〈原著論文〉

A Rasch-Based Validation of the Vocabulary Size Test with High School Students

Robert J. S. ROWLAND

Abstract

The primary purpose of this study was to provide further validity evidence for an abridged, 40-question version of the Vocabulary Size Test (VST). The secondary purpose was to build on a validity argument for the VST made by Beglar (2010). The VST was administered to a class of 15 and 16-year old high school students ($N=43$). Data analysis was performed based on the Rasch model and results examined with reference to Messick's framework of validity. The study results indicate that (1) the items and test-takers largely performed as expected based on hypothesis, (2) most of the test items fit the Rasch model well (3) performance of items that did not fit the model well could be explained by knowledge of L1 cognates on the test and low-frequency items with high-frequency occurrence in classroom learning.

Key words: Vocabulary Size Test, Rasch model, Validity, Unidimensionality, Monolingual English version

Introduction

Building vocabulary is an important part of any language learner's goals. The job of an educator is to guide learners along a path to greater vocabulary knowledge in a principled fashion. There are two important questions that an educator must ask when designing a vocabulary curriculum to increase learner vocabulary.

The first question is with regard to which words are most useful for a learner to learn. Studies in corpus linguistics suggest that there are words in a given language that appear in texts with higher frequency than other words. West[(1)] compiled the *General Service List of English Words*, which contains 2,000 high-frequency word families. Schmitt and Schmitt[(2)], however, argued that a larger list of 3,000 word families would be a better goal for learners.

人文学部・欧米文化学科
論文受理日 2017 年 11 月 6 日

Nation's[3] findings support these claims, as he suggested that knowledge of the most frequent 3,000–4,000 words plus proper nouns are necessary for 95% coverage of a variety of media of English including spoken English, Children's movies, Newspapers and Novels. It follows, then, that learners would make the most effective use of their study time by focusing on learning these high-frequency words.

The second question is with regards to how learners can track their learning of these words over time. One way to assess a learner's vocabulary learning is by examining their receptive vocabulary. Two popular tests for measuring receptive vocabulary are *The Vocabulary Levels Test* (VLT)[4] and *The Vocabulary Size Test* (VST)[5]. The VLT was designed as a diagnostic test for vocabulary knowledge. According to Nation[6] this test does not measure a learner's total vocabulary size, but rather indicates which level of vocabulary frequency a learner should focus their studies on. A learner with large gaps in high-frequency vocabulary would be better off focusing their efforts on acquiring these words before moving to less frequent words. The VLT is a good starting point for a short-term class, but offers no mechanism for reliably monitoring vocabulary knowledge growth over time. The VST, on the other hand, was designed to measure a learner's total vocabulary size. Beglar's[7] Rasch-based validation of the VST with a large, diverse group of learners suggested that the VST may be a useful tool for tracking changes in a learner's vocabulary size over time. Thus, the VST is the more appropriate test of the two to administer to a group of learners who can be tracked over an extended period.

Beglar[8] presented Rasch-based validity evidence for the VST in light of Messick's[9] construct-centered approach to validity. Messick's framework places the constructs to be measured by a test at the center of the development of methods for measurement and scoring. There are 6 facets of construct validity: content, substantive, structural, generalizability, external and consequential. The content facet is composed of content relevance, representativeness and technical quality. Content relevance and representativeness stipulate that test items test specifically the constructs they are intended to measure. Technical quality refers to the quality of test construction and whether or not this quality interferes with construct measurement. The substantive facet refers to how information is found on whether or not the examinees engaged in the targeted cognitive processes the items were designed to measure. The structural facet examines whether or not the test accurately targets the single construct it was designed to target, that is, whether or not the test exhibits psychometric unidimensionality. The generalization facet refers to the level to which measurements and interpretations can be generalized across different populations of test takers or different sets of tasks with identical

design parameters. In other words, the generalization facet is concerned with the invariance of the instrument. External validity examines the extent to which behaviors which were not expected to interact with the construct affect measurements. Consequential validity examines the validity of actions taken in light of interpretations based on measurements and whether or not they are appropriate.

Beglar[10] made a strong argument for the validity of the VST based on an interpretation of a Rasch-analysis examined in relation to Messick's construct validity. His study had 197 participants separated into four groups by English proficiency. The four groups were 1) native English speaking doctoral students 2) high proficiency English speaking Japanese doctoral students 3) intermediate proficiency English speaking Japanese students of an immersive English language program and 4) low proficiency English speaking Japanese students at a Japanese university. While Beglar concluded that the VST appears to be a valid instrument for measuring the vocabulary sizes of a wide range of proficiencies of native and non-native English speakers in tertiary education, he provides no evidence or suggestion for whether or not the same conclusion may be drawn for students in secondary education. The current study draws on the methods of analysis of Beglar to investigate how one form of the VST functions with a group of secondary school EFL students in Japan.

Method

Participants

One group of 10th-grade students at a private high school in Japan ($N=43$) participated in this study. There were 32–female students and 10–males. These students were English majors who, in addition to electing to replace higher maths classes with English composition and reading classes for a total of 9 hours of EFL classes per week, had also been placed in the school's highest academic track based on junior high school final marks and performance on the school's entrance exams. All students had completed at least three years of English education as per the compulsory curriculum in junior high school. 6 students in the class had spent a year or more in an English first language schooling system in another country. No other students had ever lived outside of Japan. All students were informed that the purpose of the test was to gather preliminary information on their vocabulary size and to follow up with a second test at the end of the academic year to measure growth.

The instrument

The primary assessment instrument was the *Vocabulary Size Test* (VST)[11]. The VST is a 140–item multiple-choice test intended to test the vocabulary size of native and non-native English speakers. The test items consist of 14–sets of ten questions, each set drawing from one of Nation's[12] fourteen, 1000–word British National Corpus (BNC) frequency word lists. The lexical unit for these lists is the word family. Empirical evidence suggests that the word family is a psychologically real unit[13]. Nation and Webb[14] say word families are appropriate for receptive vocabulary tests because they are more inclusive than other word units, such as lemmas. Their argument is that learners beyond the absolute beginner level will have some word building ability, such as knowledge of word parts, derivational and inflectional forms, and therefore may be able to accurately guess the meaning of an unknown word. Word families included in these lists have been set to level 6 of the Bauer and Nation's[15] scale of levels. Level 6 is highly inclusive and list members met the criteria of frequency, regularity, predictability and productivity.

The lists were originally sequenced according to frequency in the 100–million word BNC. Nation and Beglar[16] suggested that, because these lists were largely based on a written corpus, informal words which typically appear at earlier stages of development (i.e. *hello, cat*) appear in later frequency lists, while more formal words (i.e. *civil, commission*) appear on earlier lists. To calibrate the test to represent more accurately natural receptive vocabulary development, the first twelve 1000–word lists were re-sequenced referencing the 10–million word spoken English sub-corpora of the BNC. Though reordering did not result in drastically different lists, these lists appeared to be more reasonable representations of words likely to be incidentally encountered in a variety of contexts.

The 140–items of the VST are multiple-multiple choice questions. Each question prompt places the target word in a sentence with limited context. Placing items in limited context hints at the target word's part of speech as well as provides contextual orientation of word usage. Each question has 4 choices. This format was chosen because it is familiar to a wide variety of test takers, to maintain control for equal cognitive demand for each question through a careful drafting procedure, to make marking consistent and simple, and to make learners demonstrate knowledge of each individual item. Here is an example of item 4 from the second 1000–word level:

4. DRAWER: The **drawer** was empty
   a. sliding box
   b. place where cars are kept
   c. cupboard to keep things cold
   d. animal house

Choices for each question were drafted with limited vocabulary. Each choice was written to define the target using vocabulary from higher frequency word lists than the target, and relying on the first 2,000 words as much as possible. Members of each set of choices are all interchangeable with the target in the context sentence, which demands test takers to exhibit well-developed knowledge of the target word to answer correctly.

Procedures

Form A of the VST was administered to the learners in one, 45–minute session. Only the first four sections of the test, a total of 40 items, were used. Nation and Beglar[17] and Beglar[18] argued that there was little value in having low-level learners sit all fourteen sections of the test, as it was unlikely that valuable data would be obtained from the lower frequency levels and suggested that the first four word lists were likely the most appropriate for these learners. Items were then scored dichotomously, results recorded in a text data file, and exported to WINSTEPS 3.81.0[19]. Items were coded in groups of ten. The ten items from the first 1,000–word list were assigned the label "f#" where "#" was the item number with each level between one and ten. The second, third and fourth sets of 10–items were coded in the same way as "S#," "T#" and "O#." Students were each assigned a number from 1 to 42. Possible scoring codes were "a," "b," "c," "d," and in the event of an unanswered question, "X." Unanswered questions were scored incorrect as it was assumed that lack of knowledge of a word lead to a blank answer. Table 1 includes the answer key and a small segment of the data set.

Table 1 *VST Form A answer key and example student data*

Answer key	c c b d a d b d d d a a d a a b a c b d d d a c d d d a b c b c d b c a a a a a
Student 1	c c d b a d b d d b c a a a a c d c d b d c d b b b d a a b a d a d c a a b a d
Student 2	c c b d a d b d d d a a d a a a a c b d d d a c d d d a b c a b b b c a a a a b
Student 3	c c b x d a d x b d x d a x x x a x d c x x x x x b x a b c x x x x c a a a x x

The data was then analyzed using the Rasch dichotomous model. The mathematical formula for this model is

$$P_n = \exp(B_n - D_i) / [1 + \exp(B_n - D_i)]$$

Beglar[20] explains that P_{ni} is defined as "the probability of person n with ability B_n succeeding on item i, which has a difficulty of D_i; exp = the exponent of the natural constant e = 2.71828". This model was chosen because it enables the construction of linear item and person measures, which in turn enables the relation of the item and person hierarchy to hypotheses about the function of the item on the latent trait. Dimensionality of the data set can then be determined by examining the discrepancies between the expected and observed responses.

Results and discussion

The results of analysis will be discussed in relation to three aspects of Messick's[21] framework for construct validity: content, substantive and structural.

Construct validity: Content facet

The content facet of construct validity has 3 components: content relevance, representativeness, and technical quality. Content relevance has already been discussed at length. This test was designed carefully to measure vocabulary knowledge of words divided into sections by frequency of occurrence on Nation's[22] BNC frequency word lists. These careful design principles allow us to assume that the content relevance of questions of this test to the construct they were intended to measure exists.

To examine the representativeness of the test, one must determine 1) whether there are a sufficient number of items to measure the construct 2) whether the empirical item hierarchy shows sufficient spread and 3) whether significant gaps exist in the item hierarchy. Answers to these questions can be found by examining the Wright map (fig. 1) and the student and item fit statistics (table 2).

Figure 1 is an item-person map that displays the linear relationship between the calculations of the 43 test takers and 40 items. It shows that there are items represented throughout the range of low to high difficulty. The largest clustering of items, within one standard distribution of the mean of difficulty, coincides with the largest clustering of students. The 40–items for this test were selected on the assumption that students had at least some knowledge of words from the first four of Nation's[23] BNC frequency lists, and the item distribution seems to show that there was an appropriate distribution of difficulty of items for these learners.

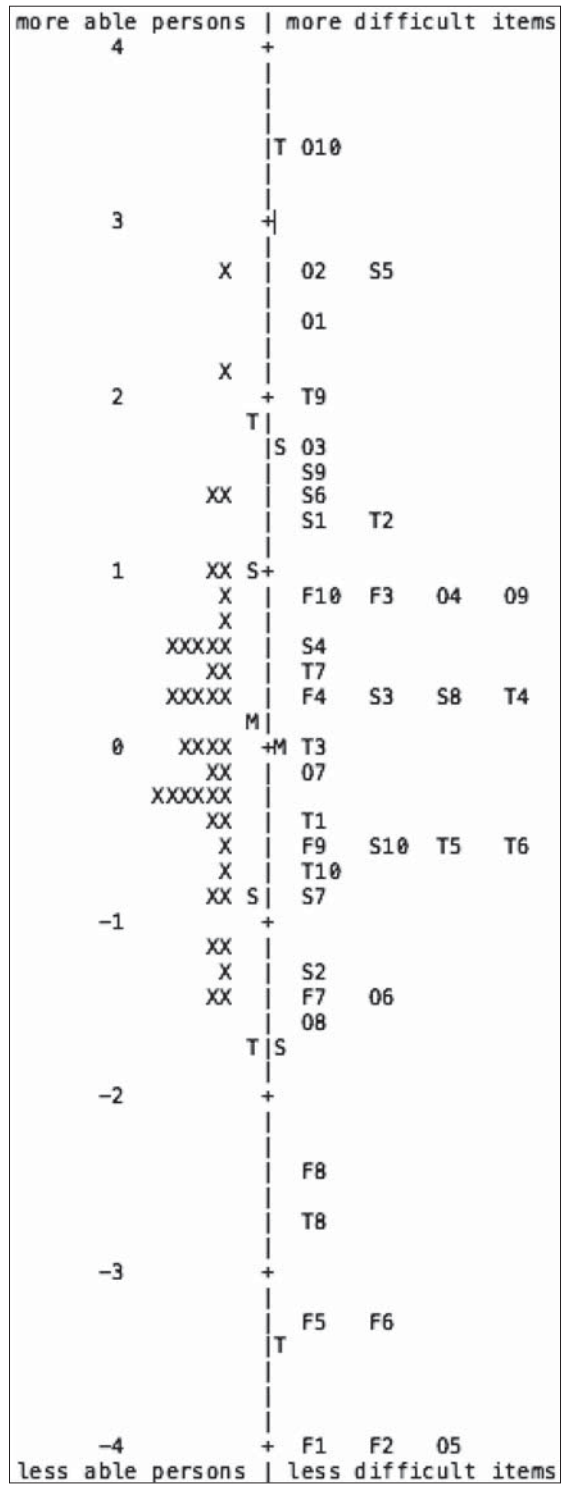

Figure 1. Wright map of person and item measures

Table 2 *Student and Item Fit Statistics*

Student	43 input		43 measured		Infit		Outfit	
	Total	Count	Measure	Realse	IMNSQ	ZSTD	OMNSQ	ZSTD
Mean	21.1	40.0	.1	.4	1.00	−.1	.97	.0
S.D.	5.1	.0	.9	.0	.26	1.3	.47	.9
Real RMSE	.4	True S.D.	.8	Separation	1.74	Student reliability		.75
Item	40 Input		40 measured		Infit		Outfit	
	Total	Count	Measure	Realse	IMNSQ	ZSTD	OMNSQ	ZSTD
Mean	22.7	43.0	−.3	.5	1.00	.0	.97	.0
S.D.	12.2	.0	2.0	.3	.14	.8	.35	.9
Real RMSE	.6	True SD	1.9	Separation	3.14	Item Reliability		.91

Examination of item fit (table 2) indicates 3.14 strata of item separation. These findings are consistent with Beglar's[24] who found 7.29 for the 140–item version of the same test. Previous research into similar tests further suggests that two strata of item separation are the minimum requirement to infer adequate representation of the measured construct[25]. The item hierarchy, therefore, shows sufficient spread to represent the construct.

Examination of the item hierarchy further indicates that there are few significant gaps. There is, in fact, considerable redundancy in the items found within the ability range of the learners. Furthermore, 10–items per level appears to be a sufficient number of items to test receptive vocabulary size accurately, as is indicated by the acceptable level of standard error of measurement of all learners (from .4 to .5, $M = .4$, $SD = 0$).

To examine the technical quality of the test, the individual item fit statistics were examined for misfit. Following Bond and Fox[26] Mnsq fit standards were set between 0.7 and 1.3 and Zstd set between −2.0 and 2.0 for both infit and outift. Only two items, O3, *candid* (Infit Mnsq = 1.32, Infit Zstd = 1.2) and T4, *scrub* (Infit Mnsq = .78, Infit Zstd = −2.2) displayed numbers outside of acceptable infit range. Further analysis of the distractors for O3, *candid* revealed that, while 19% of all test takers chose the correct response "say what you really think," 40% of all test takers answered option a, "be careful." The context sentence provided was "Please be" so it is likely that students guessed the option that seemed most familiar to them within their limited learning experience. As this distractor is not semantically similar to the correct response, it can be assumed that the distractor is well designed and was overly distracting only because of the students' limited knowledge. Closer examination of the distractors T3, *scrub* indicated that while

47% of test takers answered correctly with no regular pattern to the choosing of other distractors. According to Bond and Fox[27] this item's fit statistics indicate that it fits the model too closely.

As for outfit, five items, O1, *compound* (Outfit Mnsq = 2.14, Outfit Zstd = 1.8), F9, *standard* (Outfit Mnsq = 1.57, Outfit Zstd = 2.5). O10, *allege* (Outfit Mnsq = 1.47, Outfit Zstd = .8), T9, *rove* (Outfit Mnsq = 1.44, Outfit Zstd = 1.1), and O3, *candid* (Outfit Mnsq = 1.39, Outfit Zstd = 1.1) all displayed statistics indicating underfit. This means that the responses were too erratic to be meaningfully predicted by the Rasch model. The high percentage of lower frequency words suggests it is likely that low ability students guessed these questions correctly, creating misfit[28]. An additional five items, O5, *quiz* (Outfit Mnsq = .30, Outfit Zstd = −.4), F8, *shoe* (Outfit Mnsq = .54, Outfit Zstd = −.6), F5, *poor* (Outfit Mnsq = .31, Outfit Zstd = −.7), F6, *drive* (Outfit Mnsq = .31, Outfit Zstd = −.7), and S9, *microphone* (Outfit Mnsq = .66, Outfit Zstd = −1.1) displayed statistics indicating overfit. These statistics indicate performances that are "too good to be true" and likely due to lack of item independence. All of these words are L1 cognates, so students were likely to have gotten these questions correct reliably across the full population.

Construct validity: Substantive facet

To determine the substantive validity of the VST, the degree to which the items adhered to the difficulty estimate hypothesis was examined. It was assumed that each consequent 1,000–word list would be increasingly more difficult than the one preceding it. This was based on the idea that the frequency of exposure to linguistic items can have a direct impact of acquisition[29]. Mean difficulties were calculated for each of the 4–sets of ten questions. The results can be seen in table 4.

As predicted, the first 1,000 level items displayed the lowest difficulty estimate. Measurements of the lower three frequency levels, however, did not conform to the hypothesis. There are several possible explanations for this phenomenon. The first is that development of the students' vocabulary knowledge has does not adhere to frequency predictions because of the vocabulary they have studied. Two items in the 4,000–word list (*O5* = quiz; *O8* = vocabulary) are words which students have undoubtedly had a number of exposures disproportionate to standard frequency predictions due to classroom language learning, thus resulting in a lower difficulty. A second explanation is the presence of L1 cognates present in the higher frequency level lists (e.g. S9 = *microphone;* T8 = dash; O5 = *quiz;* O6 = input O8 = vocabulary). Coxhead et. al[30] suggest that, though the presence of cognates may skew item difficulty predictions based

Table 4　*Calibrated item difficulty in logits*

Item number within level	First 1,000 (F)	Second 1,000 (S)	Third 1,000 (T)	Fourth 1,000 (O)
1	−5.2	1.3	−0.4	2.4
2	−5.2	−1.3	1.3	2.7
3	0.8	0.2	0	1.8
4	0.2	0.6	0.2	0.8
5	−3.2	2.7	−0.5	−4
6	−3.2	1.4	−0.5	−1.4
7	−1.4	−0.9	0.4	−0.2
8	−2.5	0.2	−2.8	−1.6
9	−0.5	1.6	2	0.9
10	0.8	−0.5	−0.8	3.5
Average	−1.94	0.53	−0.11	0.49
SD	2.16	1.18	1.22	2.19

on frequency, it would be inappropriate to remove them completely from the test as cognates are part of the learners' true vocabulary size. Beglar[31] also found that the difficulty of second through fourth 1,000 word lists items were impossible to distinguish and that it was only at higher levels that incremental difficulty was observable, so perhaps the second, third and fourth lists need to be reexamined for the purpose of offering them independently of the full 140–item test.

A second way that the substantive aspect was examined was by examining whether learner English ability was a predictor for performance on the VST. Data from the six students with experience in English L1 schooling (2, 36, 37, 39, 4 and 35) were separated from the pool, their abilities were averaged, and compared to the rest of the students. The students pulled from the pool had an average ability of 1.6 logits ranging from 1.0 to 2.7 logits. The other 37 students' ability average was −.18 logits, ranging from −1.5 to .9 logits. Learners with experience living in the English L1 environment scored, on average, much higher than other students. These results seem to corroborate Beglar's[32] findings that higher English ability is a predictor of success on the VST.

Construct validity: Structural aspect

The structural aspect of construct validity was examined by measuring the psychometric unidimensionality of the test items. This can be examined by looking at item fit in conjunction

with the total amount of variance accounted for by the Rasch model[33] The total amount of raw variance accounted for by the Rasch model was 37.7% , which is well below the 60% threshold of acceptability proposed by Linacre[34]. Of the 62.3% of the variance unexplained by the model, variance explained by the first five contrasts was 7.4% , 5.5% , 5.2% 4.7% and 4.1% , indicating the likelihood of at least one other meaningful psychometric dimension. Suggestion of a presence of an additional psychometric dimension is unsurprising considering the large number of cognates evenly distributed throughout the four levels of the test. Assisting knowledge from the L1 is a likely culprit for causing variance that could not be adequately predicted by the Rasch model.

Conclusion

The purpose of this study was to explore validity evidence for one form of *Vocabulary Size Test* with a group of younger learners. Rasch analysis results indicate that this version of the test has a sufficiently robust item hierarchy to measure the vocabulary size of this particular group of students. A large number of items, however, showed significant misfit, likely due to the large number of L1 cognates and a large number of students guessing. The data suggest that the relationship between BNC corpus frequency and item difficulty is tenuous for items the 2nd, 3rd and 4th lists. However, this test may show a correlation between prolonged exposure to English and greater vocabulary knowledge with a more robust data set. With some items rewritten to reduce the impact of cognates on unexplained variance, this test would be a more meaningful measure of learner vocabulary size for teachers of a similar cohort of students.

There were several limitations of this investigation. First and foremost, the researcher's limited understanding of the Rasch model and interpretation of output data severely handicapped a meaningful analysis of results. In addition, the small sample size of learners and small item pool may have exacerbated small discrepancies in the data. Further studies could examine the generalizability aspect of the construct validity of this test by testing for differential item functioning on this test, either by randomly dividing the test into two equal parts and testing them against each other, or by comparing the results of this test with the results of the same test with a different cohort. Further following Beglar's[35] line of investigation, this study's data could be analyzed for responsiveness and interpretability. A fuller analysis of this test data would shed more light on the usefulness of the VST for measuring the vocabulary sizes of young learners.

Notes

(1) Michael West, *A general service list of English words* (London: Longman, Green & Co., 1953)
(2) Norbert Schmitt and Diane Schmitt, "A reassessment of frequency and vocabulary size in L2 vocabulary teaching" in *Language Teaching* 47, no. 4 (2012): 5.
(3) Paul Nation, "How large a vocabulary is needed for reading and listening" in *Canadian Modern Language Review* 63, no. 1 (2006): 70.
(4) Paul Nation, "Testing and teaching vocabulary" in *Guidelines* 5, no. 1 (1983): 19.
(5) Paul Nation and David Beglar, "A vocabulary size test" in *The Language Teacher* 31, no. 7 (2007): 11.
(6) Paul Nation, *Learning vocabulary in another language* (Cambridge: Cambridge University Press, 2013): 36.
(7) David Beglar, "A Rasch-based validation of the Vocabulary Size Test" in *Language Testing*, 27 (2010): 116.
(8) Beglar, "Rasch-based", 105–113.
(9) Samuel Messick, "Validity of psychological measurement: Validation of inferences from persons' responses as scientific inquiry into score meaning" in *American Psychologist*, 50 (1995): 742.
(10) Beglar, "Rasch-based", 116.
(11) Nation and Beglar, "size test", 11.
(12) Nation, "How large", 62.
(13) Nagy *et al.*, "Morphological families in the internal lexicon" in *Reading Research Quarterly*, 24 (1989): 275.
(14) Paul Nation and Stuart Webb, *Researching and analyzing vocabulary* (Boston: Heinle, 2011), 136.
(15) Laurie Bauer and Paul Nation, "Word families" in *International Journal of Lexicography*, 6, (1983), 254.
(16) Nation and Beglar, "A vocabulary", 10.
(17) Nation and Beglar, "Size test", 11.
(18) Beglar, "Rasch-based", 104.
(19) Michale Linacre, "Winsteps (3.81.0)". Windows. Michael Linacre, 2007.
(20) Beglar, "Rasch-based", 105.
(21) Messick, "Validity", 745.
(22) Nation, "How large", 63.
(23) Nation, "How large", 63.
(24) Beglar, "Rasch-based", 107.
(25) Judith Runnels, "Using the Rasch model to validate a multiple choice English achievement test" in *Journal of Applied Measurement*, 2, no. 3 (2012), 146.
(26) Trevor Bond and Christine Fox, *Applying the Rasch model: Fundamental measurement in the human sciences 3rd ed.*(New York: Routledge, 2015): 272.
(27) Bond and Fox, "Applying", 273.
(28) Bond and Fox, "Applying", 273.
(29) Nick Ellis, "Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition" in *Studies in Second Language Acquisition*, 24 (2002): 145.
(30) Averil Coxhead, Paul Nation, and Dalice Sim, "Creating and trialling six versions of the Vocabulary Size Test" in *TESOLANZ Journal*, 22 (2014), 14.
(31) Beglar, "Rasch-based", 109.

(32) Beglar, "Rasch-based", 110.
(33) John M. Linacre, "A user's guide to WINSTEPS" Retrieved from www.winsteps.com (2017), 196.
(34) Linacre, "User's guide", 386.
(35) Beglar "Rasch-based", 113–114.

Bibliography

Bauer, Laurie, and Paul Nation. "Word families." *International Journal of Lexicography* 6, (1983): 253–279.

Beglar, David. "A Rasch-based validation of the Vocabulary Size Test." *Language Testing* 27, (2010): 101–118.

Bond, Trevor G., and Christine M. Fox. *Applying the Rasch model: Fundamental measurement in the human sciences.3rd* ed. New York: Routledge, 2015.

Coxhead, Averil, Paul Nation, and Dalice Sim. "Creating and trialing six versions of the vocabulary size test." *TESOLANZ Journal* 22, (2014): 13–27.

Ellis, Nick. "Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition." *Studies in Second Language Acquisition* 24, (2002): 143–188.

Linacre, Michael. "A user's guide to WINSTEPS ver. 3.81.0." Chicago: www.winsteps.com/ (2007)

Linacre, Michael. "Winsteps (3.81.0). Windows. Michael Linacre, 2007.

Messick, Samuel. "Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning." *American Psychologist* 50, (1995): 741–749.

Nagy, William E., Richard C. Anderson, Marlene Schommer., Judith A. Scott., and Anne C Stallman. "Morphological families in the internal lexicon." *Reading Research Quarterly* 24, (1989): 263–282.

Nation, Paul. "Testing and teaching vocabulary." *Guidelines* 5, no. 1 (1983): 12–25.

Nation, Paul. "How large a vocabulary is needed for reading and listening." *Canadian Modern Language Review* 63, no. 1 (2006): 59–82.

Nation, Paul. *Learning vocabulary in another language*. Cambridge: Cambridge University Press, 2013.

Nation, Paul, and Stuart Webb. *Researching and analyzing vocabulary*. Boston: Heinle, 2011.

Nation, Paul, and David Beglar. "A vocabulary size test." *The Language Teacher* 31, no. 7 (2007): 9–13.

Runnels, Judith. "Using the Rasch model to validate a multiple choice English achievement test." *International Journal of Language Studies* 6, no. 4 (2012): 141–153.

Schmitt, Norbert, and Diane Schmitt. "A reassessment of frequency and vocabulary size in L2 vocabulary teaching." *Language Teaching* 47, no. 4 (2012): 484–503.

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

聖学院大学論叢　第 30 巻　第 2 号　2018 年

高校生対象ラッシュモデルに基づく Vocabulary Size Test の妥当性検証

ローランド・ロバート・J. S

抄　録

本研究の第一目標は省略された 40 問の Vocabulary Size Test（VST）の妥当性を検証することでたった。第二目標は Beglar（2010）の同テストの妥当性の主張を強固させることであった。VST を 15 から 16 歳の高校生（N=43）で実施した。データ分析をラッシュモデル測定で行って，その結果を Messick の妥当性の定義を対象に調査した。本研究の結果は以下のことを示した（1）問題と受験者は概ね仮説に基づいた通りであった（2）ほとんどの問題はラッシュモデル測定に一致した（3）ラッシュモデル測定妥当性程度の低いものについては，日本語に外来語であるもの，英語学習環境によく出る言葉のいずれかで説明することができた。

キーワード：英語の語彙サイズ，ラッシュ・モデル，妥当性，一次元的