Vocabulary Learning and Instruction
Volume 4, Issue 1, October 2015
http://vli-journal.org

Article

Psychometric Properties of Word Association Test with Regard to Adolescent EFL Learners

Hye Won Shin
Teachers College, Columbia University
doi: http://dx.doi.org/10.7820/vli.v04.1.shin

Abstract

This study reports on the psychometric evaluation of a second language vocabulary test. The test measured target words within a given lesson using a sample of 142 adolescent EFL learners. The test was a modified word association test designed to asses students' ability to identify paradigmatic and syntagmatic relations between words. The study was an outgrowth of research examining how different instructional techniques affect deep word knowledge in elementary school-aged EFL learners over time. Results demonstrate that the data fit a two factor dimension and support adequate levels of reliability and validity.

1 Introduction

Second language (L2) teachers and researchers have widely affirmed that vocabulary knowledge is central to developing L2 proficiency. Consequently, in the last decade researchers have looked to identify conceptual frameworks to assess vocabulary knowledge. L2 vocabulary researchers, in particular, have proposed varying frameworks, one of which, the word association test (WAT), assesses "network knowledge" of words.

Although there are various conceptualizations of lexical knowledge, there is still debate among scholars about what exactly "deep word knowledge" entails. The struggle over how to define "depth" is part of the challenge. Read (2000, 2004) attempts to clarify by suggesting network knowledge, or how well learners can relate a word to other words they know, as a measure for assessing depth of word knowledge. Network knowledge is based on the assumption that words people learn do not exist as isolated elements. Richer associations among related words, a more connected, and better organized lexicon – these things are part of higher vocabulary knowledge.

The WAT format, in which learners must identify words that are semantically associated with a given target word, is well suited to assessing this facet of deep word knowledge. Word association tasks generally utilize a multiple-choice response paradigm and have the potential to not only measure word *meanings* but also some of their *uses* as well (see, for example, Greidanus & Nienhuis, 2001; Qian, 1999; Schoonen & Verhallen, 2008).

The WAT format has been shown to yield reliable and valid information about depth of L2 vocabulary (Greidanus & Nienhuis, 2001; Read, 1993; Schoonen &

Please cite this article as: Shin, H.W. (2015). Psychometric properties of word association test with regard to adolescent EFL learners. *Vocabulary Learning and Instruction*, *4* (1), 9–15. doi: 10.7820/vli.v04.1.shin

edit			
arithmetic	film*	pole	publishing*
revise*	risk	surface	text*

Figure 1. Word association multiple-choice format (Read, 1993).

Verhallen, 2008). Read (1993), for example, designed his word association task to evaluate college students' knowledge of academic English words. It consisted of a target word followed by eight other words, some of which showed paradigmatic, syntagmatic, and/or analytic relationships to the target word (see Figure 1). The L2 learners had to identify the four words which were in fact related to the target word. The high correlation found between two forms of the test (i.e., Form A, Form B) indicates that there is almost a perfect relationship for the forms ($r = 0.97$). Additionally, forms A and B showed high correlations with the vocabulary section of the English Language Institute Proficiency Test ($r = 0.76$ and $r = 0.81$, respectively), which can be interpreted as evidence for the concurrent validity of Read's WAT.

Research on how to measure the deep lexical knowledge of students at an early stage of their L2 learning has been done by Schoonen and Verhallen (2008). They designed their WAT for children aged 9–12 (Grades 3–6) learning Dutch as an L2. It was in fact a simplified version of Read's (1993) test with each target item having three correct answers out of six options. Using this six-option version, the two forms (Form A and Form B) of their test appeared reliable: item-total correlations ranged from 0.75 to 0.83, and Item Response Theory (IRT) reliability was as high as 0.92. Moreover, the test appeared valid on the basis of concurrent validity with a definition test for Grade 3 ($r = 0.82$) and Grade 5 ($r = 0.85$) students.

In short, WATs appear to be a well-developed measure for assessing depth of word knowledge. This study extends prior research on WAT. Here, I investigate the reliability and validity of a modified WAT within elementary school classrooms to ascertain the psychometric properties of WATs on adolescent learners of English as a foreign language.

2 Purpose of This Study

Evidence about the reliability and validity of WATs is accumulating. However, no studies have examined the psychometric properties of a modified WAT for adolescent EFL learners. The purpose of this study, therefore, was to ascertain the internal consistency, construct validity, and factor structure of a WAT designed for adolescent learners of English as a foreign language.

The following research questions guided this study:

(1) How reliable is a modified WAT for measuring word knowledge in adolescent EFL learners?
(2) How valid is a modified WAT for measuring word knowledge in adolescent EFL learners?

3 Method

3.1 Participants

This study included 152 sixth grade students from a single elementary school in Seoul. Ten students, or 6.6% of the total number, did not complete the test and were therefore excluded from analysis. Based on the proportion of students who qualified for the free or reduced-price lunch program, which is indicative of low socioeconomic households, it is fair to call the students at the school relatively affluent. About 1.65% of the sample was eligible for free or reduced-price lunches. The sample was 51% female. All of the students had at least three years of English as a foreign language, as prescribed by the national curriculum of South Korea.

3.2 Procedures

An elementary school was recruited for the study. Data collection was conducted by administering the tests on two separate occasions to all sixth graders in the school. During each testing session of 40 minutes, overseen by the students' English teacher during regular class time, participants completed a norm-referenced test that assessed vocabulary knowledge. In a subsequent session, the participants completed the modified word association measure.

3.3 Measures

3.3.1 Word association test. Word knowledge was assessed with a researcher-developed WAT. In a six-choice multiple-choice response format, this word association task allowed students to demonstrate deep word knowledge of paradigmatic relations (e.g., the fact that the words *edit* and *revise* are similar in meaning) and syntagmatic relations (e.g., the fact that the words *edit* can come right before the word *text*). Of the 21 test items, 15 items targeted paradigmatic relations and 6 items targeted syntagmatic relations. Students were presented with six possibly associated words and instructed to select the word with a paradigmatic or syntagmatic link to the target word. The remaining five options were unrelated distractors.

3.3.2 The Gates-MacGinitie reading test, fourth edition. Global vocabulary skill was assessed using the Gates-MacGinitie Reading Test, Fourth Edition (GMRT-4), a valuable and commonly used reading assessment tool that evaluates the general level of reading achievement of a native speaker in a given grade or grade range. It is an 82-item multiple-choice task, with 43 and 39 test items for word knowledge and reading comprehension, respectively.

3.4 Data Analysis

Internal consistency was established for the total WAT after removal of items using Cronbach's α. Item-level analyses were adopted as a step in the investigation

illustrating whether the revised word association format exhibited psychometrically sound properties for this population of students and for the purposes of this study. Scores on the GMRT-4 were correlated with the WAT's paradigmatic relations, the WAT's syntagmatic relations, and the WAT as a whole, in order to establish convergent validity.

Confirmatory factor analysis (CFA) was used to investigate and identify the factors underlying the word association for students in an FL setting. Specifically, CFA was used to validate the two-dimensional structure of the WAT. As described above, WAT was designed to measure two subscales related to word knowledge in elementary school. Paradigmatic and syntagmatic items were designed to contribute to two factors, and as such, items were hypothesized to load to two factors. CFAs were carried out in Mplus version 7.0 (Muthén & Muthén, 1998–2012) to examine the two factor model.

4 Results

4.1 Item Analysis and Internal Consistency

A summary of item analysis statistics of WAT is compiled in Table 1. The WAT with 21 items had a Cronbach's α of .80. The internal consistency for both paradigmatic and syntagmatic relations exceeds .95. Despite evidence of internal consistency, a set of items was removed for further analysis based on difficulty and item-to-total correlation. A closer look revealed that all items exceeded the positive index values of 0.30 limit, a discrimination index deemed appropriate for a researcher-developed test item (Ebel & Frisbie, 1986; Nunnally, 1978), with the exception of 4 items: 2 (*deliver*), 4 (*lay*), 11 (*subtract*), and 21 (*prescribe*). Therefore, the final item set characteristics of the WAT included 17 items, with 15 paradigmatic relations and 6 syntagmatic relations. The internal consistency computed for the sampled test-takers after the removal of four items was likewise, high, with coefficient α of 0.83.

Table 2 presents the mean, standard deviation, minimum, maximum, kurtosis, and skewness for 17 target items of the WAT. On this test task, the highest possible score was 17.00; the range of scores actually acquired by the participants ranged from 0.00 to the maximum score of 17.00. The mean was 7.46 and the standard deviation was 4.22. In addition, a positive kurtosis of 2.11 and a skewness of 0.28

Table 1. Summary of Item Analyses on WAT (N = 142)

	No. of items	Cronbach's α	No. of items removed	No. of items retained	Cronbach's α (after items removed)
WAT	21	.80	4[a]	17	.83
Paradigmatic relations	6	.95	–	–	–
Syntagmatic relations	15	.96	–	–	–

Note: N = number of participants.
[a]2 (deliver), 4 (lay), 11 (subtract), 21 (prescribe).

Table 2. Descriptive Statistics of WAT, and Paradigmatic and Syntagmatic Relation Scores for Pretest

	WAT (k = 17)	Paradigmatic relations (k = 6)	Syntagmatic relations (k = 11)
N	142	142	142
Mean	7.46	2.73	4.73
SDs	4.22	1.84	2.70
Minimum	0.00	0.00	0.00
Maximum	17.00	6.00	11.00
Kurtosis	2.11	1.91	2.19
Skewness	0.28	0.21	0.27

were observed. The skewness, which is slightly positive, implies that the test was somewhat difficult for the students. This is not surprising, given that the target words were selected from a list of words which had not yet been covered in class. In general, the 142 students' vocabulary knowledge was normally distributed.

4.2 Convergent Validity with GMRT-4

To investigate whether a modified WAT measured what it intended to measure (that is, test validity), correlations between WAT scores and GMRT-4 test scores, a measure of the same or related construct test (Cronbach, 1971; Messick, 1989), were obtained. While the GMRT-4 subtests included vocabulary knowledge and reading comprehension, correlations with the mostly aligned construct, namely, vocabulary knowledge should be found. The convergent validity of the researcher-developed measure and the norm-referenced test was moderately positive for the WAT and the GMRT-4 vocabulary knowledge (r = .66, p = .01); paradigmatic relations and the GMRT-4 vocabulary knowledge (r = .58, p = .01); and syntagmatic relations and the GMRT-4 vocabulary knowledge (r = .66, p = .01). In general, there is initial evidence to establish convergent validity between the WAT and the GMRT-4 subtest scores.

4.3 Confirmatory Factor Analysis

A CFA employing maximum likelihood estimation with robust standard errors and a mean adjusted chi-square statistic test was undertaken to test the two dimensional measurement model. The CFA standardized factor loadings are presented in Figure 2. The model revealed item fit, with the two underlying factors assessing construct of word knowledge: comparative fit index (CFI) = 0.82, and standardized root mean square residual = 0.06, $p < .05$. Although CFI is slightly smaller than the recommended standard of .90 (Hu & Bentler, 1999; In'nami & Koizumi, 2011), Bollen (1989) has suggested that a value around .85 indicates a good fit. Therefore, the model met recommended criteria for an acceptable fit to the data. The diagram also displays the standardized factor coefficients for each item. The factor loadings of the items ranged from .35 to .77. All loadings were statistically significant. The two factors were also significantly correlated, r = 0.83, $p < .001$.

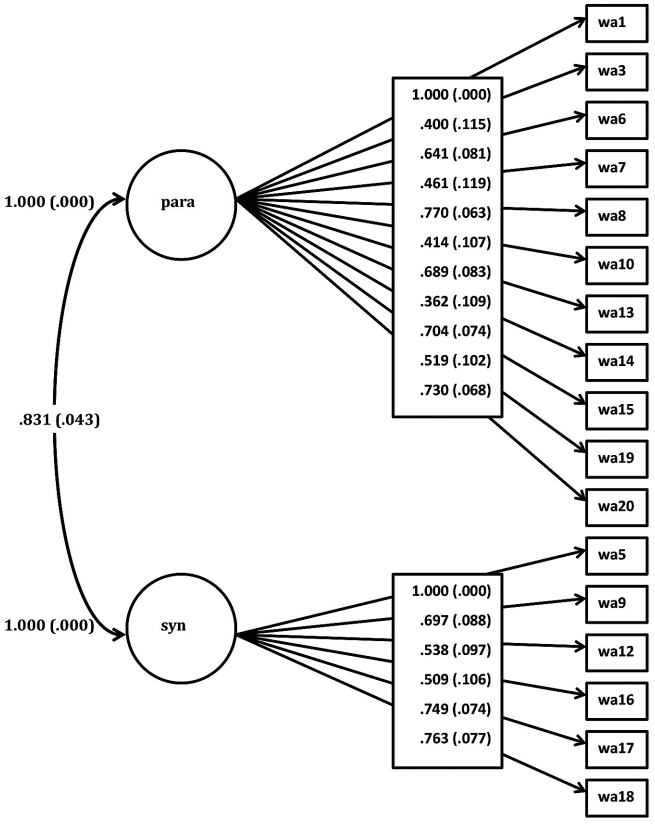

Figure 2. Diagram of two-factor CFA model.
Note: para denotes paradigmatic and syn denotes syntagmatic.

5 Conclusion

The primary focus of this study was to establish evidence of the validity and reliability of WAT in a sample of adolescent EFL learners. Aforementioned pieces of psychometric evidence make it clear that the abilities to identify paradigmatic and syntagmatic relations are tapping into rather different dimensions of deep word knowledge.

Acknowledgement

I am indebted to Dr. Rie Koizumi for her invaluable comments on an earlier version of this paper.

References

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement* (4th ed.). Toronto: Prentice Hall.

Greidanus, T., & Nienhuis, L. (2001). Testing the quality of word knowledge in a second language by means of word associations: Types of distractors and types of associations. *The Modern Language Journal*, *85*, 567–577. doi:10.1111/0026-7902.00126

Hu, L. Z., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi:10.1080/10705519909540118

In'nami, Y., & Koizumi, R. (2011). Structural equation modeling in language testing and learning research: A review. *Language Assessment Quarterly*, *8*, 250–276.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus*. Los Angeles, CA: Author.

Nunnally, J. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*, 282–308. doi:10.3138/cmlr.56.2.282

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*, 355–371. doi:10.1177/026553229301000308

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 & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 209–227). Amsterdam: Benjamins.

Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing*, *25*, 211–236. doi:10.1177/0265532207086782

Online edition : ISSN 2187-2759 Print edition : ISSN 2187-2767

Vocabulary Learning and Instruction

A Journal of Vocabulary Research

Volume 4 Number 1 October, 2015