Validity of the *Vocabulary through Reading Test*

Daniel Isbell

Northern Arizona University

Abstract

Second-language (L2) vocabulary is commonly tested as an underlying trait, typically conceptualized as a size of one's total lexical knowledge and commonly operationalized by selecting the correct meaning for a given word. However, there are concerns about how learners might cope with the sizable vocabulary demands of academic study that lie outside of the realm of L2 instruction. Furthermore, there have been calls in the field of L2 vocabulary testing for higher-context vocabulary tests that shift focus toward the interaction between examinee and words. This provided the impetus for the creation of a test that measures the ability of an examinee to identify definitions for new terms in a text. The present paper examines the fitness of such a test by exploring its reliability, capacity to discriminate test-takers of varying ability, correlation with general reading comprehension, and effect of test-taker background on test results. Evidence for use was found in terms of reliability, ability to discriminate test-takers, and correlation with the reading test. Evidence of test bias related to background was found, but was somewhat mitigated by characteristics of the sample. Implications are relevant for L2 vocabulary researchers, language testers, and test stakeholders.

Keywords: L2 vocabulary, vocabulary testing, definition structures, vocabulary acquisition, academic text, English for academic purposes

Background

In university study, second-language (L2) users are required to read a text and be able to understand definitions for new key terminology. Few tests, if any, exist that specifically target this crucial ability. The present paper describes the development and use of a test which asks L2 learners to read an introductory-level university textbook excerpt and identify definitions for contextually-supported technical terms (Vocabulary through Reading Test, or VRT).

Previous vocabulary research and tests have been mainly concerned with the size and depth of learner vocabulary (Read, 2007; Schmitt, 2008). But what of words that learners do not yet know? It has been suggested that low-frequency yet important technical terms in disciplinary texts are out of the realm of L2 vocabulary instruction (Read, 2000). It would follow that learners need to autonomously learn new terms (Zimmerman, 2009) in university study, just as their native speaking peers must do. The VRT, in turn, targets one avenue for learning new terms independently: identifying definitions while reading a text. Furthermore, the ability to connect definitions and terms in a text requires not only lexical knowledge but also discourse and syntactic knowledge to recognize that a definition is being signaled and where the definition begins and ends, making this ability of interest to L2 reading as well.

Research Questions

To provide validity evidence for the VRT, the proposed study will address the following research questions:

1. Does the VRT effectively distinguish test-takers across the range of ability to recognize definitions in texts? (RQ1)
2. Is the VRT a reliable measure? (RQ2)

3. What is the relationship between the ability to recognize definitions in texts and general reading comprehension ability? (RQ3)
4. Is the ability to recognize definitions in texts affected by learner background? (RQ4)

Methods

Participants

Participants were a group of 158 test takers at the (PIE) in August 2013. These test takers represent both sexes (male = 87, female = 71) and include L1 speakers of mainly Arabic (n = 37), Chinese (n = 80), and Brazilian Portuguese (n = 24). The sample's test forms are archived by the PIE.

Measures

Vocabulary through reading test. The VRT is composed of one reading passage and 10 dichotomously scored items. The reading passage was chosen from a textbook used in one of NAU's 100-level biology courses, and was edited for length in order to have ten key terms (replaced with non-words) with supporting definition structures. Test-takers responded by producing short definitions. Responses were scored correct if they captured the "core meaning" (gist) of the word in the context of the passage.

Reading subtest. The PIE's reading subtest is used to make inferences about a test-taker's reading comprehension ability. The PIE's Fall 2013 placement battery reading subtest was a dichotomously scored multiple choice (4 option) format test. The test was machine scored, with 0 to 40 as the range of possible scores, based on an answer key.

Procedures

Several variables were examined in this study. First, the ability to recognize definitions in texts was considered and operationalized as a score on the VRT. Reading comprehension was operationalized by a score on the PIE placement battery's reading subtest. Additionally, background is a variable operationalized as the self-reported sex and first language (L1).

RQ1 can be phrased operationally as "Do scores on the VRT effectively make distinctions among test-takers?" The distribution of scores within the group was examined.

RQ2 was considered from two perspectives: internal consistency and inter-rater reliability. In operational terms, RQ2.1 is "Are scores on the VRT internally consistent?" Scores were analyzed by computing Cronbach's alpha. RQ2.2 is expressed operationally as "Are scores on the VRT consistent between raters?" The consistency of VRT scores was analyzed by computing the inter-rater reliability (via Cohen's Kappa) of the scores assigned by two different raters for 25 (15.8% of total) test-takers (i.e., a between-groups design).

RQ3 is expressed operationally as "What is the relationship between VRT scores and reading subtest scores?" It was hypothesized that there is a relationship. The two variables were examined in a repeated-measures design via Pearson correlation.

RQ4 is considered in two facets: sex and L1. RQ4.1 is expressed in operational terms as "Are VRT scores affected by sex?" RQ4.2 is expressed operationally as "Are VRT scores affected by L1?" with the null hypothesis that there is no effect of L1 on VRT scores. For sex, VRT scores were analyzed with a Case II independent t-test. VRT scores of the three largest L1 groups were compared via one-way ANOVA.

Results

To answer RQ1, item statistics as well as descriptive statistics were computed for the VRT. Item statistics, including difficulty and discrimination are grouped by subconstruct in Table 1. Difficulty ranged from .34 to .65 with a mean of .51. Discrimination values ranged from .28 to .63 with a mean of .46.

Table 1

VRT Item Statistics

Subconstruct	Item Number	P	D
Explicit Definitions			
	78	0.65	0.45
	79	0.53	0.45
	81	0.61	0.51
	85	0.47	0.46
	87	0.46	0.45
Extended Definitions			
	80	0.64	0.46
	82	0.35	0.28
	83	0.34	0.40
	84	0.69	0.63
	86	0.37	0.53
Mean		.51	.46

Note. P = item difficulty; D = item discrimination (point biserial).

Descriptive statistics for the complete VRT are displayed in Table 2. For the total test, the complete range of scores was used and the mean (5.12) was near the median (5.00). The SEM for the whole test is 1.33.

Table 2

VRT Descriptive Statistics

	N	K	Min	Max	Mean	SD	Reliability	SEM
Total	158	10	0	10	5.12	2.843	.79	1.33
Explicit	158	5	0	5	2.73	1.619	.67	0.93
Extended	158	5	0	5	2.39	1.534	.65	0.91

RQ2 was broken down into two subquestions. RQ2.1 asked whether scores on the VRT are internally consistent. Cronbach’s alpha for the VRT was .79. RQ2.2 asked whether scores on the VRT are consistent between raters. Cohen’s kappa was used to investigate this question;

average inter-rater agreement for the VRT was .79. Agreement values for each item are found in Table 3; kappa values ranged from .54 to 1.

Table 3

Inter-rater Reliability

Item	78	79	80	81	82	83	84	85	86	87	Mean
kappa	1	.54	.60	.68	.72	.65	.92	.82	1	1	.79

Note. Kappa = Cohen's kappa.

RQ3 asked if there is a relationship between VRT scores and Reading Subtest scores. Pearson product-moment correlation between the two measures yielded a statistically significant result of .52 (N = 158, $r_{critical}$ = 0.19, p < .05). The strength of association (r^2) of this relationship is .28.

RQ4 was broken down into two subquestions. RQ4.1 asked if VRT scores are affected by sex. Results of the t-test are presented in Table 4. The difference between groups was found to be significant, allowing a rejection of the null hypothesis. The strength of association (eta^2) for this difference is .04.

Table 4

Comparison of VRT Score by Sex

				95% CI		
Group	n	m	SD	Lower	Upper	t
Male	95	4.65	2.94	4.05	5.25	-2.66*$^+$
Female	63	5.83	2.56	5.18	6.47	

Note. $T_{critical}$ (156 df, 2 tailed) = 2.59; *p < .05; $^+$equal variances not assumed; eta^2 = . 04.

RQ4.2 asked if VRT scores are affected by L1. The three major language groups were considered (Arabic, Chinese, and Portuguese). The ANOVA returned a statistically significant F value (see Table 5), making post hoc comparisons appropriate, shown in Table 6. All groups were found to be significantly different; namely that Portuguese speakers had higher scores and Arabic speakers had the lowest scores.

Table 5

Comparison of VRT Score by L1

Source	df	F	eta^2
Between	2	13.26*	.16
Within	138		

Note. $F_{critical}$ = 3.09; *p < .05.

Table 6

		Mean	95% CI	
Group1	Group 2	Difference	Lower	Upper
Arabic	Chinese	-2.19*	-3.58	-0.78
	Portuguese	-3.37*	-4.85	-1.89
Chinese	Portuguese	-1.19*	-2.37	-0.01

Note. *p < .05; Tamhane T2 adjustment.

Because of these results, it was suspected that another factor may be at play: overall language proficiency. This was operationalized as placement into PIE level 5/6 or into a lower PIE level. Chi-square tests revealed significant differences in the ability of the sexes (df = 1, $X^2_{observed}$ = 8.28 [Yates' correction], p < .05, phi = .24) and the L1 groups (df = 2, $X^2_{observed}$ = 9.25, p < .05, phi = .26). Females trended toward high proficiency while males trended towards low proficiency, and Chinese speakers trended towards high proficiency while Arabic speakers trended towards low proficiency. Comparing the low (n = 71) and high (n = 87) proficiency groups, an independent samples t-test (two tails, df = 156, $t_{critical}$ = 1.98, equal variances assumed) revealed a significant difference ($t_{observed}$ = 6.32, p < .05, eta^2 = .20), as suspected.

To summarize, the VRT was found to effectively distinguish test-taker ability, to be reliable, and to correlate with reading comprehension but not so much as to be redundant. Some evidence for potential test bias was found when examining group means, but overall proficiency characteristics of groups offers explanation for at least a considerable part of any differences.

Relevance to the PIE and Second Language Learning

The results of this research have implications for testing and teaching at PIE. For testing, this study presents a reliable test that distinguishes test-taker ability to identify definitions in an academic text. Information from this test may be useful in making placement decisions at the PIE, as recognizing and then learning definitions for new, unknown terms is useful in real-world academic contexts. In turn, the demonstrated differences among test-takers provide cause for teaching definition structures and in-text definition recognition strategies to help learners become more effective readers of academic texts and prepare them for the vocabulary demands of university study.

References

Grabe, W., & Stoller, F. L. (2011). *Teaching and researching reading* (2nd edition). Harlow, UK: Longman.

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

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies, 7* (2), 105-125.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research, 12*, 329-363.

Zimmerman, C. B. (2009). *Word knowledge: A vocabulary teacher's handbook*. New York, NY: Oxford University Press.