Vocabulary through Reading Test:

Multiple-Choice Version Development and Evaluation

Daniel Isbell

Northern Arizona University

Abstract

When students enter an American university, they typically take a variety of introductory courses. In these courses, information is commonly (if not primarily) provided to students through textbooks, where they begin learning foundational concepts and technical vocabulary of a field. This technical vocabulary accounts for a considerable portion of words in textbooks (Chung & Nation, 2003), much of it considered beyond the realm of L2 instruction (Read, 2000).

To this end, a short-answer test called the Vocabulary through Reading Test was developed, reported on in Isbell (2014). To explore the issue of item format, a multiple-choice version of the test was developed using responses from the short-answer version to create keys and distractors for each item. 147 different examinees took the multiple-choice version (M = 5.48, SD = 2.23). Internal consistency was somewhat lower (Cronbach's alpha = .61). Mean scores for the two forms were not significantly different, which was also true across L1 subgroups, suggesting that the two versions of the test provided similar information. Items in both formats generally performed similarly, but the multiple-choice version had lower item discrimination. Further analysis revealed some difference in scores for one of the two subconstructs, which may suggest a small degree of construct irrelevant variance due to format. These results should be of interest to higher education stakeholders and of relevance to language testers interested in item formats and the integration of reading and vocabulary measurement.

Keywords: English for academic purposes, item format, multiple choice, vocabulary, placement testing

Background

This study reports on an innovative measure of prospective ESL university students' abilities to develop knowledge of new terms presented in a text. When students enter an American university, they typically take a variety of introductory courses. In these courses, information is commonly (if not primarily) provided to students through textbooks, where they begin learning foundational concepts and technical vocabulary of a field. This technical vocabulary accounts for a considerable portion of words in textbooks (Chung & Nation, 2003), and much of it considered beyond the realm of L2 instruction (Read, 2000). However, introductory textbooks are notable for providing in-text definitions (Carkin, 2001). Assessing whether or not L2 students have sufficient ability to connect in-text definitions to technical vocabulary is thus potentially valuable information. Existing direct measures of L2 vocabulary primarily focus on existing vocabulary knowledge (Read, 2007), and common indirect measures involve the robustness of vocabulary in spoken/written production. Neither of these methods would appear to account for the learning of unknown technical vocabulary through reading. The present project aims to: 1) describe the design a test which measures the ability to correctly identify in-text definitions for technical vocabulary (Vocabulary through Reading Test, or VRT), and 2) compare two item formats for delivery of the test.

Research Questions

The present study sought answers to the following questions:

- RQ1: How well does a multiple-choice version of the VRT make distinctions among test takers of varying ability?
- RQ2: Do short-answer and multiple-choice formats of the VRT have comparable reliability?

- RQ3: How similarly do short-answer and multiple-choice format items perform?
- RQ4: How similarly do L1 subgroups perform on both versions of the VRT?

Methods

Participants

Participants for this study were 147 test takers who came to Northern Arizona University as international students for the Fall 2014 semester. This sample included 44 L1 Arabic speakers, 53 L1 Chinese speakers, and 45 L1 Brazilian Portuguese speakers (and 5 test takers who spoke Korean, Japanese, or Spanish). This sample is believed to be a fairly representative sample of PIE students, and more broadly of non-native English speaking students who come to study at universities in the US.

Instruments

A multiple-choice version of the VRT (VRT MC) was developed by the researcher as part of PIE's Placement Test. The multiple-choice VRT has one input passage and 10 items which ask students to identify the best definitions for nonwords inserted into the passage. Keys and distractors for each item were developed based on short-answer responses from a previous version of the test. This instrument, along with the other parts of the PIE Placement Test, was administered to 147 test-takers in August 2014. Data from the short-answer VRT (VRT SA, N = 158, K=10, reliability = .79) was collected in a previous project carried out at the PIE in Fall 2013 (Isbell, 2014).

Procedures

Analyses. The following analyses were performed to investigate each RQ (alpha set at $p < .05$ for all inferential statistics):

- RQ1: Descriptive statistics (mean, SD) and frequencies were calculated to evaluate how well the MC VRT distributes test-takers across the range of possible scores.

- RQ2: Crohnbach's alpha was computed to determine the internal consistency of the VRT MC and this will be compared to the previously computed alpha of the VRT SA (.79).
- RQ3: Item facility and discrimination values for each MC item were computed. These will be compared with values from the SA items. Mean IF and D values were also computed and compared. Additionally, IF and D ranks for items on each form were compared.
- RQ4: A one-way ANOVA was used to compare the means of L1 subgroups on the VRT MC. This result was compared to the same test carried out on the VRT SA.

Administration. Test takers were assigned to different rooms for administration of the Fall 2014 PIE Placement Test (the PIE Placement consists of four subtests corresponding to language skills). The VRT MC was always given after the Reading Subtest of the PIE Placement, but the exact time and order varied. These conditions closely mirror those of the VRT SA. For the MC VRT, students marked their answers on a Scantron sheet for automated scoring. Scantron sheets were collected and processed by the PIE Assessment Team.

Results

Descriptive statistics of the VRT MC and VRT SA (seen below in Table 1) address RQ1 and RQ2. The MC VRT, with a mean score of 5.48 and standard deviation of 2.23, allowed for separation of test takers across a broad range of abilities.

Table 1

VRT MC and VRT SA Descriptive Statistics

Form	N	K	Mean	SD	Reliability	SEM
MC						
Total	147	10	5.48	2.23	.61	1.40
Explicit	147	5	3.13	1.38	.50	0.98
Extended	147	5	2.34	1.20	.30	1.00
SA						
Total	158	10	5.12	2.84	.79	1.33
Explicit	158	5	2.73	1.62	.67	0.93
Extended	158	5	2.39	1.53	.65	0.91

Note. Reliability = Cronbach's alpha.

A histogram showing the distribution of VRT MC scores (Figure 1 below) shows that most test takers fell near the middle of the range of possible scores, with very few test-takers scoring 0 or 10. In comparison with scores from the VRT SA (mean score = 5.12), a t-test (two-tailed, df = 303, $t_{critical}$ = 1.98, $t_{observed}$ = -1.21, p = .227) revealed no statistical difference. Examining subconstruct scores via t-tests revealed a statistical difference in Explicit Definition scores (Bonferroni-adjusted alpha = .025, p = .020), but not in Extended Definitions (p = .773). In terms of reliability, the VRT MC had a reliability (Cronbach's alpha) of .61, which is considerably lower than that of the VRT SA (alpha = .79). Removing Item 5 would result in a Cronbach's alpha of .65; removal of other items would be largely ineffectual.

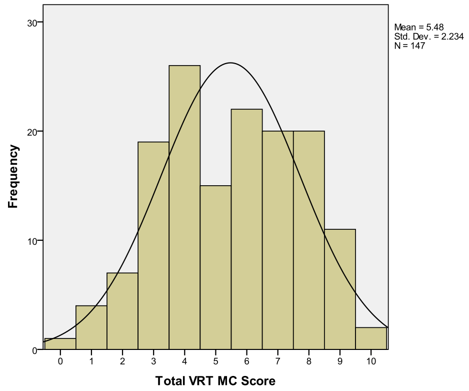

Figure 1. Distribution of VRT MC scores.

Addressing RQ3, Table 2 compares the item statistics of the VRT MC and VRT SA. In general, the MC form was slightly easier than the SA form, and this was the case in 7 out of 10 items. Item difficulty ranks showed small (2 or fewer) differences. In terms of discrimination, the SA form overall had greater discrimination (mean = .46), and this was also true for 9 out of 10 items. Discrimination ranks did not change much when considering the number of ties in the VRT SA discrimination rankings.

Table 2

Comparison of VRT MC and SA Item Statistics

Test Form		MC		SA		Difference (MC-SA)	
Subconstruct	Item	P (Rank)	D (Rank)	P (Rank)	D (Rank)	P	D
Explicit Definitions							
	1	0.79(1)	0.26(7)	0.65(2)	0.45(5)	0.14	-0.19
	2	0.63(3)	0.39(3)	0.53(5)	0.45(5)	0.10	-0.06
	4	0.62(4)	0.35(5)	0.61(4)	0.51(3)	0.01	-0.16
	8	0.54(7)	0.15(9)	0.47(6)	0.46(4)	0.07	-0.31
	10	0.55(6)	0.49(1)	0.46(7)	0.45(5)	0.09	0.04
Extended Definitions							
	3	0.59(5)	0.17(8)	0.64(3)	0.46(4)	-0.05	-0.29
	5	0.29(9)	-0.03(10)	0.35(8)	0.28(7)	-0.06	-0.31
	6	0.45(8)	0.36(4)	0.34(10)	0.40(6)	0.11	-0.04
	7	0.74(2)	0.42(2)	0.69(1)	0.63(1)	0.05	-0.21
	9	0.28(10)	0.32(6)	0.37(9)	0.53(2)	-0.09	-0.21
Mean		0.55	0.29	0.51	0.46	0.04	-0.17

Note. P = item difficulty, Rank = easiest (1) to most difficult (10); D = item discrimination, Rank = most discrimination (1) to least (10).

Last, when examining L1 subgroup performance on the VRT MC, Table 3 shows subgroup means across the two forms of the test. Generally, groups ranked similarly, though Arabic speakers performed noticeably better on the MC version while Chinese speakers performed worse. However, based on the means and 95% confidence intervals, none of the differences in score based on test format appear to be statistical.

Table 3

L1 Subgroup Means for the VRT MC and VRT SA

			95% CI				95% CI	
L1 Subgroup	VRT MC n	VRT MC Mean (SD)	Upper	Lower	VRT SA n	VRT SA Mean (SD)	Upper	Lower
Arabic	44	4.48 (1.84)	3.92	5.04	37	3.30 (2.91)	2.33	4.27
Chinese	53	4.89 (2.14)	4.30	5.48	80	5.48 (2.78)	4.86	6.09
Portuguese	45	7.11 (1.80)	6.65	7.72	24	6.67 (1.78)	5.91	7.42

For the VRT MC, a one-way ANOVA revealed a statistical difference (df1 = 2, df2 = 139, $F_{critical}$= 3.09, $F_{observed}$= 24.18, p < .001). Post-hoc comparisons are displayed in Table 4.

Table 4

Multiple Comparisons among L1 Groups on the VRT MC

		Mean		95% CI	
Group1	Group 2	Difference	p	Lower	Upper
Arabic	Chinese	-0.41	.910	-1.37	0.55
	Portuguese	-2.63*	<.001	-3.63	-1.64
Chinese	Portuguese	-2.22*	<.001	-3.18	-1.27

Note. *p < .05, Bonferroni adjustment.

Compared with the same test run for the VRT SA (Table 5), results were similar in that most pairs were statistically different, except in this case no statistical difference was found between L1 Arabic and L1 Chinese speakers.

Table 5

Multiple Comparisons among L1 Groups on the VRT SA

		Mean		95% CI	
Group1	Group 2	Difference	p	Lower	Upper
Arabic	Chinese	-2.19*	.001	-3.58	-0.78
	Portuguese	-3.37*	<.001	-4.85	-1.89
Chinese	Portuguese	-1.19*	.046	-2.37	-0.01

Note. *p < .05, Tamhane’s T2 procedure.

Relevance

This project addressed PIE Research Priority #3 regarding university preparedness. The VRT specifically targets a specialized reading skill for learning technical vocabulary in university textbooks, and thus yields pertinent information for determining an L2 English speaker's preparedness to engage in university studies.

In sum, the VRT MC appears to capture mostly the same information about test-takers as the VRT SA. The means and SDs of the two forms are similar, and total scores are not statistically different. In terms of subgroup performance, each test format found a statistical difference in scores based on L1. For the VRT SA, Isbell (2014) explained this as largely being due to general language proficiencies of the L1 groups. The VRT MC ranked L1 groups in the same order as the VRT SA, and there were no statistical differences within L1 subgroups across test formats. The VRT MC was also attractive for its expedient scoring. However, the reliability of the VRT MC and the statistical difference in one of the subconstruct scores are cause for concern. The MC format only had a reliability of .61 compared to the SA format’s .79. While .61 is respectable considering the short length of the test, a parallel form would need to be twice as long to achieve a reliability of .75 (via Spearman-Brown predictive reliability formula). The lower reliability of the test can be interpreted as a reflection of its items generally having lower discriminating power than the SA items. Both the reliability and difference in subconstruct scores may be due to guessing being a viable strategy in the MC format. Alternatively, certain skills which benefit test takers on the SA version (e.g., close reading/syntax knowledge to parse the beginning and end of in-text definitions, writing coherent phrases/sentences) may be non-factors in the MC version, where test takers may be able to make use of somewhat different abilities (e.g., comparing the content of choices to information in a the text).

References

Carkin, S. (2001). *Pedagogic discourse in introductory classes: Multi-dimensional analysis of textbooks and lectures in biology and macroeconomics* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Publication number: AAI3004014).

Chung, M., & Nation, P. (2003). Technical vocabulary in specialized texts. *Reading in a Foreign Language*, *15*, 103-116.

Isbell, D. (2014). Validity of the Vocabulary through Reading Test. *Research Projects in the PIE 2013-2014*. Retrieved from http://nau.edu/uploadedFiles/Academic/CAL/PIE/_Forms/Validity%20of%20the%20Vocabulary%20through%20Reading%20Test.pdf

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*, 105-125.