Vocabulary Learning and Instruction
Volume 1, Issue 1, August 2012
http://vli-journal.org

Identifying Dimensions of Vocabulary Knowledge in the Word Associates Test

Aaron Batty
Keio University
doi: http://dx.doi.org/10.7820/vli.v01.1.batty

Abstract

Depth of vocabulary knowledge (DVK) (i.e. how much a learner knows about the words he knows) is typically conceptualized as a psychologically multidimensional construct, including various forms of word knowledge. Read's Word Associates Test (WAT) is the most common test of DVK in the literature, assessing knowledge of words' synonyms and collocates. Despite the fact that the WAT aims to measure two dimensions of vocabulary knowledge, no studies until now have investigated whether these dimensions are psychometrically distinct. The present study seeks to fill that gap. A known-reliable-and-valid WAT developed by David Qian was administered to 530 Japanese university English majors. Confirmatory factor analysis was employed to investigate the psychometric dimensionality of the WAT. It was discovered that a bifactor model, wherein the primary explanatory factor is a vocabulary g-factor, with additional, uncorrelated factors for synonym and collocate items, demonstrated the best fit. This finding implies that although these dimensions of DVK may be somewhat distinct, they are largely subsumed by general vocabulary knowledge.

Keywords: vocabulary; depth of vocabulary knowledge; word associates test; multidimensionality; structural equation modeling.

1 Background

1.1 Depth of Vocabulary Knowledge (DVK)

Vocabulary can be separated into two broad categories of knowledge: how many words one knows (vocabulary breadth) and how well one knows those words (vocabulary depth) (Nation, 1990; Richards, 1976). Operational definitions of DVK are numerous, but most conceptualize it as a dimensional construct (e.g. Henriksen, 1999; Hunston, Francis, & Manning, 1997; Read, 1993, 1998, 2000; Schoonen & Verhallen, 2008).

Some (e.g. Hunston et al., 1997; Miller & Fellbaum, 1991) claim that DVK lies in the semantic networks which connect a word with the other information necessary to truly understand and use it, a view shared by Henriksen (1999), who argues that in order to truly know a word, the learner must engage in semantic network building, i.e. creating intentional links between the target word and other words the learner knows, including morphological similarity, syntactic similarity and, of particular relevance to the present study, collocational similarity. It is partly this kind of knowledge depth that enables what Pawley and Syder term (1983)

Please cite this article as: Batty, A. (2012). Identifying dimensions of vocabulary knowledge in the word associates test. *Vocabulary Learning and Instruction, 1*(1), 70–77. doi: 10.7820/vli.v01.1.batty

"nativelike selection"– the ability for a speaker of a language to make the same lexical choices as a "native" speaker of that language.

Although many agree that collocational knowledge is critical to DVK, few vocabulary tests attempt to address it. The only widely known test format that does make such an attempt is the Word Associates Test (WAT) (Read, 1993, 1998), which, in addition to testing synonyms, attempts to incorporate collocational knowledge of words of tested words. However, its ability to treat collocational knowledge as a separate psychometric dimension has not been established. The goal of this study is to determine if the two dimensions measured by the test, knowledge of synonyms and knowledge of collocates, can be identified as distinct psychometric constructs.

1.2 The WAT

The test presents the examinee with a stimulus word followed by four possible synonyms and four possible collocates, from which the examinee is to choose four correct associates. The stimulus word is always an adjective and the collocates are always nouns modifiable by the stimulus. The correct answer may include one synonym and three collocates, two synonyms and two collocates, or three synonyms and one collocate. This uncertainty was added in an attempt to limit the effectiveness of guessing strategies (Read, 2000, p. 184). An example is shown in Figure 1. The newer WAT has been employed most visibly by Qian to investigate the link between DVK and reading performance (Qian, 2002), and later to evaluate the format as a possible addition to the TOEFL (Qian & Schedl, 2004).

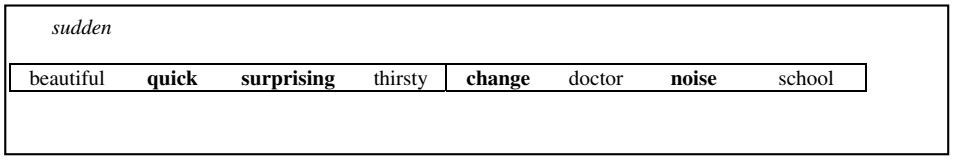

Figure 1. Example of WAT item (Read, 1998, p. 46). The words in the left box are possible synonym options; those on the right are possible collocate options. Correct answers are in boldface.

1.3 Dimensionality of the WAT

Despite the fact that the WAT is intended to measure DVK conceptualized as a dimensional construct, very little research has been conducted on this aspect. Several researchers have scaled WAT data with the Rasch model (Rasch, 1960), either dichotomously (i.e. each correct answer is treated as a separate item) (e.g. Batty, 2006), or with the Rasch partial credit model (Masters, 1982), treating each stimulus word as an item worth four points (e.g. Batty, 2008; Read, 1998). In these cases, the items—whether dichotomous or polytomous—exhibited acceptable fit statistics, indicating that the items adhered closely to the Rasch model's assumption of unidimensionality (Bond & Fox, 2007, p. 35). However, Rasch fit statistics can be deceptive, as they are sample dependent (Bond & Fox, 2007, p. 285). As such, increases in the number of items or observations will always result in improved fit to the Rasch model, which may obscure dimensionality. Factor analysis of an original, 1993 version of the WAT, found that a one-factor model exhibited better fit to the data than a two-factor model, although details are few (Schoonen & Verhallen, 2008). These findings all suggest that the WAT operates as a unidimensional

measure, despite its multidimensional theoretical underpinnings. However, to date, no one has carried out a principled and theoretically sound investigation of the dimensionality of the current WAT format.

There are three likely possibilities for the dimensional structure of the WAT. The first is a unidimensional structure, where a single vocabulary knowledge factor predicts all of the items, as illustrated by the path diagram in Figure 2.

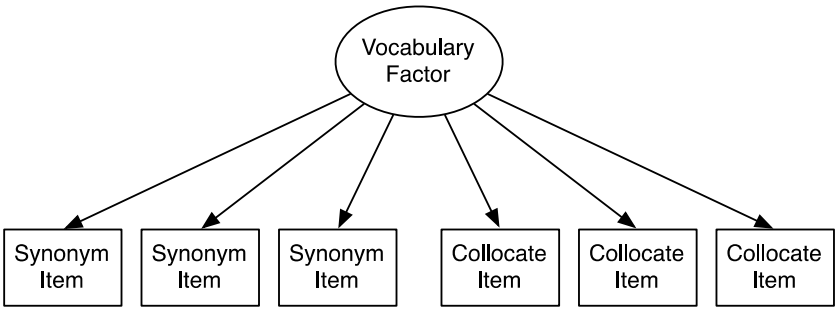

Figure 2. A simplified path diagram of a unidimensional model of the WAT.

Another possibility is a two-factor model, wherein a synonym factor predicts the synonym items, and a collocate factor predicts the collocate items. In such a model, the two factors are correlated, as we can assume that both are types of vocabulary knowledge, and are therefore related (see Figure 3).

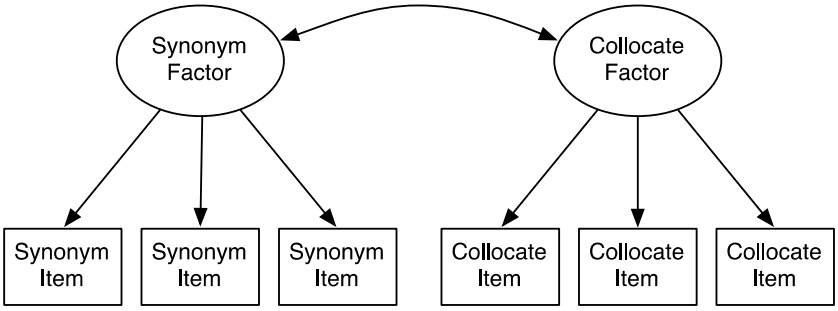

Figure 3. A simplified path diagram of a two-factor model of the WAT.

Another type of model pertinent to the WAT is the bifactor model (Holzinger & Swineford, 1937). In such a model, all the variables are assumed to load on a single general factor (or g-factor), while groups of variables additionally load on separate, smaller factors. In such a model, the subskill factors are distinct from the g-factor, rather than being subsumed by it. In the case of the WAT, such a model is acceptable from a theoretical standpoint, as it includes a single, overarching vocabulary dimension, in addition to distinct synonym and collocate factors which are pertinent only to them (see Figure 4). An advantage of such a model is that, if it displays appropriate model fit, reporting of the test's subscores in addition to the total scores is justified.

1.4 Research Question

The goal of this study is to apply tests of dimensionality to the WAT to address the following research questions: Is the WAT a unidimensional measure of

vocabulary knowledge, or do the subskills load on their own psychometric dimensions? If so, are the item types best modeled as correlated yet distinct factors, or as uncorrelated subskills subordinate to a general factor in a bifactor model?

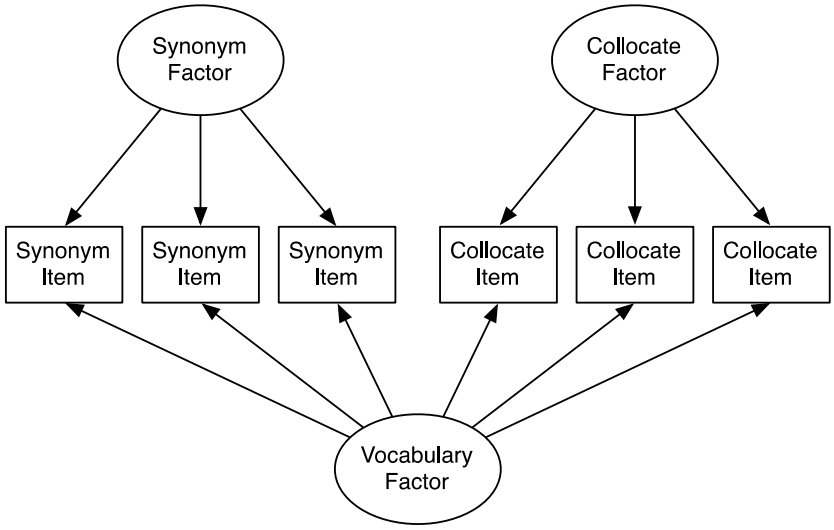

Figure 4. A simplified path diagram of a bifactor model of the WAT.

2 Method

2.1 Instrument

The WAT form used in the present study was developed by Qian for the 2002 study discussed above. In that 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, p. 118). The instrument, therefore, can be assumed to be both reliable and concurrently valid. The 40 stimulus words were selected by Qian and are described as "general academic adjectives" (Qian, 2002, p. 525).

2.2 Sample

The participants were 530 first- and second-year English and international communication majors at a foreign-language university in Japan. The sample included a wide range of English language proficiencies, as evidenced by the university's internal placement exam. The WAT was administered as an optional vocabulary test during normal class sessions, and results were delivered to the class instructors to be distributed to the students.

2.3 Data Analysis and Results

For the purpose of investigating dimensionality, each correct answer on the WAT was treated as a single dichotomous item, resulting in 160 possible items. To facilitate the creation of item parcels (see next section), the dataset was reduced to 145 items. The two synonyms and the three collocate items with the lowest point-biserial statistics were removed from the dataset, resulting in a dataset of

70 synonym and 75 collocate items. This reduced dataset was found to have a Cronbach alpha reliability coefficient of 0.89.

To ascertain whether the WAT conformed better to a one-factor or two-factor model, confirmatory structural equation modeling was employed using the Mplus software package (Muthén, Muthén, Asparouhov, & Nguyen, 2011). Mplus offers the weighted least squares with mean and variance adjusted (WLSMV) estimation method, which is widely recommended for analyses of categorical (e.g. binary) data (e.g. Brown, 2006, p. 388; Jasper, 2010). An initial two-factor analysis revealed a very high correlation between the factors ($r = 0.92$), indicating a high degree of multicollinearity in the model. Following the advice of Jasper (2010), the items were grouped into item parcels for further analysis. The items were ordered from least- to most-difficult as determined by Rasch analysis using Winsteps (Linacre, 2011) and parceled into bundles of five items and the item responses were summed.

The data were first fitted to a unidimensional (i.e. vocabulary g-factor) model. Model fit statistics is shown in Table 1. Overall model fit was less than satisfactory. The significant chi-square statistic indicated poor fit with the data (Brown, 2006, p. 81; Schumacker & Lomax, 2004, p. 82), although the chi-square statistic is often inflated by sample sizes of over 200, and so may not be a good indicator of fit for the present sample (Schumacker & Lomax, 2004, p. 100). The comparative fit and Tucker-Lewis indices (CFI and TLI, respectively) should ideally be approximately 0.95 or greater, and the root mean square error of approximation (RMSEA) should be 0.06 or less (Brown, 2006, p. 87). None of these are the case with the one-factor model. The data were then fitted to a two-factor model, wherein the synonym and collocate items load on correlated synonym and collocate factors, respectively. The fit of this model was very slightly superior to that of the one-factor model, but was still not ideal (see Table 1). Finally, the data were fitted to a bifactor model, as discussed above. This model was found to exhibit the best fit of all (see Table 1). Examining the standardized loadings on the three factors of this model (Table 2), it becomes clear that the highest item loadings overall were on the vocabulary g-factor, half of the synonym items loaded highest on the synonym factor, and only two of the collocate items loaded more highly on the collocate factor than on the g-factor.

Table 1. Results of the SEMs of the Item-Parceled WAT dataset ($N = 530$)

Model	χ^2	CFI	TLI	RMSEA
1 factor: vocabulary g-factor	1446.21*	0.840	0.828	0.073
2 factors: synonym and collocate, correlated	1199.83*	0.877	0.867	0.064
Bifactor model: vocabulary g-factor, synonym, and collocate	909.29*	0.916	0.902	0.055

Notes: All chi-square statistics calculated via the DIFFTEST operation in Mplus 6.11 (Muthén et al., 2011), under WLSMV estimation.
*$p < 0.01$.

3 Discussion and Conclusion

This study sought to determine whether the WAT was best modeled as a unidimensional, two-factor, or bifactor test of vocabulary knowledge. The data were found to best fit to the bifactor model, where the primary explanatory factor is a

single vocabulary g-factor, with additional, uncorrelated second-order subskill factors for synonym and collocate items.

Table 2. Standardized Factor Loadings of the Bifactor Model

Item parcel	Vocabulary	Synonym	Collocate
Syn. 1	0.350*	0.413*	
Syn. 2	0.362*	0.243*	
Syn. 3	0.550*	0.089	
Syn. 4	0.427*	0.346*	
Syn. 5	0.252*	0.570*	
Syn. 6	0.249*	0.469*	
Syn. 7	0.409*	0.459*	
Syn. 8	0.572*	0.238*	
Syn. 9	0.413*	0.314*	
Syn. 10	0.357*	0.360*	
Syn. 11	0.333*	0.496*	
Syn. 12	0.417*	0.218*	
Syn. 13	0.344*	0.498*	
Syn. 14	0.601*	0.155*	
Col. 1	0.409*		0.479*
Col. 2	.502*		0.184*
Col. 3	0.529*		0.056
Col. 4	0.657*		−0.198*
Col. 5	0.448*		0.253*
Col. 6	0.315*		0.488*
Col. 7	0.380*		0.195*
Col. 8	0.541*		0.316*
Col. 9	0.527*		0.015
Col. 10	0.592*		−0.046
Col. 11	0.502*		0.245*
Col. 12	0.529*		0.093
Col. 13	0.605*		0.063
Col. 14	0.446*		0.389*
Col. 15	0.632*		−0.150*

Notes: Estimation method: WLSMV. The variances of the three factors were fixed at 1. $N = 530$.
*$p < 0.01$.

These findings are of interest theoretically, as they suggest that knowledge of synonyms and collocates are distinct subskills of vocabulary knowledge. The raw-score correlation between the item types was low ($r = 0.61$) relative to inter-item correlations within the collocate and synonym items alone (0.79 and 0.80, respectively), indicating that it may be meaningful for tests to provide distinct subscores for both synonym and collocational knowledge. Further development of items and tests to do this may be a fruitful avenue for future research.

In the case of the WAT itself, however, the collocation item types loaded somewhat weakly as a separate factor, calling the ability of this particular test to reliably assess collocational knowledge into question. Despite the items' loadings on

distinct psychometric dimensions, the large vocabulary knowledge g-factor shared by both test sections muddles the interpretability of the two item types somewhat. Moreover, as the WAT is typically modeled as unidimensional (e.g. Batty, 2006, 2008; Nurweni & Read, 1999; Read, 1993, 1998), it raises the question of what is gained by its novel and potentially problematic design (e.g. guessing, examinee confusion; see Read, 1993, 1998; Schmitt, Ng, & Garras, 2011 for further discussion) over more robustly validated measures such as the Vocabulary Levels Test (Beglar & Hunt, 1999; Schmitt, Schmitt, & Clapham, 2001).

Acknowledgements

The author would like to extend special thanks to David Qian for providing a reliable and validated WAT, and Jeffrey Stewart for his valuable suggestions and contributions. Partial funding was provided by a grant from the Research Institute of Language Studies and Language Education at Kanda University of International Studies, Chiba, Japan.

References

Batty, A.O. (2006). An analysis of the relationships between vocabulary learning strategies, a word associates test, and the Kanda English Proficiency Test. *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)*, *17*, 1–22.

Batty, A.O. (2008). *Vocabulary learning strategies vs. depth of vocabulary knowledge: Do strategies have any effect?* Presented at the International Association of Applied Linguistics (AILA) 2008 World Congress, Essen, Germany.

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, *16*(2), 131–162. doi:10.1191/026553299666419728

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). London: Lawrence Earlbaum Associates.

Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303–317. doi:10.1017/S0272263199002089

Holzinger, K.J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*, 41–54. doi:10.1007/BF02287965

Hunston, S., Francis, G., & Manning, E. (1997). Grammar and vocabulary: Showing the connections. *ELT Journal*, *51*(3), 208–216. doi:10.1093/elt/51.3.208

Jasper, F. (2010). Applied dimensionality and test structure assessment with the START-M mathematics test. *International Journal of Educational and Psychological Assessment*, *6*(1), 104–125.

Linacre, J.M. (2011). *Winsteps®*. Beaverton, Oregon: Winsteps.com. Retrieved from http://www.winsteps.com/

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/BF02296272

Miller, G.A., & Fellbaum, C. (1991). Semantic networks of English. *Cognition*, *41*(1–3), 197–229. doi:10.1016/0010-0277(91)90036-4

Muthén, L., Muthén, B., Asparouhov, T., & Nguyen, T. (2011). *Mplus*. Los Angeles, CA: Muthén & Muthén.

Nation, I.S.P. (1990). *Teaching and learning vocabulary*. Boston, MA: Heinle & Heinle.

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

Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes*, *18*(2), 161–175. doi:10.1016/S0889-4906(98)00005-2

Pawley, A., & Syder, F.H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J.C. Richards & 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), 513–536. doi:10.1111/1467-9922.00193

Qian, D.D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*(1), 28–52. doi:10.1191/0265532204lt273oa

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355–371. doi:10.1177/026553229301000308

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.

Richards, J.C. (1976). The role of vocabulary teaching. *TESOL Quarterly*, *10*(1), 77–89. doi:10.2307/3585941

Schmitt, N., Ng, J.W.C., & Garras, J. (2011). The word associates format: Validation evidence. *Language Testing*, *28*(1), 105–126. doi:10.1177/0265532210373605

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55. doi:10.1177/026553220101800103

Schoonen, R., & Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing*, *25*(2), 211–236. doi:10.1177/0265532207086782

Schumacker, R.E., & Lomax, R.G. (2004). *A beginner's guide to structural equation modeling* (2nd ed.). New Jersey: Psychology Press.