The nature of vocabulary knowledge in EFL assessment

Ken Norizuki

Vocabulary acquisition plays an important part at all stages of language learning experience, and is nonetheless typically seen as a restricted domain of language assessment. The present study explores some simple measures of vocabulary knowledge as potentially powerful tools of EFL classroom assessment. In the first part a hundred-item vocabulary test was constructed and administered to the first-year university students for the purpose of re-placing them into adequate general language proficiency streams for their second-year EFL course. The results were compared with those of the language test battery used for the first-year placement decision. The analysis of item statistics led the author to the construction of four mini versions of vocabulary knowledge assessment instruments which were designed to tap developmental dimensions of vocabulary knowledge using the IRT Rasch analysis. The overall findings indicate that vocabulary research merits further pedagogical attention from all EFL stakeholders.

I. Background

1. Vocabulary research

Research on vocabulary assessment starts by determining how to define vocabulary knowledge and to measure the construct. According to Anderson and Freebody (1981), cited by Read (2000), it is useful to distinguish between the breadth, or size of vocabulary knowledge, on the one hand, and the depth, or quality of that knowledge, on the other hand.

Assessing the breadth of vocabulary knowledge raises a question of how to choose which words to test. Obviously, it is not a realistic idea to test all the words that any individuals might or might not know. Alternatively, one may opt to construct a word list arranged by frequency or difficulty levels, from which words will be randomly selected for the test, and to design a measure of vocabulary knowledge defined.

The Eurocentres Vocabulary Size Test (EUVST), published by Meara and Jones (1990), is one solution. The EUVST is a Yes/No test, which requires the test-takers to respond "Yes" or "No" to words they know or ones they do not. About a third of the items in the test are 'non-words' (Read 2000), to counterbalance the effect of "invalid Yes" responses. This type of test can include more items than conventional multiple-choice vocabulary tests and is easier to construct (Shillaw 1996). (See Barrow, et al. 1999 for a similar survey conducted over Japanese college students).

Another test of interest is the Vocabulary Levels Test (VLT), devised by Nation (1990). Unlike the EUVST, all the items in the test are real words and the test-takers are required to match the words to the definitions. The VLT has gained esteem worldwide and has been found to correlate well with more productive measures of vocabulary knowledge (Laufer and Nation 1995).

A major pitfall shared by the above tests is that these tests in particular and these test types in general may not seem to suit learners of English at lower intermediate or lower levels, especially in exposure-poor, EFL learning backgrounds. Having to look through a large number of words mixed with non-words or to identify synonyms from the numerous distractors, low proficient learners are likely to be hampered by their poor reading ability.

Meanwhile, assessing the depth of vocabulary knowledge demands multiple measures of vocabulary knowledge. Laufer (1998) and Laufer and Paribakht (1998), for example, employed three different measures of vocabulary knowledge, which were supposed to test passive, controlled active, and free active dimensions.

This line of research merits further investigation in exposure-limited EFL backgrounds. Some modifications may need to be made to include more passive and more controlled dimensions of its knowledge.

Another approach to vocabulary knowledge depth is the use of the Vocabulary Knowledge Scale (VKS), defined by Paribakht and Wesche (1993). The scale is presented with a list of words, and the subjects are asked to self-rate their knowledge of the words in five categories (Read 2000). This type of self-assessment questionnaire may well be combined with tests of vocabulary knowledge.

With all the issues outlined above, it is natural to expect that vocabulary assessment can be more integrative and global than it appears to be and it alone can serve as an easier alternative to a language test battery in a relatively low-stakes test setting. (See Meara and Jones 1988). This possibility is worth discussing in later sections.

2. Purpose

The purpose of this study is to explore the usefulness of different types of vocabulary assessment instruments. To that end, the following three research questions were formulated.

1) Can a vocabulary test be a reliable and valid measure of general language proficiency?
2) Can different types of vocabulary assessment procedures test the breadth and depth of vocabulary knowledge in a consistent and meaningful manner?
3) Can a self-assessment questionnaire work well with the corresponding tests of vocabulary knowledge?

II. Method

1. Participants

The participants in the first part of this study were 134 students who enrolled for the first-year EFL course at the Faculty of Management, Shizuoka Sangyo University in the 1998 academic year. All of them analyzed took a language placement test battery at the entry of the course and a vocabulary test at the exit of the course.

The participants in the second part of the study were 31 students who enrolled for the second-year EFL course at the above-mentioned institution in the 1999 academic year. The majority of them were those who failed in the course in previous years; the rest were third-year transfer students and second-year students. They completed all three different measures of vocabulary knowledge analyzed in the present study.

2. Materials

The materials in this study were the following five methods of vocabulary assessment developed by the author.

1) A 100-item vocabulary test (VPT: Vocabulary Placement Test) was developed as part of a placement procedure for those who wished to attend the second-year EFL course in the 1999 academic year. The first 80 items in the test require the students to match single English words with Japanese translation equivalents accompanied by English synonyms or definitions for exchange students who might prefer to identify TL equivalents. The words were the university learners' most essential, two-asterisk or single-asterisk words sampled from Kiyokawa's (1988) word list. The remaining 20 items in the test were matching of idiomatic expressions with single-word equivalents.

2) Four versions of 20-item vocabulary assessment tools were constructed to measure different dimensions of vocabulary knowledge. The 20 single-words were selected on the basis of item difficulty and discrimination statistics obtained from the VPT analysis. The first version (V1) was designed to test controlled active vocabulary knowledge, following the format adopted by Laufer (1998) and Laufer and Paribakht (1998). A substantial number of candidates had to be eliminated, though, because of their zero raw scores, which were not amenable to Rasch scale analysis intended. Eventually, this version of the assessment method was excluded from the present analysis.

3) The second version of the method (V2) was designed to test "relatively" passive vocabulary knowledge. The test-takers were asked to match English words with synonymous English expressions. What is meant by being relatively passive is that students need to activate processing of their TL vocabulary knowledge in semantic networks, which is assumed to require more active knowledge than TL-L1 matching items.

4) The third version of the method (V3) was intended to measure the most passive dimension of vocabulary knowledge analyzed here. The test comprised of TL-L1 matching, i.e., translation items.

5) The fourth version (V4) was a self-assessment questionnaire in which the students were asked to choose one of four steps or categories that best represents how well they know the words in the test. At step 0 the word is not recognized at all. Step 1 means that the word is not known but the meaning could be inferred from the word form or context. Step 2 means that the word is known but cannot be used in writing. At step 3 the word is known and can be used in writing. The questionnaire was placed immediately after the TL-L1 matching test on the same survey sheet. Ten to fifteen minutes of regular class sessions were used respectively for versions 1 and 2. About fifteen minutes of a single class session was used for versions 3 and 4.

3. Data Analysis

Descriptive statistics were calculated for the VPT sample alongside a KR-20 reliability and a correlation as a measure of external validity. The V2 and V3 test items were scored as correct or incorrect and underwent an IRT Rasch dichotomous analysis (Wright and Stone 1979; Henning 1987). The V1 items, which were scored both dichotomously and polytomously with partial credits given for sensible, incorrect responses, were excluded from the present study for the reason mentioned above. The V4 questionnaire was analyzed with the Rasch partial-credit model (Wright and Masters 1982; Pollitt and Tang 1994; McNamara 1996) to capture the perceived distance between one step and the next. The item step threshold was adopted as cumulative step difficulty for the sake of vertical comparisons between different measures of item and step difficulty.

The Rasch analysis was conducted using the computer program MINIFAC, version 3.2. The MINIFAC is an abridged version of FACETS, which "has the same functionality, but is limited to the number of observations it can analyze" (Linacre 1998).

III. Results

1. The VPT

Two months prior to the VPT administration, at the entry of the 1998 EFL course, a 40-item language placement test battery (LPTB) was conducted over students who intended to enroll for the course. The test, which consisted of vocabulary, short conversation, structure, reading and listening subtests, was claimed to be fairly reliable (KR-20 =.831), and valid, as demonstrated by moderately high intercorrelations among the five subtests (Norizuki 1998).

The VPT was produced as an alternative method to the LPTB for the placement purpose as mentioned in the preceding section. The VPT has some advantages over the LPTB. First, the VPT was much less demanding to construct and could be revised more easily. Second, the VPT accommodates more than twice as many items as LPTB, with virtually the same or less amount of time required (30 minutes) for the learners to complete the test. Third, the availability of a larger number of items means that a potentially larger amount of feedback could be provided to the test-takers.

As shown in Table 1, the VPT was a highly reliable test, with a KR-20 coefficient of .94. The VPT correlated fairly well with the LPTB (.76), even with a two-month interval between the two tests. This indicates that the VPT could be a good predictor of the learners' general language proficiency. In other words, the VPT could stand alone as a criterion measure for such a placement decision in a relatively low-stakes testing situation.

Item difficulty indices ranged from 1.0 to .25. Five items were answered correctly by all the test-takers. In a strict statistical view, these extreme-score items should be deleted from the test, especially when the Rasch analysis is intended. The inclusion of such items might well be justified if they help ease anxiety and frustration observed among low proficient test-takers.

Meanwhile, item discrimination indices (point-biserial correlations) ranged from -.11 to .69. According to Ebel (1979), cited by Beglar and Hunt (1999), items below .19 need to be revised or eliminated. Twelve items, including five perfect-score items, are under suspicion on this ground. All of them are items of extreme easiness, with item difficulties of .99 or higher, thus allowing even low proficient learners to answer correctly. As the depressed discrimination indices were not due to the unexpected response patterns, they could be retained for the reason mentioned above.

Table 1 Descriptive statistics and reliability and validity estimates for the VPT

Mean	SD	Number of items	Number of persons	KR-20	Corr. with the LPTB
72.75	15.12	100	134	.94	.76**

**:p<.01

2. V2 - 4

Table 2 shows twenty words sampled from the VPT and their item difficulty and discrimination statistics obtained from the analysis. These words to be measured in different dimensions of vocabulary knowledge were intended to represent different parts of speech (nouns, verbs, adjectives, etc.) and varied levels of difficulty (.28 to .96). All the item discrimination indices exceeded .20 (.22 to .61).

The V2 (synonym matching) and V3 (translation matching) tests were analyzed with the Rasch dichotomous model. Table 3 displays the V2 and V3 item difficulty indices expressed on the latent scales. In most cases, the V3 items were found easier than the V2 items. The V3 mean item difficulty was almost identical to the mean ability level centered at 0.00 in this analysis.

Table 2 Twenty words sampled from the VPT and their item statistics

	capital	president	another	technical	wrong	serious	hardly	improve
DIFF	.96	.85	.92	.72	.78	.68	.66	.67
DISC	.22	.29	.37	.26	.34	.50	.54	.47
	furniture	narrow	attitude	environment	proof	below	audience	region
DIFF	.52	.65	.51	.66	.43	.63	.47	.58
DISC	.47	.37	.41	.55	.47	.61	.41	.43
	establish	gain	indicate	permit				
DIFF	.31	.50	.28	.35				
DISC	.39	.35	.36	.35				

Table 3 The V2 and V3 item difficulty indices (Rasch dichotomous analysis)

	capital	president	another	technical	wrong	serious	hardly	improve
V2	-1.19	-1.83	.20	- .68	-.23	.65	2.12	.98
V3	-1.01	-1.31	-2.10	-1.59	-.38	-.24	1.58	.65
	furniture	narrow	attitude	environment	proof	below	audience	region
V2	1.17	.20	2.47	.06	.20	.20	.65	1.17
V3	1.37	-.69	1.37	.98	-.09	-.24	.20	.81
	establish	gain	indicate	permit	Mean	SD		
V2	.98	1.58	2.12	.65	.57	1.06		
V3	1.37	-.09	.50	-.53	.02	1.03		

Table 4 The V4 self-assessment questionnaire item step thresholds (Rasch partial-credit analysis)

	capital	president	another	technical	wrong	serious	hardly	improve
S1	-1.96	-1.55	-1.75	-2.02	-1.19	-1.23	-2.46	-1.07
S2	.05	- .05	.13	.17	.38	.43	.75	.45
S3	1.91	1.59	1.61	1.84	.87	.85	1.70	.70
	furniture	narrow	attitude	environment	proof	below	audience	region
S1	- .98	-1.39	-.98	-1.85	-1.04	-1.93	-.94	-1.21
S2	.47	.33	.98	.56	.29	.60	.43	.24
S3	.63	1.09	*	1.30	.80	1.34	.62	.99
	establish	gain	indicate	permit	Mean	SD		
S1	-1.19	-1.53	-1.40	-1.08	-1.44	.33		
S2	.19	.39	.32	.04	.36	.25		
S3	1.01	1.16	1.10	1.04	1.17	.40		

*No subjects rated themselves as having reached step 3 for this item.

Table 5 Intercorrelations among V2 and V3 difficulty and V4 threshold estimates

	V2	V3	V4S1	V4S2	V4S3
V2		.73**	.23	.66**	-.42
V3			.24	.66**	-.48*
V4S1				-.00	-.89**
V4S2					-.27

**:p<.01; *:p<.05

The V4 self-assessment questionnaire was subjected to the Rasch partial-credit analysis. Unlike the difficulty parameter for the dichotomous model, the partial-credit step difficulty parameter refers to how much more ability is required to progress from one step to the next. Table 4 presents the cumulative difficulty of a particular step, or item step threshold as an alternative measure, which is defined as "the ability level that is required for an individual to have a 50 per cent chance of passing that step" (McNamara 1996). The use of item step threshold enables us to interpret step difficulty in relation to the learners' ability.

Intercorrelations among the V2 and V3 difficulty estimates and the V4 step thresholds are reported in table 5. The V2 and V3 test measures were fairly well correlated (.73). These two test measures had moderately high correlations with the self-assessment questionnaire step 2 measures (.66 both). In contrast, negative correlations were found between the test measures and the questionnaire step 3 measures (-.42 and -.48). The results suggest that the learners' perceived ability to understand the word (step 2) is well represented by synonym-matching and translation-matching tests while the ability to use the word (step 3) is not well recognized by the learners and/or is not positively affected by the ability to tackle passive vocabulary tests. As step 3 might have been better reflected by the test of controlled active vocabulary knowledge which has been excluded from the present analysis, it is hard to reconcile the data with the analysis of the other measures.

Figure 1 is a graphical summary of V2 and V3 difficulty estimates and the V4 step 1 and step 2 thresholds.

Figure 1 Graphical summary of dimensions of vocabulary knowledge

	V2: synonym matching	V3: translation matching	Step 1: guess the word (SA)	Step 2: understand the word (SA)	
HARD 2.5	attitude				HARD 2.5
2	hardly/indicate				2
1.5	gain	hardly furniture/attitude/establish			1.5
	furniture/region				
1	improve/establish	environment region		attitude	1
0.5	serious/audience/permit	improve indicate		hardly below environment serious/improve/furniture/audience wrong/narrow/gain/indicate proof/region another/technical/establish capital/president/permit	0.5
	another/narrow/proof/below	audience			
0	environment	proof/gain			0
	wrong	serious/below wrong			
-0.5	technical	permit narrow			-0.5
-1	capital	capital	furniture/attitude/audience improve/proof/permit wrong/establish serious/region narrow indicate president/gain		-1
		president			
-1.5		technical			-1.5
	president		another environment capital/below technical		
-2		another			-2
-2.5 EASY			hardly		-2.5 EASY

Aligning the four scales in this manner enables us to compare the relative difficulty measured between synonym-matching and translation-matching items, on the one hand, and the perceived difficulty involved in guessing the word (V4 Step 1) and understanding the word (V4 Step 2), on the other hand.

As for the test item measures, synonym-matching items were generally more difficult and more widely distributed than translation-matching items in the difficulty scales. Despite some differences in the order of difficulty, the overall picture depicts the possibility that there might be a latent developmental sequence of vocabulary acquisition.

The two steps of the questionnaire items, associated with guessing the word and understanding the word, were found to function distinctively, as displayed by two separate clusters along the scales. This corroborates our intuition that tapping different dimensions of vocabulary knowledge is a key to synthesize a complex picture of vocabulary acquisition.

Looking at individual words listed carefully, one can notice that some words were consistently high or low in the light of the difficulty order while others varied a lot from one scale to another. The word, "attitude", for example, was measured and rated as one of the most difficult words in all the four scales. When learners come across many words of this difficulty level in reading, they may need to benefit from explicit input to facilitate their text comprehension. On the contrary, the word, "hardly" was found to be one of the most difficult in the two test scales and the self-assessment step 2 scale, but was found by far the easiest word in the step 1 scale. A cause for this mismatch is readily accountable from a fact that a large proportion of participants in this study misinterpreted "hardly" as synonymous with "difficult" in the synonym-matching and translation-matching tests. When the learners lack the knowledge of the word, they may try to guess the word with reference to its form or others, which often ends in failure.

IV. Discussion

The first research question concerned the reliability and validity of a vocabulary test as a measure of general language proficiency. The VPT recorded a very high reliability coefficient and had a fair amount of correlation with a language test battery as a valid measure of general language proficiency. The test can possibly work well as a convenient measure of general language proficiency in a low-stakes testing situation.

The second research question addressed the issues of the breadth and depth of vocabulary knowledge. The VPT of its kind proves to be a potentially good and efficient means of assessing the breadth of vocabulary knowledge. From our future research perspectives, this type of test may be revised and expanded through a variety of item and test analysis techniques increasingly made available to test users. The advent of IRT analysis, for example, allows items derived from different tests to be equated on a common latent scale. This analysis is no longer bounded by a particular sample of test items or a particular group of test-takers. With a large number of items pooled in a test bank, multiple forms of a vocabulary test given to different groups of examinees can be compared to one another, enabling the test users to make more reliable estimates of the breadth of the learners' vocabulary knowledge (Beglar and Hunt 1999).

The depth of vocabulary knowledge was explored with applications of IRT Rasch analysis to different tests of vocabulary knowledge and a self-assessment questionnaire. A test of controlled active vocabulary knowledge was found to be too difficult for the learners examined and had to be excluded from our analysis. This poses a problem of eliciting controlled active vocabulary knowledge in a context-deprived prompt. In a similar vein, the questionnaire's highest response category (Step 3) correlated negatively with the other measures and was judged to be unsuitable for the comparative analysis. The remaining measures functioned consistently well in terms of the overall difficulty order. Synonym-matching items were generally more difficult than translation-matching items, and the questionnaire step 1 and step 2 measures formed two distinct clusters along the common difficulty scales. A few words of fluctuating difficulty order, as in the case of the word, "hardly" might be ascribable to the learners' errors associated with conceptualization at some stages of vocabulary acquisition. Teachers need to treat these words with caution and to think of offering the learners diagnostic feedback.

The third research question relates to the usefulness of the self-assessment questionnaire in association with the tests. The questionnaire's moderately high correlations with the tests might justify its use. This being so, the questionnaire step threshold estimates obviously differ in distribution pattern from the test counterparts. The two narrowly-packed clusters along the scales indicate that the learners' perceived difficulty was not yet as accurate as the tested difficulty. This "inability" to know oneself well is a matter of human nature and should not necessarily be regretted. By sensitizing the learners to both the breadth and the depth of their vocabulary knowledge, self-assessment is expected to exercise a definitive role in vocabulary learning as well as vocabulary assessment.

V. Conclusion

Vocabulary knowledge is of general concern for learners at all stages of linguistic development, and has nevertheless been deemed as a rather peripheral area of interest in language assessment. The present study exhibits some directions to integrate simple, somewhat crude measures of vocabulary knowledge as potentially powerful tools of EFL assessment. This kind of vocabulary research should prove useful to EFL test developers and teaching professionals.

The study is still at its preliminary stage and, as such, some of its limitations should be acknowledged. First, being able to identify one right definition for a given word, the learner may not understand the word used in other meaning senses (Schmidt 1999) or in derived and inflected forms (Beglar and Hunt 1999). Second, only a small number of words could be selected out of hundreds of or thousands of words in the learners' potential lexicon, especially in the comparative latent trait analysis. Third, probably most importantly, the relationship between the breadth and the depth of the individual learners' vocabulary knowledge remains a hypothetical construct, and needs to be determined precisely in relation to specific educational goals.

These problems will be major educational issues for all EFL stakeholders involved. It is hoped that vocabulary assessment research will play a prominent role to help bridge the gap between applied linguistic theory and EFL teaching practices.

VI. References

Barrow, J., Nakanishi, Y. and Ishino, H. 1999: Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System* 27, 223-47.

Beglar, D. and Hunt, A. 1999: Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing* 16, 131-62.

Henning, G. 1987: *A Guide to language testing.* Cambridge, Mass.: Newbury House.

Kiyokawa, H. 1988: Koukou-daigaku-you readability koushiki no kaihatsu. Wayo Women's University: *Language and Literature* 22, 43-63.

Laufer, B. 1998: The development of passive and active vocabulary in a second language: Same or different?, *Applied Linguistics* 19, 255-71.

Laufer, B. and Nation, P. 1995: Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics* 16, 307-22.

Laufer, B. and Paribakht, T.S. 1998: The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning* 48, 365-91.

Linacre, J.M. 1998: *A user's guide to FACETS* (Version 3.1). Chicago, Ill.: MESA Press.

McNamara, T.F. 1996: *Measuring second language performance.* London: Longman.

Meara, P. and Jones, G. 1988: Vocabulary size as a placement indicator. In Grunwell, P., editor, *Applied linguistics in society,* London: Centre for Information on Language Teaching and Research, 80-7.

Meara, P. and Jones, G. 1990: *Eurocentres Vocabulary Size Test* (VersionE1.1/K10). Zurich:Eurocentres Learning Service.

Nation, I.S.P. 1990: *Teaching and learning vocabulary.* New York: Heinle and Heinle.

Norizuki, K. 1998: Research issues on university English language testing. Shizuoka Sangyo University: *Kankyo to Keiei* 4, 199-208.

Paribakht, T.S. and Wesche, M. 1993: Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal* 11, 9-29.

Pollitt, A. and Tang, G. 1994: Language testing and second language acquisition research. In Boyle, J. and Falvey, P., editors, *English language testing in Hong Kong,* Hong Kong, The Chinese University Press.

Schmidt, N. 1999: The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing* 16, 189-216.

Shillaw, J. 1996: The application of Rasch modelling to yes/no vocabulary tests. Vocabulary Acquisition Research Group discussion document No.js96a, available over the Internet at <www.swan.ac.uk/cals/vlibrary/js96a.htm>

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

Wright, B.D. and Masters, G.N. 1982: *Rating scale analysis.* Chicago, Ill.: MESA Press.

Wright, B.D. and Stone, M.H. 1979: *Best test design.* Chicago, Ill.: MESA Press.