from 85% (*hensachi* ≥61 group, 1k word level) to 22% (*hensachi* ≤50 group, 6k word level). Overall, accuracy decreased as the words became less frequent as shown in Figure 2. The decline was determined to be non-linear by adding best-fit lines with the highest indices of fit. For the *hensachi* ≥61, and 51–60 groups, the lines were polynomial with coefficients of determination of $R^2 = .89$ and $R^2 = .82$, respectively. For the *hensachi* ≤50 group, the line was

Table 4. Spearman's Correlation Coefficients between Student Scores and VST Results

Factor	*n*	*p*	Correlational coefficient with VST scores	Dept *hensachi* and VST coefficient for each subpopulation
Department *hensachi*	3,427	<.001	.73*	.73*
HS English *hensachi*	826	<.001	.60*	.60*
TOEFL	259	<.001	.58*	.34*
TOEIC	412	<.001	.57*	.32*
HS 3rd subject *hensachi*	643	<.001	.50*	.64*
HS 5th subject *hensachi*	628	<.001	.49*	.61*

*$p < .008$. HS = High School.

Table 5. Percentage Item Accuracy at 1,000 Word Levels

	Word Level							
Department *hensachi*	1k	2k	3k	4k	5k	6k	7k	8k
≥61	85	71	73	68	59	42	44	48
51–60	79	54	62	56	43	35	40	42
≤50	59	34	38	36	26	22	26	25

logarithmic with a coefficient of determination of $R^2 = .84$. The trend was toward accuracy declining steeply as item difficulty increased, but then leveling off to a slower decline toward the later items.

4 Discussion

Answering research question 1, the vocabulary sizes of Japanese university students was found to be 3,715.20 word families. This estimate is greater than the 2,000–2,300 as found by Shillaw (1995) and Barrow et al. (1999). However, in these studies students were only tested on their knowledge of the most frequent 3,000 word families in contrast to the present study which tested the participants' knowledge of the most frequent 8,000 words. The mean vocabulary sizes of the ≥61, 51–60 and ≤50 *hensachi* groups are 4,903, 4,103, and 2,792 words, respectively.

Research question 2 asked: Are a university department's standardized rank scores (*hensachi*) a useful proxy for English vocabulary size? The results of the ANOVA and post hoc tests showed that VST scores were significantly higher for students in higher department *hensachi* programs. Figure 2 also shows that the

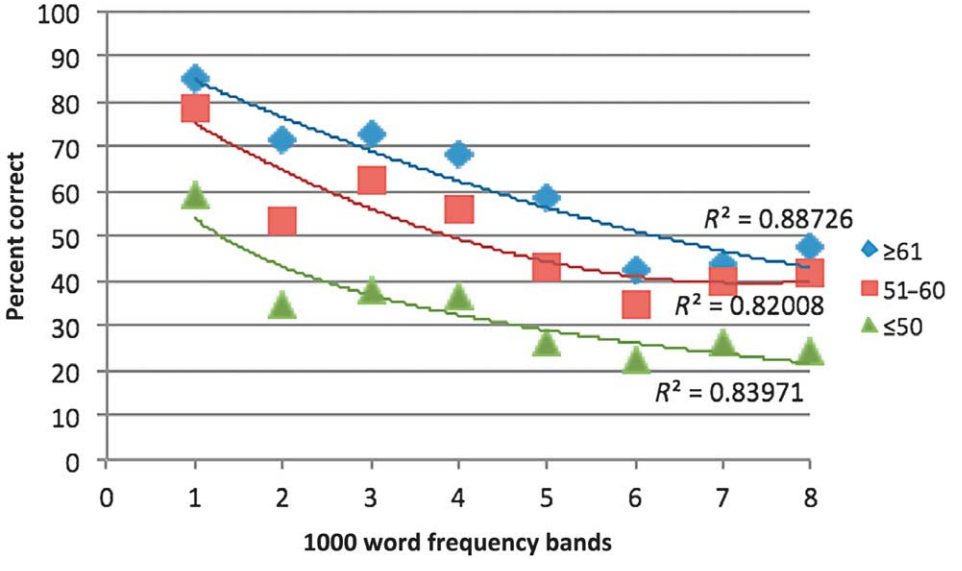

Figure 2. A scatter plot of VST item accuracy at each 1,000-word frequency level for the three *hensachi* groups with best-fit lines.

vocabulary size differences of the three *hensachi* groups are relatively consistent throughout each band of the VST. These findings are promising in that department *hensachi* scores are readily available online allowing teachers to easily gain a general understanding of their students' vocabulary size relative to other departments. However, it should be cautioned that while the mean VST scores achieved represent the approximate total written receptive vocabulary sizes of the students in the three *hensachi* ranges (4,903, 4,103, or 2,792, respectively), the results do not suggest that instructors can expect their students to have complete knowledge of the most frequent 4,903, 4,103, or 2,792 English words. The answers to the research questions, when considered along with the high internal consistency reliability, support the use of the VST for comparing written receptive vocabulary knowledge among English learners, while suggesting that it may not be appropriate for establishing specific vocabulary sizes for uses such as selecting or creating lexically appropriate materials.

In examining the correlational strength of different student scores and vocabulary size, the highest found within all participants was with their current department *hensachi*. This is explained by three factors: possibly differing sources of individual previous *hensachi* scores, the self-reported nature of all the individual scores used and inaccuracies therewith, and not the least of which the larger sample size for the comparisons between department *hensachi* and VST scores. Although the data were limited in this study, when available for an entire group of students, TOEFL scores, followed by TOEIC, appear to be better proxies for vocabulary size. When such scores are unavailable, department *hensachi* is a reasonable substitute.

For all three groups, there appears to be a lack of consistent knowledge of even the most frequent words of English. These lexical gaps were found to be even greater with lower-department *hensachi* students. This suggests that regardless of *hensachi* level or the predicted vocabulary size, vocabulary components of the curricula at Japanese universities should focus on high-frequency English words. This finding also provides negative evidence for the common assumption that texts written at the 1,000 most frequent word level are lexically appropriate for university students in Japan when conducting meaning-focused learning or fluency activities. Even if all of the words within a text were from this level, the average university student would have far less than the 98% lexical coverage recommended by Hu and Nation (2000) for unassisted comprehension. Thus some provision for dealing with unknown words should be made.

5 Conclusion

This paper reports the results of the VST taken by 3,427 students at various Japanese universities. The scores were found to be significantly higher for students in higher-department *hensachi* programs, and a strong correlation was found between the two variables, suggesting that instructors can quickly and easily gain a general understanding of students' vocabulary size simply from their department *hensachi* score. In all three groups, however, the results show a lack of consistent knowledge among the most frequent 2,000 word families found in the BNC. This emphasizes the need to focus on supporting high-frequency vocabulary acquisition regardless of university rank. Alternatively, it questions the appropriateness of

measuring Japanese university students' lexical knowledge based on word frequency within the BNC. This frequency list, for example, ranks the word "nil," at the 2,000 word level, as more frequent than "quiz," at the 5,000 word level, while containing more loanwords than may be proportionate to the BNC as a whole. The results support the use of the VST for separating learners by vocabulary knowledge, but suggest that it may not be appropriate for establishing students' vocabulary size for deciding and creating appropriate materials.

5.1 Limitations and Future Research

Preliminary in-depth interviews conducted with participants after completing the test suggest that, as is to be expected on a multiple choice test, guessing can inflate VST scores, and especially so for test takers with limited lexical knowledge. The issue of student guessing within these data is being addressed for future research, and while the analyses described in this paper indicate that this does not hurt the test's ability to separate the students according to ability, it may cast doubt on the test's ability to accurately estimate the latent vocabulary size. This is in line with the results discussed by Stewart (2014), who also expresses concern that student guessing causes an overestimate of vocabulary size using the VST.

Another limitation is the difficulty of determining the degree to which participants took the test seriously. Although it has yet to be explored in depth, test fatigue appears to have affected the $\leq$50 *hensachi* group to a greater extent based on the number of unanswered items. It is possible that more participants could and should have been removed had it come to light that they had given up or guessed wildly.

In addition to the results presented in this paper, there are a number of directions that would be interesting to pursue and would shed further light onto both Japanese students' vocabulary sizes as well as the VST in general. First, it would be ideal if the participants included more English majors and 3rd and 4th year students. Second, the self-reported *hensachi* data collected for participants' individual English, three-subject, and five-subject *hensachi* scores were of limited value as it was not ascertained which examination boards' tests were taken. The correlations between these scores and the VST may have been higher had this information been accounted for. The issue of loanwords in the test possibly inflating Japanese students' scores should also be explored. While the use of loanwords in the test is not itself a problem, it can become problematic if they make up a greater proportion of the test questions than the corpus it represents.

References

Barrow, J., Nakanishi, Y., & Nishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System*, *27* (2), 223–247. doi:10.1016/S0346-251X(99)00018-4

Educational Testing Service. (2005). *TOEFL iBT scores: Better information about the ability to communicate in an academic setting*. Retrieved from hhl.de/fileadmin/texte/_relaunch/Conversion_Table_TOEFL_(PBT,CBT,iBT).pdf

Field, A. (2009). *Discovering statistics using SPSS*. London, UK: Sage.

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13* (1), 403–430. Retrieved from nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehchao.pdf

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31* (7), 9–13. Retrieved from jalt-publications.org/files/pdf/the_language_teacher/07_2007tlt.pdf

Newfields, T. (2006). Assessment literacy self-study quiz #1 [Suggested answers]. *Shiken*, *10* (2), 25–32. Retrieved from jalt.org/test/SSA1.htm

Shillaw, J. (1995). Using a word list as a focus for vocabulary learning. *The Language Teacher*, *19* (2), 58–59. Retrieved from jalt-publications.org/tlt/issues/1995-02_19.2

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11* (3), 271–282. doi:10.1080/15434303.2014.922977

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15* (2), 130–163. Retrieved from http://nflrc.hawaii.edu/rfl/October2003/waring/waring.html