Vocabulary depth in written and oral assessment

Aaron Olaf Batty
Kanda University of International Studies

Reference Data:
Batty, A. O. (2007). Vocabulary depth in written and oral assessment.
In K. Brandford-Watts (Ed.), JALT *2006 Conference Proceedings*. Tokyo: JALT.

Measures such as the word associates test (WAT) have been developed to address the construct of vocabulary knowledge depth, but the question of how well traditional written and oral test tasks address it within a large-scale, medium-stakes exam has heretofore been unexplored. This study examines this question in the context of the Kanda English Proficiency Test (KEPT), which employs a timed-essay writing task and an oral task, both of which are assessed for vocabulary. A known valid and reliable WAT was administered to 198 university students two weeks prior to KEPT administration, the resultant data were scaled using the Rasch model, and multiple regression was performed. Although performance on the WAT was found to be significantly predictive of vocabulary scores on the written section of the KEPT, the relationship with the oral vocabulary score was non-significant. The paper concludes with a discussion of the implications of these results.

Vocabulary knowledge depth（語彙知識の深さ）を判断するために、word associates test（単語と単語の関連を知るテスト）（WAT）のようなテストが作られてきました。しかし、１０００人以上の大規模な人数によるmedium-stakes テスト（high-stakesは入学試験レベル、low-stakesは授業の中の単語試験レベル）の範囲内で、従来のライティングやスピーキングの試験を使って、どの程度うまく語彙知識の深さを判断できるか、誰も研究していませんでした。この研究は神田外語大学英語能力テスト（KEPT）を使用して、語彙知識の深さを判断しようと試みたものです。KEPTは制限時間のあるライティングテストを用い、それによって語彙力の評価をします。またスピーキングテストも行い、それもまた語彙力を評価します。２００６年１月、１９８人の神田外語大学の学生にKEPT を実施しました。またKEPTを実施する２週間前に、valid and reliable（有効で信頼性のある）WATも、同じ生徒たちに実施しました。それらのテスト結果のデータはRaschモデルにより分析されました。また、従来の英語能力を測るテストは、語彙知識の深さを測ることが出来るかどうかを知るために、KEPTの結果データとWATの結果データを重回帰分析を用いて調べました。WATにおける得点は、KEPTのライティング部門の語彙の得点に統計的に有意に関係していることが分かりましたが、スピーキング部門の語彙の得点においては、その関係が見られませんでした。この論文は、これらの結果に関する考察について書かれています。

Depth of vocabulary knowledge (DVK)

Vocabulary knowledge can be conceptualized as having two basic dimensions: breadth and depth. Whereas breadth is *how many* vocabulary items a learner knows; depth is understood as *how well* said learner knows them (Paribakht & Wesche, 1996; Qian, 2002). For a learner to truly know a word, he or she must know many things *about* it: spelling, morphology, acceptable inflectional use, word class, etc. Although linguists and cognitive scientists have devised complex-but-precise means of describing a word's semantic features (Hatch & Brown, 1995), others, such as Miller and Fellbaum (1991), have proposed a *semantic network* model of word knowledge, wherein a variety of facts about a word combine to create the full meaning of said lexical item. These approaches to word meaning, no matter how complete or useful to the researcher, are perhaps beyond what the typical teacher or student wants or needs when approaching the question of vocabulary knowledge depth. For this reason, Nation (1990) developed an oft-cited simple list of eight types of word knowledge, ranging from pronunciation through *associations,* such as synonyms and other words which typically occur with a given word. It separates word knowledge into both receptive and productive knowledge, and approaches vocabulary knowledge from the dimensions of form, position, function, and meaning. Furthermore, it is critical to remember that words do not exist in isolation and that collocations are a key type of vocabulary knowledge. Hunston, Francis, and Manning (1997) have argued that a word's meaning often depends on factors beyond the word boundary, into the sentential context. This is a key component to word knowledge, and key to *nativelike* selection (Pawley & Syder, 1983).

DVK assessment

The question of how best to assess knowledge of these various forms of word knowledge (i.e. depth of vocabulary knowledge—DVK) has been debated since the rise of the field of psychometrics and objective testing (Read, 2000). The central problem has typically been one of logistics: how to test so many types of knowledge about a single lexical item?

This problem has been most popularly addressed by John Read's word associates test (WAT) originally proposed in 1993 and revised in 1998, which seeks to strike a balance between DVK and size of words sampled. In the original version, the examinee was presented with a stimulus word followed by eight possible *associates.* These associates fell into three categories:

1. *Paradigmatic,* wherein the words are synonymous or otherwise similar in meaning. This category includes synonyms, hypernyms, meronyms, etc.
2. *Syntagmatic,* wherein the words are collocates often appearing together in a sentence.
3. *Analytic,* wherein the words share some association, such as *edit* and *publishing.*

An example item can be seen below (Read, 1993, p. 366):

diffuse

circulate	government	holiday	light
optional	scatter	tolerate	vague

In the above item, the correct answers are *circulate, scatter, light,* and *vague.*

After using the format a number of times, Read concluded that due to the fact that all word classes are represented, this type of format can be very difficult to analyze, since different word classes have different categories of associates. Moreover, some words are highly polysemous, while others essentially have only one meaning, limiting the number of paradigmatic associates available (1993). Furthermore, the measure was excessively difficult to write, a fact to which the present researcher can attest (Batty, 2006). Read revised the format in 1998 in the course of investigating its construct validity. In the revised version Read limited the selection of stimuli words to adjectives only, in order to standardize the categories of associates possible for each item. Furthermore, the possible answers are split into two boxes, with possible synonyms on the left and possible collocates—in this case, nouns which can be modified by the stimulus adjective—on the right. The examinee is to choose a total of four words from the two boxes. There may be three of one and one of the other, or two of each. See the example item below (Read, 1998, p. 46):

sudden

beautiful	quick	surprising	thirsty	change	doctor	noise	school

The answers to the above item are *quick, surprising, change,* and *noise.*

Part of the impetus for this major change to the format was to reduce the effect of guessing by making the number of each kind of associate vary from item to item, from one to three. Later work by Qian and Schedl has confirmed this effect of the format change (2004).

In 2002, Qian sought to evaluate the WAT in his investigation of the link between vocabulary knowledge and academic reading performance. In order to demonstrate construct validity, a pre-1995 TOEFL vocabulary subtest was used as a criterion measure, due to its known statistical reliability and construct validity. Furthermore, a TOEFL reading measure and a Nation Vocabulary Levels Test (Nation, 1990, 2001) were used as criteria. The resulting reliability of the DVK measure was found to be 0.88, which is acceptable, although previous use of this test had resulted in reliability coefficients above 0.90 (Qian, 1998, 1999, as cited in Qian, 2002). The results of the measure correlated significantly with the TOEFL vocabulary measure, the TOEFL reading measure, and the Vocabulary Levels Test. Furthermore, the correlation between the depth test and the TOEFL reading test accounted for 59% of the shared variance, although the R2 of the correlation between depth and the TOEFL vocabulary measure was only 0.46 (i.e. accounting for 46% of the shared variance).

In 2004, Qian again investigated the above measure with Schedl of the Educational Testing Service (ETS) as a possible component of the new TOEFL. One of the primary concerns of the researchers was whether the number of test items could be expanded enough to provide the kind of item pool size required by ETS for the TOEFL while still retaining reliability. Qian's version of the WAT had remained static since it had gained reliability, but a new TOEFL measure would have to use new words from the TOEFL word list. Once again, only adjectives were used in the creation of the new measure. The new test performed much like the original with a Cronbach alpha reliability of 0.91

for the entire measure. Once again, correlations between the measure and the TOEFL sections were significant. Although concerns over the difficulty of the innovative test format for the examinees are common, exit interviews in this study indicated that although the examinees may be confused at first, after reading the directions several times, the format becomes fairly intuitive. Once again, strong correlations between the DVK test and the TOEFL vocabulary and reading tests were observed and the DVK test and vocabulary sections were found to be amply predictive of each other and of performance on the reading section (Qian & Schedl, 2004). Overall, Read's WAT has been shown to be a valid and reliable test of DVK.

The Kanda English Proficiency Test

The Kanda English Proficiency Test (KEPT) was established in 1989 and is a large-scale (over 1000 examinees), video-mediated, norm-referenced, medium-stakes test of general English proficiency. It is administered twice annually, once to current students at the end of the academic year (January) and once again to entering first-year students at the beginning of the academic year (March). The results are used by the university for the purposes of separating the students of the English and Intercultural Languages and Cultures departments into four ability streams for their first- and second-year English classes, tracking increases in student proficiency, and evaluating the English education program in the interest of constant improvement. The KEPT is divided into five sections (reading, grammar, listening, writing, oral), and there are five versions of the test. Each version is thematically centered on a single country, with the readings and listenings being related to that country. The tests are administered in rotation so that no student can take the same version twice in her college career. Research on the test has been published in *Language Testing* (Bonk & Ockey, 2003) and has been found to be highly predictive (74%) of the TOEFL (Bonk, 2001).

Of central importance to the present study are the written and oral sections of the KEPT. These sections do not follow the theme of the rest of the test. The written section consists of a traditional academic (i.e. *five-paragraph-style*) essay written in thirty minutes. It is double-rated for paragraph structure, essay structure, grammar, vocabulary, and content. The oral section is comprised of a group oral discussion task, wherein four examinees discuss amongst themselves a topic presented to them while two raters observe. The examinees are rated on pronunciation, fluency, grammar, vocabulary, and communicative strategies. Raters are normed in two sessions in the two days prior to KEPT administration. Ninety-five percent of raters' vocabulary scores on the written section are within a point or less of each other; in the case of the oral, the agreement within one point is 98%. All scores are subjected to Rasch analysis.

Research question and hypotheses

Despite the widespread use of the kinds of vocabulary assessment which appears on the KEPT, the relationship between these scores and vocabulary knowledge depth as measured by a WAT has heretofore been unexplored. The following research question is therefore posed:

How well do the vocabulary categories of the written and oral sections of the KEPT address the construct of vocabulary knowledge depth?

The following hypotheses are stated:

1. *Scores on the vocabulary category of the written section of the KEPT will be significantly predictive of scores on the DVKT.*
2. *Scores on the vocabulary category of the oral section of the KEPT will be significantly predictive of scores on the DVKT.*

The present study

Setting

The setting for the present study was Kanda University of International Studies (KUIS), a mid-size, four-year foreign language university in Chiba, Japan, offering majors in English Language, International Communication (IC), and International Languages and Cultures. The vast majority of the English instruction occurs at the English Language Institute (ELI), where the students are instructed by over fifty native-English speakers from around the English-speaking world. Communicative language teaching and learner autonomy are stressed throughout the program.

Participants

The participants for this study were 198 second-year English and IC students of the above Japanese university (38 male, 160 female), aged 19 to 23 (mean = 19.9, SD = 0.61). They had studied English for a mean of 8.74 years (SD = 2.18), typically beginning in the first year of middle school. They were less than a month from completion of their second year of university at the time of testing.

Instruments

Depth of Vocabulary Knowledge Test (DVKT)

The test of vocabulary knowledge depth used in the present study was developed by Qian for his 2002 article in *Language Learning*. In Qian's study, a reliability coefficient of 0.88 ($N = 217$) was observed and it correlated significantly with the results of a Nation Vocabulary Levels Test (Nation, 1990, 2001), an accepted and acceptable measure of vocabulary size (Read, 2000). The instrument, therefore, can be assumed to be both reliable and valid. The test was comprised of forty items with four correct answers apiece for a total of 160 points. Guessing was not penalized. The stimulus words were selected by Qian and are described as "general academic adjectives" (2002, p. 525).

The KEPT

Written and oral vocabulary KEPT scores were collected from the January 2006 administration of the KEPT.

Method

The participants were presented with the DVKT during normal class sessions as an optional assessment of their vocabulary knowledge. No students approached for participation in this study opted out. An explanation and the instructions were provided in both English and Japanese.

The administrator also explained the test format in English and led the class in answering two sample items. The classes were then allowed twenty minutes to complete the test. Due to the novel test design, which is difficult to convey in explanation, and the extremely low-stakes condition of the test, any error introduced by students communicating between class administrations of the DVKT is assumed to be insignificant.

Results

Descriptive and reliability statistics on DVKT

Descriptive and reliability statistics can be seen in Table 1. The mean score on the instrument was 105 out of a possible 160 (66%), and with a standard deviation of 15.5, indicating that the distribution of scores was quite uniform. The scores ranged from 56 (35%) to 133 (83%), indicating the difficulty of the measure. Finally, the reliability coefficient (Cronbach's alpha) obtained was 0.89.

Table 1: Mean, standard deviation, range, and reliability of the DVKT

Variable	Max. Poss. Score	Mean (%)	SD	Range (%)	Reliability (α)
DVKT	160	105 (66)	15.5	56 (35) – 133 (83)	0.89

Rasch analysis of the DVKT data

Rasch analysis was performed on the DVKT data to eliminate poorly-performing items and to assign an ability score to each participant. The Rasch model is a model of item response in assessments, and is used to determine an examinee's *ability,* which is understood to be a value irrespective of score on any particular measure. In the Rasch model, an examinee's probability of answering any particular item is modeled as the difference between his ability, as determined by the other items, and the difficulty of the item in question, as determined by the other examinees' performance on it. Items which do not fit this ideal model of response and whose deviation from this ideal model of response is statistically significant are understood to be too unreliable to be allowed to contribute to the determination of the examinees' ability scores and are therefore stricken and the analysis is performed again until the remaining items display an acceptable *fit* with the model.

Rasch analysis resulted in the removal of 35 poorly-fitting items from the final measure, with 125 items remaining in the final version of the data. Descriptive and reliability statistics post-Rasch analysis can be seen in Table 2. The mean ability score was 1.07 with a standard deviation of 0.58. The ability scores ranged from -0.65 to 2.56. Reliability was determined via the person separation index, which can be interpreted similarly to Cronbach's alpha, but is based on the linear transformation of the data resulting from Rasch analysis as opposed to the raw scores of the test itself. In this case, the reliability dropped to 0.85, which is still within the range of acceptable reliability for this kind of measure.

Table 2: Ability score mean, standard deviation, range, and reliability of the DVKT after Rasch analysis

Variable	Mean Ability Score	SD	Range	Reliability (PSI)
DVKT	1.07	0.58	-0.65 – 2.56	0.85

Addressing the research question

Multiple regression was employed to investigate the KEPT's ability to address the DVK construct as measured by the DVKT (Table 3). This revealed a significant, albeit weak, relationship between the KEPT writing vocabulary score and the DVKT, and a non-significant relationship between the KEPT speaking vocabulary score. These results confirm the first hypothesis (that KEPT writing vocabulary scores will be predictive of DVKT scores) and reject the second (that KEPT speaking vocabulary scores will be predictive of DVKT scores). The implications of these findings will be discussed in the next section.

Table 3: Multiple regression of KEPT vocabulary scores and DVKT

	Unstandardized Coefficients		Standardized Coefficients	*t*	Sig.
	B	Std. Error	Beta		
(Constant)	.112	.276		.405	.686
KEPT Writing Vocabulary Score	.270	.086	.222	3.133	.002
KEPT Speaking Vocabulary Score	.121	.086	.100	1.412	.159

Dependent Variable: DVKT

Discussion

The fact that the KEPT written vocabulary score was predictive of the DVK score, while the oral vocabulary score was not, is regrettable but not surprising. Oral raters of the KEPT have long complained of the difficulty of rating so many categories *real time* for four examinees simultaneously, as opposed to rating the same number of categories on a single essay with unlimited rating time. Post-administration analyses have also long indicated that the grammar and vocabulary categories of the oral section are quite confounded. Current research within the KEPT committee is examining this issue and the categories are likely to be revised in the next administration of the test.

In addition to the difference in rating methods, the different task types may require or invite better or more complex samples of vocabulary knowledge. Both the written and oral prompts are intended to allow examinees of all abilities to respond appropriately; however, this necessarily means that the topics must be immediately approachable. This is especially true of the oral prompts, which are often based on the students' lifestyles (e.g. "Could you live without a mobile phone?"). Such questions may not prompt much diversity in lexical usage between examinees of differing abilities. Furthermore, since the format is a group discussion, vocabulary is almost necessarily conversational as opposed to academic, as to produce higher-level vocabulary in such a situation would be unnatural (and even un-nativelike) and may lead to misunderstanding among the other examinees taking part in the discussion.

Moreover, as the DVKT is a test of vocabulary knowledge *depth,* the kind of semantic networking it seeks to probe

is largely outside of the single-word boundary, which may be measured by the raters in another category (likely "grammar"). The KEPT rating rubric for vocabulary makes no mention of collocation or other types of word knowledge, which may explain some of the dissimilarity, but it is important to note that the DVKT has been found to be valid when compared to a Nation Levels Test, so even if the KEPT oral section's vocabulary category was only concerned with vocabulary size, we should expect to see a significant relationship between it and the DVKT. Given all these factors, it seems safe to say that the KEPT oral vocabulary category is adequately addressing neither the construct of vocabulary knowledge depth nor, in all likelihood, vocabulary size.

Limitations of the study

Although the DVKT was administered to 299 students in all, due to the fact that the KEPT is not required of second-year students of the International Communication department, only 198 of those administered the DVKT then completed the KEPT. A larger sample size may reveal different results. Furthermore, it should be remembered that although the WAT has been found to be both reliable and valid, this validity has been demonstrated against tests of vocabulary breadth as opposed to depth. It is possible that the WAT as a measure of vocabulary knowledge depth may need to be re-thought, as Read himself has recently opined (2004). Ultimately, although the findings of the present study are fairly clear, further research is necessary.

Acknowledgements

The present author would like to thank the Research Institute of Language Studies and Kanda University of International Studies for their generous financial support of this project and extend his very warm and heartfelt thanks to Dr. David Qian of Hong Kong Polytechnic University for providing his word associates test to be used in this study.

Aaron Olaf Batty is currently a lecturer at Kanda University of International studies. His research focuses on language assessment and vocabulary. <abatty@kanda.kuis.ac.jp>.

References

Batty, A. O. (2006). The depth of vocabulary knowledge and vocabulary learning strategies of Japanese EFL junior college students. *The Journal of Kanda University of International Studies (*神田外語大学紀要*), 18,* 261–284.

Bonk, W. J. (2001). Predicting paper-and-pencil TOEFL scores from KEPT data. *Studies in linguistics and language education of the Research Institute of Language Studies and Language Education, Kanda University of International Studies (*神田外語大学言語教育研究所言語教育研究*), 12,* 65 – 85.

Bonk, W. J. and Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89–110.

Hatch, E. and Brown, C. (1995). *Vocabulary, semantics, and language education.* New York: Cambridge University Press.

Hunston, S., Francis, G., and Manning, E. (1997). Grammar and vocabulary: Showing the connections. *ELT Journal, 51*(3), 208–216.

Miller, G. A., and Fellbaum, C. (1991). Semantic networks of English. *Cognition, 41,* 197–229.

Nation, I. S. P. (1990). *Teaching & learning vocabulary.* New York: Newbury House Publishers.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Paribakht, T. S., and Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady and T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (174-191). New York: Cambridge University Press.

Pawley, A., and Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards and R. W. Schmidt (Eds.), *Language and communication.* (pp. 191–226). New York: Longman.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning 52*(3), 512–536.

Qian, D. D., and Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing, 21*(1), 28–52.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*(3), 355–371.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. J. Kunnan (Ed.), *Validation in language assessment.* (pp. 41–60). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Read, J. (2004) Plumbing the depths: how should the construct of vocabulary knowledge be defined? In P. Bogaards and B. Laufer (Eds.), *Vocabulary in a second language.* (pp. 209–227). Philadelphia: John Benjamins Publishing.