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

Papers by Tim Stoeckel & Phil Bennett
Hye Won Shin
Kurtis McDonald & Mayumi Asaba
Stuart McLean, Brandon Kramer & Jeffrey Stewart
Anna C. S. Chang
Yuko Hoshino
Tatsuya Nakata
Andrew Gallacher

Commentaries by Rie Koizumi
Stuart Webb

Vocabulary Learning and Instruction

Volume 4, Number 1,

October 2015

doi: http://dx.doi.org/10.7820/vli.v04.1.2187-2759

VLI Editorial Team

Editor: Raymond Stubbe

Editorial Board: Jeffrey Stewart, Luke Fryer, Charles J. Anderson, Aaron Gibson, Peter Carter

Reviewers: Paul Meara, Norbert Schmitt, John A. Read, Stuart Webb, John P. Racine, Ishii Tomoko, Tim Stoeckel, Dale Brown, Joh Clenton, Stuart McLean, Peter Thwaites, Tatsuya Nakata, Kiwamu Kasahara, Masumi Kojima, James Rogers, Yuko Hoshino, Vivienne Rogers

Copy Editors: Alex Cameron, Andrew Gallacher, Peter Harold, Mark Howarth, Linda Joyce, Tim Pritchard, Zelinda Sherlock, Andrew Thompson, Alonzo Williams

The Editorial Team expresses a sincere thank you to *Mana Ikawa*, who designed the cover for the print version of VLI.

Copyright © 2015 *Vocabulary Learning and Instruction*, ISSN: Online 2187-2759; Print 2187-2767. All articles are copyrighted by their respective authors.

Vocabulary Learning and Instruction

Volume 4, Number 1, October 2015
doi: http://dx.doi.org/10.7820/vli.v04.1.2187-2759

Table of Contents

Articles	Page
Letter from the Editor *Raymond Stubbe*	iv
A Test of the New General Service List *Tim Stoeckel & Phillip Bennett*	1
Psychometric Properties of Word Association Test with Regard to Adolescent EFL Learners *Hye Won Shin*	9
"I Don't Know" Use and Guessing on the Bilingual Japanese Vocabulary Size Test: A Preliminary Report *Kurtis McDonald & Mayumi Asaba*	16
An Empirical Examination of the Effect of Guessing on Vocabulary Size Test Scores *Stuart McLean, Brandon Kramer & Jeffrey Stewart*	26
Second Language Vocabulary Assessment Studies: Validity Evidence and Future Directions *Rie Koizumi*	36
An Investigation of Different Text Levels on L2 Learners' Vocabulary Learning Rates in an Extensive Reading Program *Anna C-S. Chang*	47
Measuring Knowledge of Words with Multiple Meanings *Yuko Hoshino*	58
Are Learners Aware of Effective Ways to Learn Second Language Vocabulary from Retrieval? Perceived Effects of Relative Spacing, Absolute Spacing, and Feedback Timing on Vocabulary Learning *Tatsuya Nakata*	66
Mastery Sentences: A Window into the Interplay between Word Knowledge Types *Andrew Gallacher*	74
Researching Vocabulary in the EFL Context: A Commentary on Four Studies for JALT Vocabulary SIG *Stuart Webb*	83

Letter from the Editor

Dear Readers,

It is my great pleasure to offer our readers the papers presented at the *4th JALT Vocab SIG Vocabulary Symposium*, held in Fukuoka, Japan, on June 20, 2015. In these pages you will find a variety of articles with topics ranging from guessing on the *Vocabulary Size Test* to the effectiveness of retrieval-based vocabulary learning techniques. You will also find two expert commentaries by Rie Koizumi and Stuart Webb.

If you have not already looked, there is also a new article by George Higginbotham, Ian Munby and John P. Racine on the *Online First* page of the *VLI* website as well as descriptions of four SLA PhD programs, three focusing on Vocabulary and one on Reading.

Please enjoy the pages to follow, and keep those submissions to *VLI* coming.

Raymond Stubbe,

Editor, VLI

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

Article

A Test of the New General Service List

Tim Stoeckel[a] and Phil Bennett[b]
[a]*University of Niigata Prefecture;* [b]*Miyazaki International College*
doi: http://dx.doi.org/10.7820/vli.v04.1.stoeckel.bennett

Abstract

This paper introduces the New General Service List Test (NGSLT), a diagnostic instrument designed to assess written receptive knowledge of the words on the New General Service List (NGSL) (Browne, 2014). The NGSL was introduced in 2013 as an updated version of West's (1953) original General Service List. It is comprised of 2,800 high frequency headwords plus their inflected forms and is designed to provide maximal coverage of modern English texts. The test introduced here is divided into five 20-item levels, each assessing a 560-word frequency band of the NGSL. Using a multiple choice format, the NGSLT is intended to assist teachers and learners in identifying gaps in knowledge of these high frequency words. Data from 238 Japanese university students indicate the NGSLT is reliable (α = .93) and that it measures a single construct. A comparison of NGSLT and Vocabulary Size Test (Nation & Beglar, 2007) scores for a small group of learners shows that the NGSLT provides more detailed diagnostic information for high frequency words and may therefore be of greater pedagogic use for low and intermediate level learners. Ongoing developments include parallel versions of the NGSLT as well as a separate instrument to assess knowledge of the New Academic Word List. Both the NGSLT and New Academic Word List Test are freely downloadable from the NGSL homepage (www.newgeneralservicelist.org).

1 Introduction

The impact of vocabulary knowledge on more general language proficiency is now widely acknowledged. Studies by Stæhr (2008) and Milton, Wade, and Hopkins (2010) have shown healthy correlations between written and aural receptive vocabulary size and tests of the four main language skills. Additionally, the findings of corpus-driven analyses (Hanks, 2013; Hoey, 2005) are now putting vocabulary knowledge at the heart of the language learning process by revealing the extent to which lexical choice influences and determines syntactic structure in the creation of meaning.

Accordingly, vocabulary assessment can fulfill important roles in language education for teachers, learners, and researchers. In classroom settings, having reliable estimates of vocabulary knowledge enables teachers to provide suitable materials for learners' needs, to judge the efficacy of a course of study, and to set appropriate goals for further development. Vocabulary goal-setting is particularly important, as successful learners will have a desire to build vocabulary, an

Please cite this article as: Stoeckel T. & Bennett, P. (2015). A test of the new General Service List. *Vocabulary Learning and Instruction*, *4* (1), 1–8. doi: 10.7820/vli.v04.1.stoeckel.bennett

awareness of the particular words that are most likely to benefit them, and the capability to achieve their goals (Dörnyei, 2005; Nation, 2001). For research purposes, vocabulary tests could be used to better understand the relationship between lexical knowledge and other skills, to assess the impact of learning experiences on lexical development, and to measure lexical growth.

Vocabulary tests are typically based on the frequency model – the notion that the more often a word is encountered, the more likely it is to be known. While other factors, such as polysemy, cognate status, and orthographic or phonological form can affect the difficulty of individual words, for 60–80% of learners, it seems that knowledge of bands of words grouped by frequency generally decreases as the bands become less frequent (Brown, 2012; Milton, 2009). Additionally, high-frequency words offer a disproportionally high percentage of text coverage, making them important to prioritize over mid- and low-frequency words. Thus, the frequency model remains a useful principle in developing vocabulary assessment tools and word lists.

2 The General Service List and New General Service List

Since frequency has such an impact on vocabulary assessment, it is important that frequency counts are accurate and reflect modern usage patterns. For many years, the General Service List (GSL; West, 1953) was used to provide this information. The GSL was not based purely on frequency, although this was one of its criteria. This list contains around 2,000 headwords, with related forms grouped underneath into word families. Although the GSL was originally intended to aid the development of reading materials, it later became the basis for test development (Schmitt, Schmitt, & Clapham, 2001) or a prerequisite for further word lists (Coxhead, 2000). However, the rapid growth of large computerized corpora and revised opinions over what should be considered a "word" have made the GSL appear somewhat dated.

The New General Service List (NGSL; http://www.newgeneralservicelist.org) is an attempt to address these concerns while preserving the goal of the GSL, which is to include those items that provide maximal coverage of texts with as few headwords as possible. The NGSL is based on frequency and dispersion data in a 273 million-word sample of the Cambridge English Corpus. The sample contains texts drawn from fiction, journals, magazines, non-fiction, radio, documents, TV, spoken interactions, and learner output.

The NGSL also differs from the GSL in its grouping of words. The original GSL was inconsistent in how it categorized derived forms under each headword. Although an updated version of the GSL (Bauman & Culligan, 1995) brought more consistency to the organization of the list, it has been recognized that, for learners below advanced levels, knowledge of a headword does not guarantee understanding of all of the derived forms in a word family (Milton, 2009; Vermeer, 2004). The NGSL resolves this issue by grouping words into "modified lemmas." A lemma is simply a headword and its inflected forms, with different parts of speech belonging in separate lemma groups. The NGSL's modified lemma approach varies this slightly by considering a headword as all parts of speech with the same written form, and then including all of the various inflected forms. For example, the modified lemma for *stage* would include *stages*, *staged*, and *staging* as verbal inflections and *stages*, *staging*, and *stagings* as

nominal inflections. While the NGSL contains around 400 more word families than the GSL, it has approximately 800 fewer lemmas and provides 6% more coverage of a sub-section of the Cambridge English Corpus (Browne, 2014).

3 Currently Available Tests

Two well-known vocabulary assessment instruments are the Vocabulary Levels Test (VLT; Nation, 1983; Schmitt, Schmitt & Clapham, 2001) and the Vocabulary Size Test (VST; Nation & Beglar, 2007), both of which assess written receptive vocabulary knowledge. The VLT provides diagnostic information for the first, second, fifth, and tenth 1,000-word frequency bands and has a separate section for the Academic Word List. The test uses a 6:3 matching format, with six answer choices provided from the same frequency band, three of which must be matched to given definitions.

The VST assesses learners' knowledge of the first 14 (or more recently, 20) 1,000-word frequency bands based on word families appearing in the British National Corpus. The test uses a typical four-option multiple-choice format and is intended to provide an estimate of overall vocabulary size, rather than a reliable indication of how well each frequency band is known.

4 The New General Service List Test

This paper introduces the New General Service List Test (NGSLT), a diagnostic instrument designed to assess written receptive knowledge of the words on the NGSL. This purpose differs from that of the VLT, which provides a broad estimate of learners' vocabulary profiles across selected, non-consecutive frequency bands, and from that of the VST, which estimates overall vocabulary size. The NGSLT also differs from these other tests in that it is based on an underlying list of modified lemmas rather than word families. As such, there are fewer assumptions associated with correctly answered items on the NGSLT. The NGSLT is intended to diagnose mastery of each of five 560-word levels of the NGSL and to identify gaps in knowledge of these high-frequency words. The size of each of these levels constitutes a manageable goal for one semester of intensive study, especially considering that non-beginners are likely to already know some words.

Test items follow the same specifications as those of the VST (Nation & Beglar, 2007). The stem consists of the target word followed by a sentence which uses this word in a non-defining context. Answer choices include three distractors

different: They are **different**.
a. easy to see
b. large
c. not easy
d. not the same

Figure 1. Example NGSLT item.

and the correct answer. When the target word has more than one possible meaning or use, the sample sentence and correct answer are based on the more common meaning or use as determined by consultation of concordance lines in the Corpus of Contemporary American English (http://corpus.byu.edu/coca/). Figure 1 depicts an example test item.

Items were written with high-frequency vocabulary. Whenever possible, items testing words in the first three bands were written only with words from the first two of these bands. Exceptions were the inclusion of *dirty*, *bottom*, *better*, and *repeat* in items which tested words in bands of the same or higher frequency as these words. Each item testing words in the fourth and fifth bands was written exclusively with words of higher frequency than the word being tested.

The test contains 100 items, 20 at each level. The items on the test were selected from a bank of over 200 items, each of which targets a randomly chosen member of the NGSL. These items have been piloted in Japanese colleges and analyzed to determine how likely learners are to know them (for details, see Bennett & Stoeckel, 2013). The items selected represent the range and average item difficulty at each level. As such, we are reasonably confident that the test is representative of the range of word difficulties at each level for Japanese learners.

We are currently collecting data for a validation study of the NGSLT, and preliminary results from 238 learners in four Japanese colleges are favorable. Taken as a whole, the test provides a reliable estimate of knowledge of the NGSL ($\alpha = .93$). The reliability coefficients for individual 560-word frequency bands are somewhat lower but acceptable, ranging from .70 to .80. Overall item quality also appears to be good. An inspection of point measure correlations as well as Rasch infit and outfit statistics has flagged just one item (Level 1: *teacher*) as misfitting. This misfit appears to have been caused by incorrect responses from three high-ability examinees. Data from a subset of our sample shows that NGSLT results also correlate with scores on the Test of English for International Communication (TOEIC), a test of general English proficiency: $r(33) = .72$, $p < .001$.

5 New General Service List Test Score Interpretation

For diagnostic purposes, one way to use the test is to examine learners' scoring profiles in order to identify the point at which they no longer have mastery of around 80–85% of the words in a band. This threshold is based on Milton (2009), who found that it is not unusual for proficient learners' scores to plateau at around this mark, rather than at 100% in tests of high-frequency words. Our own observations are consistent with this; students who have TOEIC scores of 800 or higher sometimes score as low as 75% on individual frequency bands on the NGSLT. Because high-frequency words are of such importance, it is probably beneficial for learners scoring less than 80% or so on a band to review a list of the complete word band (available at www.newgeneralservicelist.org) and to highlight unknown words in that band for further study. If teachers take this approach, they should exercise care in selecting a higher threshold for mastery, as learners could be presented with word lists that are not challenging enough to provide a suitable goal for a particular program of study (Dörnyei, 2005).

Table 1. A Comparison of TOEIC, VST, and NGSLT Results

			NGSLT (% score)				
			560-Word Level				
Examinee	TOEIC	VST[a]	1	2	3	4	5
1	345	2,400	**100**	**85**	80	70	55
2	265	2,400	75	65	40	45	30
3	485	3,400	**100**	**90**	**90**	75	55
4	405	3,400	**90**	80	**85**	75	70
5	310	3,400	**90**	75	70	50	55
6	240	3,400	80	70	65	45	40
7	950	6,200	**100**	**95**	**95**	**100**	**100**

Bold font indicates a score of 85% or higher on an NGSLT level
[a]Estimated vocabulary sizes based on the first eight levels of the VST (see McLean, Hogg, & Kramer, 2014, for reasons to administer abbreviated VST)

NGSLT results may be more informative pedagogically than VST scores, particularly for learners of low to intermediate proficiency. The VST supplies estimates of overall vocabulary size but not knowledge of individual frequency bands. This is illustrated in Table 1, which shows VST and NGSLT results for seven learners who are representative of a group of 33 students who took both tests. TOEIC scores are also shown in order to provide information regarding overall English proficiency. The first two learners obtained identical estimated vocabulary sizes of 2,400 word families with the VST, but NGSLT results show not only that each of these learners has gaps in knowledge of high-frequency words but also that their learning needs differ significantly. While learner 1 appears to have mastered the first two word bands, learner 2 could benefit from further study of the most frequent 560 words. Entries three through six show that even when VST results indicate knowledge of over 3,000 words, large gaps can persist among the most frequently occurring words. The last entry in the list illustrates the limitations of the NGSLT for diagnostic purposes with highly proficient learners.

The detailed profile of high-frequency vocabulary knowledge which the NGSLT provides can assist in pedagogical decisions such as appropriateness of extensive and intensive reading materials and vocabulary learning goals. Regarding the last of these, there is now a variety of useful study materials for the NGSL that can be accessed from the NGSL homepage (www.newgeneralservicelist.org/vocabulary-links/). Among these are free online flashcard programs which are divided into 50- or 100-word sets according to frequency level, and with results of the NGSLT, teachers and students can identify the most appropriate groups of words to study.

6 Future Research and Development

Though the NGSLT in its current state is capable of providing teachers and learners with valuable pedagogical information, there are ways to make it more useful. To minimize the risk of a testing effect when the instrument is repeatedly

used in educational programs, parallel forms should be developed. To this end, we hope to release a second 100-item form later this year. It would also be beneficial to create bilingual versions of the test with answer choices in the L1. As a test of high-frequency vocabulary, the NGSLT is intended for learners of low to intermediate proficiency, whose answers under a monolingual format may at times reflect lack of understanding of lexical or syntactic elements of the answer choices rather than knowledge of the target word (Elgort, 2013; Karami, 2012; Nguyen & Nation, 2011).

For learners of higher proficiency levels in academic contexts, we have also recently completed and begun piloting a two-level, 40-item test to assess written receptive knowledge of the New Academic Word List (NAWL; available at www.newgeneralservicelist.org/vocabulary-links/). This list was developed by the authors of the NGSL and is comprised of words that frequently occur in academic texts and that are not included in the NGSL (www.newacademicwordlist.org). The NGSL and NAWL tests are identical in format, meaning they can be used together seamlessly. For learners who have previously demonstrated partial mastery of the NGSL, teachers can easily tailor diagnostic tests to include only the upper levels of the NGSL plus the NAWL.

Our use of a multiple-choice format identical to that of the VST was intentional. Though it is becoming clear that score interpretations may need to be revised to reflect the possibility of correctly guessing unknown words (Stewart, 2014), the multiple-choice format is familiar to students, requires more than simple recognition of word form, and is quick and easy to mark. More importantly, the creation of a large pool of items written to the same specifications opens up the possibility of first calibrating these items to a single scale under item response theory, and then using them flexibly to assess different areas of lexical knowledge as needed.

Finally, further research of the NGSL itself is also needed. As stated, when compared to the original GSL, the NGSL has been reported to provide approximately 6% better coverage (84.24% versus 90.34%) of a sub-section of the Cambridge English Corpus (Browne, 2014). Though this is an impressive advantage, it must be remembered that the criteria for selecting members of the NGSL was derived from this same corpus, so we might expect superior coverage. Though Browne (2014) has also reported notable improvements in coverage with corpora of texts from *Scientific American* and *The Economist*, coverage of more general sources needs to be more thoroughly investigated. In addition, the degree to which the construct *modified lemma* is useful pedagogically needs to be examined. There is evidence that when compared to the word family, it is a step in the right direction because fewer assumptions are made regarding knowledge of related word forms (Gardner, 2007). However, the degree to which knowledge of a modified lemma's headword can be associated with all of its constituents has not yet been established.

References

Bauman, J., & Culligan, B. (1995). *The general service list*. Retrieved from http://jbauman.com/aboutgsl.html

Bennett, P., & Stoeckel, T. (2013). Developing equivalent forms of a test of general and academic vocabulary. In N. Sonda & A. Krause (Eds.), *JALT2012 Conference Proceedings* (pp. 636–644). Tokyo: JALT.

Brown, D. (2012). The frequency model of vocabulary learning and Japanese learners. *Vocabulary Learning and Instruction*, *1*(1), 20–28. doi:10.7820/vli.v01.1.brown

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction*, *3*(1), 1–10. doi:10.7820/vli.v03.2.browne

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238. doi:10.2307/3587951

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the vocabulary size test. *Language Testing*, *30*, 253–272. doi:10.1177/0265532212459028

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*, 241–265. doi:10.1093/applin/amm010

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. Cambridge, MA: MIT Press.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Abingdon: Routledge.

Karami, H. (2012). The development and validation of a bilingual version of the vocabulary size test. *RELC Journal*, *43*(1), 53–67. doi:10.1177/0033688212439359

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction*, *3*(2), 47–55. doi:10.7820/vli.v03.2.mclean.et.al.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, MA: Multilingual Matters.

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in English as a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. del Mar Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–98). Bristol, MA: Multilingual Matters.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

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

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13. Retrieved from http://jalt-publications.org/tlt/issues/2007-07_31.7

Nguyen, L. T. C., & Nation, I. S. P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal*, *42*(1), 86–99. doi:10.1177/0033688210390264

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), 55–88. doi:10.1177/026553220101800103

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*, 139–152. doi:10.1080/09571730802389975

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*, 271–282. doi:10.1080/15434303.2014.922977

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 173–189). Amsterdam: John Benjamins Publishing Company.

West, M. (1953). *A general service list of English words.* London: Longman, Green.

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, $p < .001$.

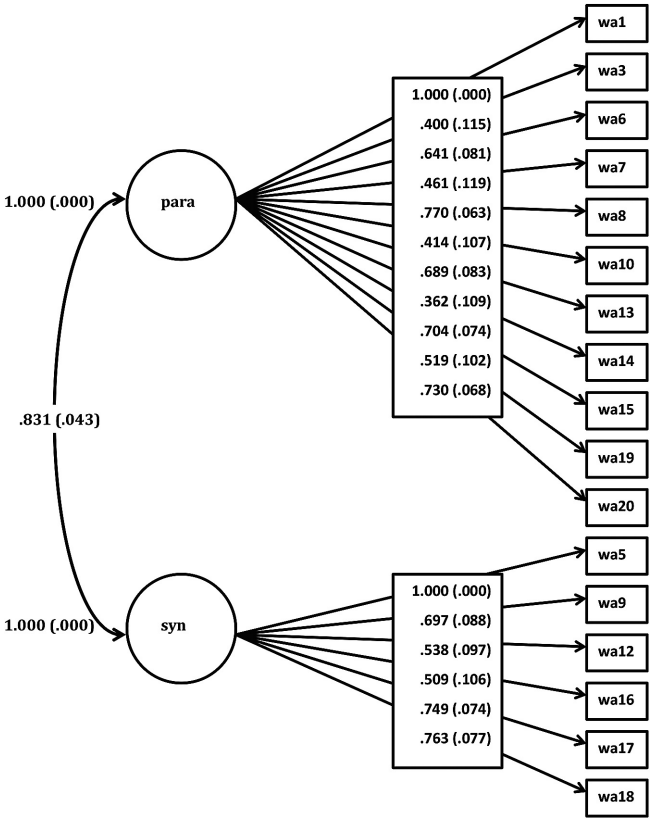

Figure 2. Diagram of two-factor CFA model.
Note: para denotes paradigmatic and syn denotes syntagmatic.

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

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

Article

"I Don't Know" Use and Guessing on the Bilingual Japanese Vocabulary Size Test: A Preliminary Report

Kurtis McDonald[a] and Mayumi Asaba[b]
[a]*Kobe College;* [b]*Kwansei Gakuin University*
doi: http://dx.doi.org/10.7820/vli.v04.1.mcdonald.asaba

Abstract

This preliminary report outlines an investigation into "I don't know" use and guessing on the 14,000-word family, 140-item bilingual Japanese Vocabulary Size Test (VST) translated by Sasao and Nakata from the original monolingual English version. Four first-year Japanese university students completed the modified Japanese VST in two passes: without guesses on unknown items and with guesses on these items. Individual semi-structured retrospective interviews were then conducted to identify how the guesses were determined. Findings suggest that "I don't know" use was largely consistent with learner proficiency and word family frequency levels and that guesses were more likely to be informed than uniformed. Using the classification of reasoning behind the guesses made, various vocabulary size estimates can be determined for each learner, with much greater differences between the more strict and more sensitive estimates found among the lower proficiency learners.

1 Background

With developments in corpora and vocabulary load analyses, interest in measuring second language (L2) learners' vocabulary sizes has grown in recent years as have proposals for how this can best be accomplished. The Vocabulary Size Test (VST; Nation & Beglar, 2007) represents one such proposal that has received increasing attention since its introduction. Designed as a relatively concise and easily-administered multiple-choice test, the original version of the VST seeks to assess written receptive vocabulary knowledge of the most frequently used 14,000 English word families of the spoken section of the British National Corpus using 140 randomly sampled items, 10 from each 1,000-word level. Each item presents a word in a simple, non-defining sentence stem and offers four potential definition options to choose from. Multiplying the number of items answered correctly by 100 provides a vocabulary size estimate up to the 14,000-word family level.

As use of the VST has increased, so has scrutiny into the validity and reliability of the vocabulary size estimates it generates. Although Beglar (2010) provided persuasive initial validation evidence for the VST using the Rasch model, subsequent research has identified a host of issues to be considered. Issues relating to the language in which the answer options should be presented (Elgort, 2013; Karami, 2012; Nguyen & Nation, 2011; Stewart, 2009), the degree to which

Please cite this article as: McDonald, K. & Asaba, M. (2015). "I Don't Know" use and guessing on the bilingual Japanese vocabulary size test: A preliminary report. *Vocabulary Learning and Instruction*, *4* (1), 16–25. doi: 10.7820/vli.v04.1.mcdonald.asaba

guessing might come into play (Stewart, 2014; Zhang, 2013), and the inclusion of "I don't know" among the answer options (Lucovich, 2014; Zhang, 2013), among others, have all been explored in recent literature, yet there are few informed claims of the impact of such factors on the validity and reliability of the vocabulary size estimates garnered by the VST.

This study seeks to contribute to the growing body of literature in this area by providing preliminary answers to the following research questions in relation to the bilingual Japanese VST:

(1) To what degree do learners use the "I don't know" option?
(2) How do learners answer self-identified unknown items?
(3) How much do vocabulary size estimates differ when different approaches to scoring guesses are applied?

1.1 Bilingual Versions

A review of the literature reveals growing support for the view that bilingual versions of the test can provide vocabulary size estimates that are more fair and accurate, particularly for lower proficiency learners (Elgort, 2013; Karami, 2012; Nation & Coxhead, 2014; Nation & Webb, 2011; Nguyen & Nation, 2011; Stewart, 2009). Bilingual versions of the VST utilize the same overall test format but present the answer options in the learners' first language (L1), such as the following example from the third 1,000-word family level:

Monolingual English version	Bilingual Japanese version
JUG: He was holding a **jug**.	JUG: He was holding a **jug**.
a. a container for pouring liquids	a. 水入れ
b. an informal discussion	b. おしゃべり
c. a soft cap	c. ベレー帽
d. a weapon that explodes	d. 拳銃

Figure 1. Example Bilingual Question.

Although both the item stems and answer options are presented in simplified English on the original monolingual English VST, Nation and Coxhead (2014) argued that bilingual versions are "more suitable than the monolingual test for native speakers of a particular L1" since they help avoid conflating L2 grammar and reading skills with vocabulary knowledge (p. 400). Beyond construct concerns, ease of test administration and affective considerations might also support the use of bilingual versions of the VST with certain populations of learners. Though questions about the validity of the figures garnered from bilingual versions of the VST remain, the limited amount of research conducted in this area to date has demonstrated that lower proficiency learners can expect higher vocabulary size estimates from bilingual versions of the VST, with increases in scores ranging from

10% (Elgort, 2013; Kobeleva as cited in Nation & Coxhead, 2014) to 17% (McDonald, 2014).

1.2 Guessing

The impact of guessing on the vocabulary size estimates provided by the VST is another key area of inquiry that has been raised in recent literature (McDonald, 2015; Nation & Webb, 2011; Nguyen & Nation, 2011; Stewart, 2014; Zhang, 2013). As shown in the example item above, the original VST was deliberately designed without an "I don't know" option in order to encourage informed guessing based on partial and intuitive knowledge (Nation, 2012; Nation & Webb, 2011). However, as a multiple-choice test, Stewart has argued that the degree to which vocabulary size estimates can be skewed by guessing randomly or through the use of test strategies remains unclear, particularly for lower proficiency learners. Adding to the noted concerns about guessing, Stewart warned that bilingual versions of the VST likely make it easier for test strategies such as the elimination of recognized distractors to be employed even if the target word is completely unknown.

1.3 The "I Don't Know" Option

The inclusion of "I don't know" as a fifth answer option on all items has been proposed as one way to reduce guessing on the VST (Lucovich, 2014; Zhang, 2013). Zhang found that the inclusion of an "I don't know" option on the monolingual English VST "slightly improved reliability and discrimination capacity" and significantly lowered the subsequent overall scores by reducing "both random successful guesses and successful guesses guided by partial knowledge" (p. 808), particularly when coupled with the threat of a scoring penalty for wrong guesses. Contrasting the effects of the modified "I don't know" versions of the VST he used with the specifications of the original VST, Zhang concluded that the different versions of the test likely align with different purposes warranting more or less sensitivity to partial and subconscious knowledge.

In order to better understand how individual language learners selected their answers and employed the "I don't know" option on the 20,000-word family version of the monolingual English VST, Lucovich (2014) conducted one-on-one interviews with two high proficiency L2 English learners. Lucovich found that the learners relied on knowledge, partial knowledge, test strategies, indeterminate informed guesses, and uninformed guesses to comparable degrees when considering their answers on both the original version and the modified "I don't know" version of the VST and that the use of the "I don't know" option did function to supplant uninformed guessing to some degree, though not entirely.

2 Method

This study extends a larger research project which seeks to compare responses on modified "I don't know" versions of the bilingual Japanese VST and the monolingual English VST from 118 Japanese first-year students from various

departments at a small, private women's college in western Japan. In the current study, four volunteers out of the original 118 participants agreed to take the 140-item Japanese bilingual VST (translated by Sasao and Nakata and available on Paul Nation's website), which was modified to include a fifth "I don't know" option ("わかりません"), and submit to one-on-one, semi-structured retrospective interviews. The tests and interviews were conducted in January 2015 following previous administrations of shortened 100-item "I don't know" versions of the tests in April 2014 (Japanese VST) and September 2014 (English VST) in their required general English classes taught by one of the authors. All tests were administered on paper with as much time as needed given.

For the purposes of this study, several changes were made to the testing procedure when the participants were tested in January following methods first outlined by McDonald (2015). Besides being administered the full, modified "I don't know" version of the Japanese VST individually, the participants also received additional test instructions. As in previous administrations of the test, the participants were explicitly instructed not to guess on items they were unsure about, but to select "I don't know" in these instances instead. However, following their first pass through the test under these conditions, the participants were instructed to return to all of the items which they had originally responded to with "I don't know" and to complete a second pass through these items by selecting what they considered to be the best answer from the original four answer options by circling it in red.

Immediately after completing the test under these conditions, the learners participated in individual retrospective interviews to discuss the reasoning they employed to select answers on all of the items where they had initially chosen "I don't know." Interviews were conducted in Japanese by one of the authors and were audio-recorded upon the participants' permission. All participants were asked a series of questions about the "I don't know" items including:

(1) Have you ever seen or heard the word before?
(2) Do you understand all of the answer options?
(3) Did you eliminate any answers from consideration? If so, why?
(4) Why did you decide on the answer you selected?

A brief description of each participant and their test and interview times are listed in Table 1 with pseudonyms used.

Table 1. Summary of Participant Details, Test Times, and Interview Times

Participant	Major	TOEIC (IP) average	Test time (1st pass)	Test time (1st + 2nd pass)	Interview time
Rena	English	775	30 minutes	45 minutes	60 minutes
Risako	Intercultural studies	543	21 minutes	40 minutes	24 minutes
Rika	Psychology	380	19 minutes	32 minutes	37 minutes
Mari	Bioscience	305	14 minutes	26 minutes	46 minutes

Note: The TOEIC (IP) scores listed represent the average from two administrations of the test conducted in April 2014 and February 2015.

3 Findings and Discussion

Various analyses were applied to the data to answer the research questions posed in this study. First, each participant's test responses were recorded from both their first and second passes through the modified "I don't know" version of the 140-item bilingual Japanese VST administered in January. The number of "I don't know" responses from the first pass was tallied and the two sets of responses were used to tabulate scores both with and without guessing on the self-identified unknown items for each individual. Next, the interview data were subjected to content analysis and coding of the reasons given for the answers participants selected as their second pass guesses, completed first by the author who conducted the interview and subsequently confirmed by the other author. Minor discrepancies were noted and discussed until mutual agreement on all coding decisions could be reached. Finally, the coding of all second pass guesses was then linked to the answer selections in order to determine their contributions to a range of vocabulary size estimates if these responses were included or excluded from scoring decisions.

3.1 To What Degree Do Learners Use the "I Don't Know" Option?

All participants used the "I don't know" option extensively on the January administration of the bilingual VST. However, a noticeable difference can be seen between the total number of "I don't know" responses made by Rena (23) and the other participants, whose figures were much closer to one another: Risako (75), Rika (85), and Mari (81). This difference seems to coincide with comparable differences in the participants' overall English proficiency levels, as interpreted through their average TOEIC (IP) scores, as well as qualitative differences observed by both authors in the interview data. However, it must be acknowledged that this difference is likely also attributable to some degree of difference in the way the "I don't know" option was used among these participants, which could range from interpretations such as "I have no idea" on one end of the spectrum to "I'm not completely sure" on the other. Though beyond the scope of the current study, test taker criteria for "I don't know" use is certainly worthy of further investigation.

Figure 1 displays the distribution of "I don't know" responses on the January administration of the test by 1,000-word family frequency level for each participant. The overall trend of "I don't know" use progresses largely as one would anticipate among the less commonly used words. It is worth noting, however, that there are only three instances where individual participants responded that they did not know any of the words at any particular word family frequency level (Rika: 10k, 14k; Risako: 14k).

3.2 How Do Learners Answer Self-identified Unknown Items?

The analysis of the interview data allowed each guess on the self-identified unknown items to be classified into four main categories suggested by previous research and considered most pertinent to the aims of the current study:

(1) *True partial knowledge-informed guess*: A guess based on accurate self-perceived knowledge of the target word, one or more of its word parts, or a concept clearly associated with the word.

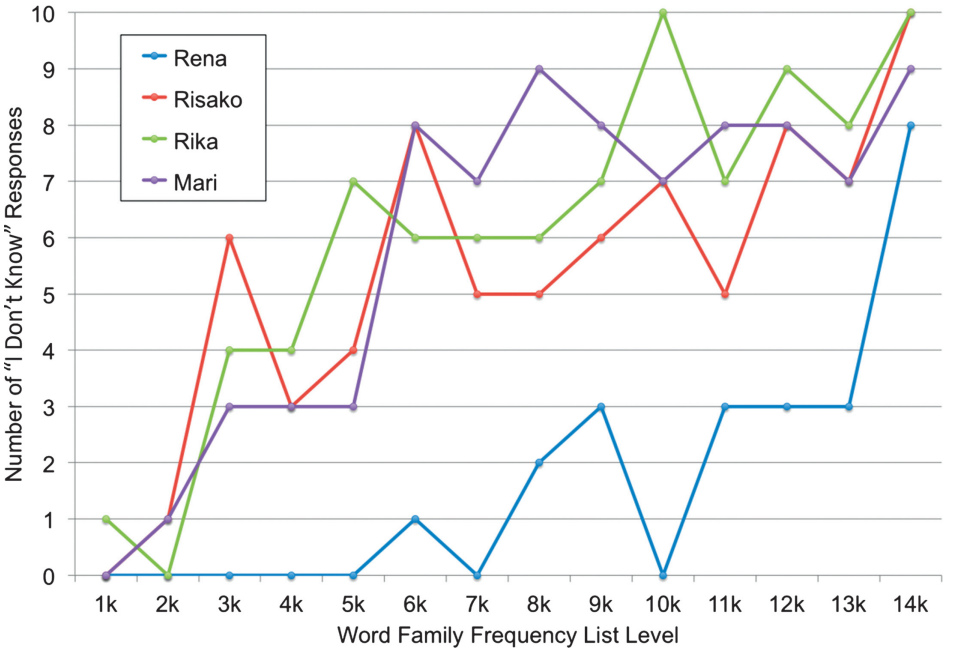

Figure 2. Number of "I don't know" responses for each word family frequency list level on the bilingual Japanese VST administered in January (k = 140).

(2) *False partial knowledge-informed guess*: A guess based on inaccurate self-perceived knowledge of the target word, misidentified or misunderstood word parts, or concepts mistakenly thought to be associated with the word.
(3) *Test strategy-informed guess*: A guess based on information gleaned from the item stem, correct answer option, and/or distractor(s).
(4) *Uninformed guess*: A completely random guess made without the use of any type of knowledge or test strategy.

However, the interview data also made it clear that the learners often used more than one type of reasoning when compelled to select an answer on an item they had first regarded as unknown (e.g., using both partial knowledge and test strategies to narrow the list of answer options). As such, it was often impossible to distinctly categorize answers into one category alone. Instead, for the purposes of this study and in keeping with the specified goals of the VST outlined by Nation (2012), a determination was made to prioritize any use of partial knowledge of a target word (whether true or false) over concurrent use of any test strategies which also may have contributed to the ultimate selection of one answer option over the others within our coding. Likewise, instances in which learners could not express the reasoning employed beyond describing some type of intuition were also coded as either true or false partial knowledge depending on the correctness of the answer selected.

Table 2 presents the number of guesses categorized by type for each participant. Using Rika's data as an illustrative example, of the 85 items that she originally responded to with "I don't know," she made completely uninformed

Table 2. Types of Guesses for Revisited "I Don't Know" Items on the Bilingual Japanese VST

			Informed guesses			
Participant	Total number of "I Don't Know" responses	Uninformed guesses	All informed guesses	True partial knowledge	False partial knowledge	Test strategy use alone
Rena	23	0	23	4	0	19
Risako	75	34	41	21	5	15
Rika	85	29	56	21	14	21
Mari	81	15	66	18	28	20

Note: $k = 140$. Total number of "I Don't Know" Responses = Uninformed Guesses + All Informed Guesses. All Informed Guesses = True Partial Knowledge + False Partial Knowledge + Test Strategy Use Alone.

guesses on 29 items and informed guesses of some kind on 56 of the items when these items were revisited. Of the 56 informed guesses, 21 could be attributed to true partial knowledge, 14 to false partial knowledge, and 21 to test strategy use alone.

The data presented in Table 2 show that all four participants employed some kind of informed guessing more often than completely uninformed or random guessing. At the two ends of the spectrum, a high of 100% of Rena's 23 total responses to self-identified unknown items could be attributed to informed guesses while a low of 55% of Risako's 75 total guesses could be ascribed to informed reasoning of some kind.

Although the separation of partial knowledge from test strategy use on items where both types of information seem to have contributed to the selection of an answer presents challenges to the type of classification attempted in this study, the distinction between demonstrations of any measure of partial knowledge of the target word (e.g., recognizing a familiar word part) and attempts to employ test strategies completely unrelated to an understanding of the word in question (e.g., eliminating a distractor considered unsuitable for the context provided by the item stem) was relatively clear. Likewise, separate figures for guesses guided by both true and false partial knowledge could also be tallied. As shown in Table 2, the distribution of true partial knowledge-informed guesses to false partial-knowledge informed guesses seems to coincide closely with the assessed TOEIC (IP) proficiency levels of all four participants. Indeed, Rena, the most highly proficient learner, never used false partial knowledge to inform her guesses while Mari, the lowest proficiency learner of those interviewed, applied false partial knowledge (largely due to confusion of the target word with other unrelated words or faulty intuitions) more often than true partial knowledge when making guesses.

3.3 How Much Do Vocabulary Size Estimates Differ When Different Approaches to Scoring Guesses Are Applied?

Table 3 displays the scores for each participant when the various self-identified guesses are included or excluded from scoring. As the data reveal, the difference in an individual's score can increase dramatically when various types of guesses are included. Although Rena's score only increases by seven points (8%)

Table 3. Scores for Each Participant on the Bilingual Japanese VST

Participant	Score without guesses	Score with true partial knowledge-informed guesses	Score with all partial knowledge-informed guesses	Score with all informed guesses	Score with all guesses
Rena	84	85	85	91	91
Risako	54	70	72	80	89
Rika	51	69	70	76	83
Mari	50	62	72	80	81

Note: $k = 140$.

when her score including all guesses is compared with her score excluding all guesses, the other three participants' scores all improve by over 30 points (38–39%). When these figures are multiplied by 100 to establish vocabulary size estimates as the design of the VST suggests, this equates to a difference of over 3,000-word families between the strictest and most lenient estimates.

The range of scores displayed in Table 3 illustrates the increasing degrees of sensitivity (from left to right) able to be distinguished when responses to self-identified unknown items are identified and the reasoning behind the answers ultimately selected are categorized. Although Nation (2012) provided no definitive guidance on what level of sensitivity would be ideal, considering both the stated goals of the VST and the need for the vocabulary size estimates it generates to have validity as measures of vocabulary knowledge within its approach, it can be argued that the most sensible figures would include only those answers based on perceived knowledge (the original answers selected on items perceived as known) and those informed by *true partial knowledge* (when items originally identified as unknown were revisited on the second pass), represented by the scores listed in the third column of Table 3. At the very least, the danger of attempting to draw any firm conclusions about these learners' vocabulary sizes from the raw scores that include all of their guesses should be clear since these scores almost completely obfuscate the very real differences in vocabulary knowledge held between these participants. While the validity of the vocabulary size estimates that any of the scores would suggest remains undetermined, the ability to calculate estimates exclusively tied to knowledge of some kind, be it self-perceived knowledge or true partial knowledge, seems to mark a step in the right direction and one certainly worthy of further research.

4 Conclusion

The findings outlined in this preliminary report suggest that conscientious Japanese university students who take a modified "I don't know" version of the bilingual Japanese VST do make use of this option in ways largely consistent with their proficiency levels, at least in one-on-one administrations of the test. If compelled to respond to unknown items, reasoning based on both true and false partial knowledge and/or test strategies is employed, when available, and random guessing is used if not. Classifying the reasoning behind the guesses made on unknown items allows for a number of different scores to be calculated, which can

accommodate a range of scoring approaches from more strict to more sensitive. The findings of this study suggest that the differences in the scores calculated at each end of this scoring spectrum may vary little for higher proficiency learners but a great deal for lower proficiency learners. The larger differences seen among the lower proficiency learners' scores point to the challenges of validity determinations of the estimates garnered from multiple-choice tests like the VST where learners apply varying degrees of true and false knowledge, true and false partial knowledge, test strategies, and uninformed guesses in unique ways. All of these issues warrant greater consideration if the value of the vocabulary size estimates produced by the VST for language learners is to be more fully understood.

References

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, *27*(1), 101–118. doi:10.1177/0265532209340194.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the vocabulary size test. *Language Testing*, *30*, 253–272. doi:10.1177/0265532212459028.

Karami, H. (2012). The development and validation of a bilingual version of the vocabulary size test. *RELC Journal*, *43*(1), 53–67. doi:10.1177/0033688212439359.

Lucovich, D. (2014). Test taking and DK use on the vocabulary size test. *Vocabulary Learning and Instruction*, *3*(2), 69–77. doi:10.7820/vli.v03.2.lucovich.

McDonald, K. (2014, September). How do monolingual and bilingual versions of the Vocabulary Size Test compare? Paper presented at the 18th Annual Conference of the Japan Language Testing Association, Ritsumeikan University, Shiga, Japan.

McDonald, K. (2015). The potential impact of guessing on monolingual and bilingual versions of the vocabulary size test. *Osaka JALT Journal*, *2*, 44–61.

Nation, I. S. P. (2012). *Vocabulary size test information and specifications*. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13.

Nation, P., & Coxhead, A. (2014). Vocabulary size research at Victoria University of Wellington, New Zealand. *Language Teaching*, *47*, 398–403. doi:10.1017/S0261444814000111.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle.

Nguyen, L. T. C., & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal*, *42*(1), 86–99. doi:10.1177/0033688210390264.

Stewart, J. (2009). A comparison of test scores between monolingual and bilingual versions of the Vocabulary Size Test: A pilot study. In A. M. Stoke (Ed.), *JALT2008 conference proceedings* (pp. 1262–1269). Tokyo: JALT.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*, 271–282. doi:10.1080/15434303.2014.922977.

Zhang, X. (2013). The "I don't know" option in the vocabulary size test. *TESOL Quarterly*, *47*, 790–811. doi:10.1002/tesq.98.

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

Article

An Empirical Examination of the Effect of Guessing on Vocabulary Size Test Scores

Stuart McLean[a], Brandon Kramer[b] and Jeffrey Stewart[c]
[a]*Kansai University;* [b]*Momoyama University;* [c]*Kyushu Sangyo University*
doi: http://dx.doi.org/10.7820/vli.v04.1.mclean.et.al

Abstract

The Vocabulary Size Test (VST) was created to provide a reliable estimate of a second language learner's written receptive vocabulary size, measuring from the most frequent fourteen 1,000 word families of the spoken subsection of the British National Corpus. While some have recommended that users should limit the amount of the test taken to only slightly above a student's level, others argue that learners should take every level of the test. However, this raises concerns that correct responses on lower frequency levels could largely be attributed to guesses rather than vocabulary knowledge. In this paper we analyze a data set of 3,373 Japanese university students' responses to the first eight levels of the original VST under the 3PL model, in order to determine the minimum expected score on the test for learners of low ability, examine the proportion of low-level students' scores on the lowest frequency level tested that can be attributed to guessing under the 3PL model, and conduct a model fit comparison to determine whether the 3PL model offers a significantly better description of the data than the Rasch model. The results indicate that a substantial portion of lower level learners' scores on items testing low-frequency words can be attributed to guessing and support the position that students should not sit every level of the test. The authors recommend using the results of the 3PL analysis in order to determine which sections of the test learners of different proficiency levels should sit.

1 Background

1.1 Introduction

The Vocabulary Size Test (VST) was created to provide a reliable estimate of a second language learner's written receptive vocabulary size, measuring from the most frequent fourteen 1,000 word families of the spoken subsection of the British National Corpus (Nation & Beglar, 2007). As the original VST tests 140 of these 14,000 most common words, vocabulary estimates are derived by multiplying the raw score on the test by 100, under the assumption that each correct answer can be considered equivalent to knowledge of each tested word.

Nation (2012), Karami (2012), Nguyen and Nation (2011), and Coxhead, Nation, and Sim (2014) argue that learners should attempt every level of the test, as

Please cite this article as: McLean et al. (2015). An empirical examination of the effect of guessing on vocabulary size test scores. *Vocabulary Learning and Instruction*, *4* (1), 26–35. doi: 10.7820/vli.v04.1.mclean.et.al

learners may know some low-frequency words in a "Slumdog Millionaire" effect. Beglar (2010) and Elgort (2013), however, recommend that learners should not take more than two levels above their ability. In support of this position, Stewart (2014) further argues that we should limit the levels of the test that are administered because use of a four-option multiple-choice format implies a 25% chance that a learner can guess correctly. Therefore, correct answers by lower level learners at more challenging levels of the test could likely be attributed to random guessing. Furthermore, a new version of the VST (Coxhead, Nation, and Sim, 2014) samples the first 20,000 words, initially with 10 items and currently only 5 items per 1,000-word band. This increases the likelihood that a sizeable portion of learner size estimates could be attributed to guessing.

To what degree do guesses by lower level learners on multiple-choice items testing less frequent words indicate knowledge of them, and to what degree do they constitute luck? Stewart (2014) suggested that the three-parameter logistic (3PL) model (Birnbaum, 1968), an item response model similar to the Rasch model, could be useful in helping to empirically determine the degree to which random guesses inflate scores. Although it does not account for guessing behaviors by higher proficiency learners, the 3PL model estimates the probability that low-level learners can correctly guess more difficult items entirely by chance. In this paper, we analyze a data set of 3,373 Japanese university students' responses to the first eight levels of the original VST under the 3PL model. The purpose of this paper is to determine the minimum expected score on the test for learners of low ability, examine the proportion of students' scores on the lowest frequency level tested that can be attributed to guessing under the 3PL model, and conduct a model fit comparison to determine whether the 3PL model offers a significantly better description of the data than the Rasch model.

1.2 The Rasch Model and the 3PL Model: A Brief Primer

In order to explain the method of analysis used in this paper, this section briefly compares and contrasts the Rasch model and the 3PL model. It should be noted that the 3PL model is used in this study and compared to the Rasch model for analytical purposes, rather than as an endorsement of test scoring or vocabulary size estimation under the 3PL model. For further information on both models, please consult De Ayala (2009).

The VST was validated by Beglar (2010) using the Rasch model (Rasch, 1960). Under the Rasch model, the likelihood that a given test taker will correctly answer a given test item is modeled as a logistic function of the difference between the learner's ability level (the person parameter) and the difficulty of the test item (the item parameter). This relationship is depicted in Figure 1. The horizontal axis indicates learner ability expressed in "logits." The vertical axis indicates the probability of a correct answer given the learner's ability level. Finally, the "s"-shaped line depicts the Rasch model's logistic function, which describes the relationship between ability and the probability of a correct answer on a given item.

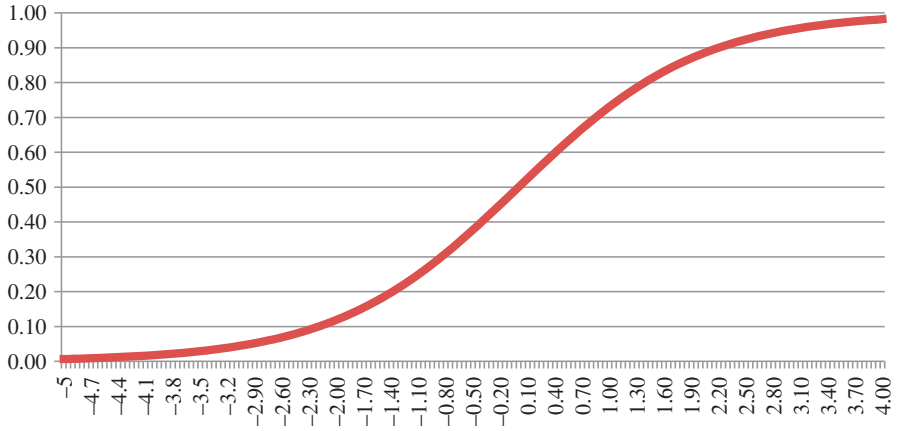

Figure 1. An Example of a Test Item Fit to the Rasch Model.

The Rasch model assumes minimal guessing; given a low enough level of ability, a learners' probability of answering given items should approach 0. However, this may not be the case with multiple-choice items. In cases where lower level learners do not know the answer to test items, they may guess the answer from the available options. Because of this, the probability of a correct guess may never actually fall to near 0. The 3PL item response model accounts for this with a ''pseudo-guessing'' parameter, which models minimum probabilities of correct answers that can be attributed to guessing, and, due to the use of a multiple-choice format, no longer drop with lower levels of ability. Figure 2 depicts a four-option multiple-choice test item, in which the minimum probability of a correct answer never falls below 0.25, even for learners of very low ability.

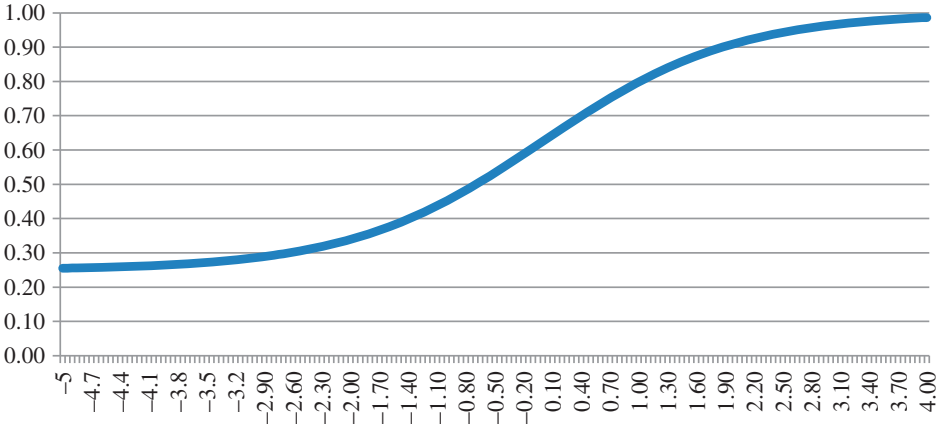

Figure 2. An Example of a Test Item Fit to the 3PL Model.

Although we typically use Rasch models in our work and research, we believe the 3PL model has useful analytical properties in regards to the validity of vocabulary size estimates made using the VST. As vocabulary size estimates derived from the test scores are based on the assumption that answering a multiple-choice question correctly implies knowledge of the tested word, a concern regarding the VST is that a substantial component of test scores – and therefore of its suggested vocabulary size estimates – could be attributable to guessing that is unrelated to proficiency, thereby inflating estimates. As the 3PL models a "flat" guessing rate for very low-level learners that is no longer related to changes in ability, if the VST has better fit to the 3PL model than to the Rasch model, it could help determine whether, and at what point, raw scores on the test cease to decrease with lower levels of proficiency, indicating the minimum vocabulary size estimates the test will provide for low-level learners.

2 Methodology

2.1 Instrument

This study utilized the first eight 1,000-word levels of Nation and Beglar's (2007) VST.

2.2 Participants

This study used a cross-sectional design with data from 3,373 Japanese university students collected through a snowball sampling approach, all of whom attempted to answer every question on the test as per Nation's (2012) instructions. The participants were from a range of Japanese universities representing a full range of abilities, as determined by examination of *hensachi* rank scores. A *hensachi* is a score assigned to individual students or school departments based on student performance on a national test standardized to the national mean across five subjects. A *hensachi* of 50 represents the mean, where one standard deviation above or below is represented by 60 or 40, respectively. The scores can range from 20 to 80, but 95.4% of all university departments fall between 30 and 70 (Newfields, 2006). Department *hensachi* scores were obtained from Benesse, a large testing company in Japan <http://shinken.zemi.ne.jp/hensachi>.

McLean, Hogg, and Kramer (2014) demonstrated that the mean *hensachi* of a participant's department was a good predictor of participants' ability, with a significant difference in VST scores between participants from three *hensachi* groups: $\geq$61, 51–60, and $\leq$50 ($F(2, 3{,}424) = 1{,}383.14$, $p < .001$, $\eta^2 = 45$). Further, participants' department *hensachi* scores correlated strongly with VST scores ($r = .73$, $p < .001$). Figure 3 shows that while the most common department *hensachi* of participants in the present study was 49 ($n = 1{,}030$), the majority of the data came from participants with average or above average department *hensachi* of 50 or above ($n = 1{,}940$), with an overall mean department *hensachi* score of 53.2 for all participants, which is somewhat above the national mean of 50.

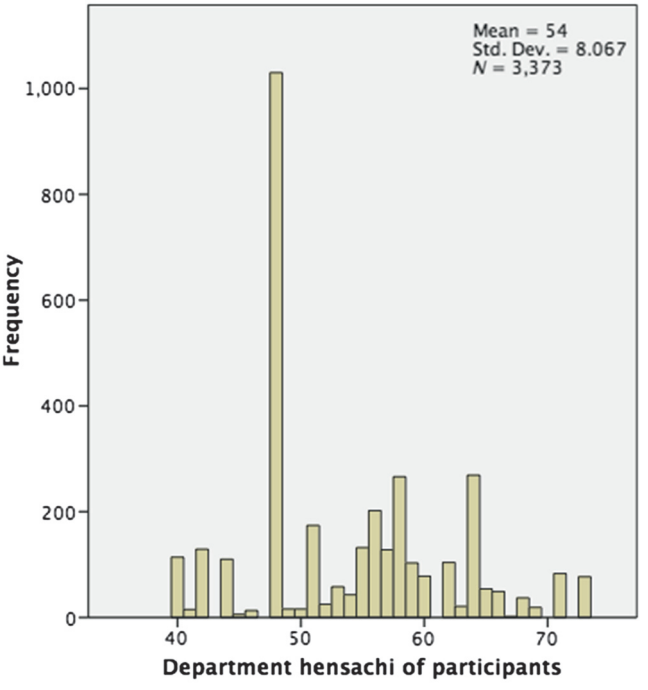

Figure 3. Histogram of Participants' Department *Hensachi*.

3 Data Analysis

3.1 Comparison of Models

The data set was analyzed with the software program IRTPro 2.1 (Cai, Thissen, & du Toit, 2011) under the Rasch model (from a technical standpoint, a two-parameter logistic model with slopes constrained to 1) and the 3PL model. However, due to the model failing to converge initially, it was necessary to remove the seven poorest items. These seven items had near-zero point-biserial correlations and negative slopes under a preliminary analysis using the two-parameter logistic model, meaning less able participants were *more* likely to answer these items correctly than more able participants, in violation of the model's assumption that guessing indicates a minimum probability of a correct answer, rather than a maximum. These items were item 10 (*basis*), 14 (*nil*), 29 (*rove*), 55 (*threshold*), 58 (*cavalier*), 65 (*bristle*), and 68 (*gimmick*). Two of these items (*basis* and *rove*) were reported by Beglar (2010) to perform poorly and underfit the Rasch model as well (standardized infit values $> + 2.00$), and Rasch analyses presented by McLean (2013) and McLean and Kramer (2014) found that all seven of the above items underfit the Rasch model (it should be noted that this study utilized the original version of the VST, and the version of the VST presently available online at Nation's website <http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-14000.pdf> includes an edited version of item 10, *basis*).

Table 1. Mean Raw Scores and Ability Estimates for Four Schools

Hensachi	Mean Score	3PL Mean Ability	Rasch Mean Ability
40	24	−1.48	−1.00
48	31	−0.66	−0.44
51	36	−0.07	−0.03
73	52	1.35	1.13

Note: The mean ability estimate for both models is 0.

Mean raw scores and scale score estimates under both models for schools with four separate *hensachi* ranks are listed in Table 1.

It is possible to graph the results of the Rasch model fit to determine what raw test score a learner at a given ability level would receive. In the Test Characteristic Curve depicted in Figure 4, the horizontal axis indicates student ability under the Rasch model, and the vertical axis indicates the raw score a learner at that ability level would be expected to receive on the test. Under this model, learners with lower ability levels than the university students tested in this study (for example, junior high or elementary school students) would be expected to receive lower test scores, with learners of very low ability expected to know almost none of the words, and therefore receive a score near 0.

However, the 3PL model's Test Characteristic Curve tells a somewhat different story (Figure 5). Under this model, a learner of very low ability would still have an expected score of 18 out of 73 or approximately 25%. As a sum score–scale score conversion table generated by IRTPro indicates that −3.349 is the lowest level of ability that the analyzed test form is capable of estimating, under the 3PL model this would appear to represent the practical minimum raw score estimate for the test, regardless of the ability level of the learner.

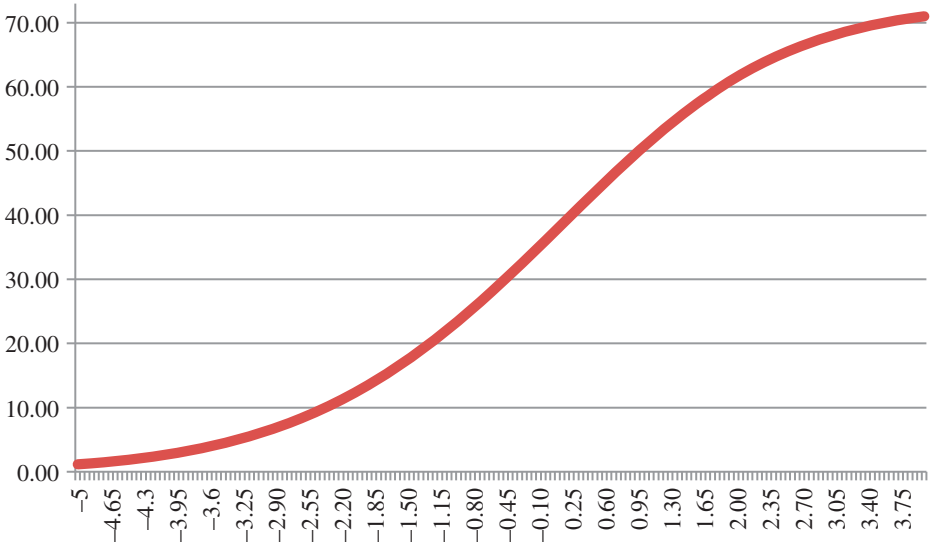

Figure 4. Test Characteristic Curve of the VST Data Set ($k = 73$) Fit to the Rasch Model.

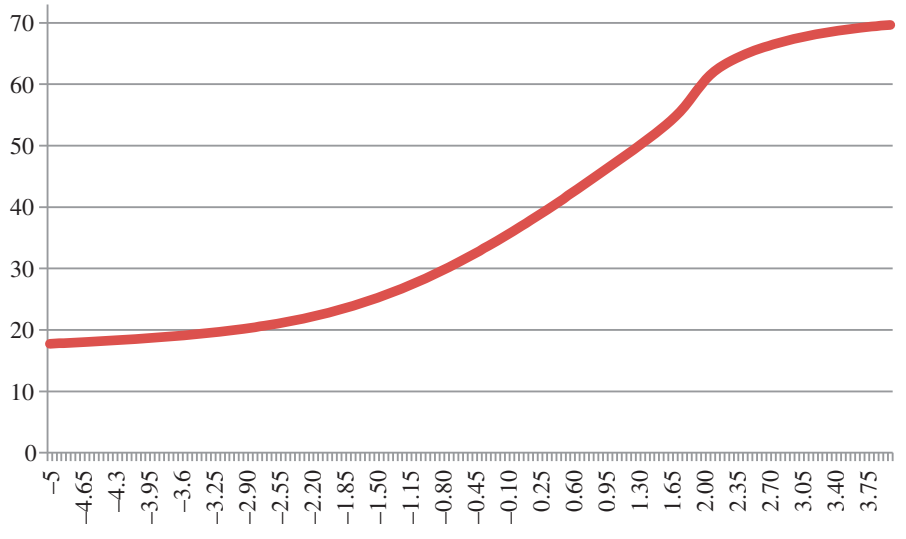

Figure 5. Test Characteristic Curve of the Data Set ($k = 73$) Fit to the 3PL Model.

3.2 Effect of Guessing on Items Testing Lower Frequency Words

A drawback of the current study is that only VST items testing the first eight 1,000-word levels were examined, and guessing would presumably be most prevalent with the lower frequency bands of the test. A further drawback is that poor items were removed from the 6k and 7k levels of the test prior to analysis, meaning it is not possible to estimate scores out of 10 for these levels. However, it is possible to examine students' scores out of 10 on the lowest frequency band tested, the 8k level, and calculate the proportion of their scores that could be attributed to uninformed guesses under the 3PL model.

As can be seen in Table 2, the mean score on the 8k level for departments with a *hensachi* rank score of 40 is 2.71 out of 10, just slightly above the estimated minimum score on this level for low-level learners. Given this, it appears that approximately 85.5% (2.32 / 2.71 $\times$ 100) of the mean raw score on this frequency band for students at this proficiency level (and therefore 85.5% of the 271 words in the band they would be estimated to know) could be attributed to uninformed guesses under the 3PL model. Even for the highest level departments tested, under the 3PL model, 41.7% of raw scores at this level can be attributed to guessing

Table 2. Mean Raw Scores Out of 10 on 8k Level of Test by Department *Hensachi*

Hensachi	Score on 8k Level	Minimum estimated score (3PL)	% of score below guess threshold
40	2.71	2.32	85.5
48	3.11	2.32	74.6
51	3.58	2.32	64.8
73	5.57	2.32	41.7

unrelated to proficiency. Although untested in this study, one can only assume that guesses would comprise even higher proportions of scores at even lower frequency levels (9k–20k levels).

3.3 Model Fit

The two item response models present differing descriptions of the VST in relation to the effect of uninformed guesses. Which item response model comes closer to describing the nature of the data? When assessing model fit, it should be noted that statistical models with more parameters (such as the 3PL) will nearly always demonstrate superior model fit to at least a negligible degree. However, if the difference is slight, the more complex model may simply overfit the data in a way that would not be generalizable to another data set (Zucchini, 2000). For this reason, IRT model fit is often assessed with fit statistics for nested models such as the Akaike Information Criterion (AIC; Akaike, 1998) or the Bayesian Information Criterion (BIC; Schwarz, 1978), which "penalize" additional parameters (see Table 3). However, under both statistics, values remain lower for the 3PL model, indicating superior fit. Given this, it appears the 3PL model provides a more accurate description of the test data.

Table 3. Model Fit Statistics

Statistics based on loglikelihood	Rasch	3PL
−2loglikelihood	275036.58	270066.76
AIC	275182.58	270504.76
BIC	275629.6	271845.82

4 Conclusion

Under the 3PL model, which had superior fit to the data, very low-level learners would be expected to receive a score of approximately 2.32 out of 10 due to guessing on the lowest frequency band of the test examined (the 8k level). As students in departments with a *hensachi* near the national mean of 50 had a mean score of 3.58 out of 10 on this word level, it seems that while average students' scores out of 10 on this word level are above chance, the bulk of their scores could be attributed to guessing that is unrelated to vocabulary knowledge.

This study has a number of limitations. Most importantly, the most difficult six levels of the original test were not taken, making the 8K level the lowest frequency word level students were tested on. Furthermore, seven items with very low point-biserial correlations could not be analyzed, and the overall mean department *hensachi* score for the sample was 53, which indicates the tested sample had a slightly higher proficiency level than the national average. However, despite these shortcomings, for most students guessing unrelated to proficiency appeared to have a greater effect on test scores on the lowest frequency band than proficiency did. Given these findings, it is difficult to endorse the position that the entire test be

given to students of all proficiency levels, and that they be encouraged to guess on items they do not believe they know the answers to. These results support the recommendation by Beglar (2010) and Elgort (2013) that learners not sit levels of the test well above their ability.

We suggest that the results of a 3PL analysis can be of use in determining precisely which sections of the VST learners of various proficiency levels should sit, as it can indicate when test items are too difficult to provide information about a given population. However, in cases where teachers wish to use a vocabulary test for pedagogical purposes, we recommend that rather than using a size test such as the VST, teachers use levels tests such as the Listening Vocabulary Levels Test (McLean, Kramer, & Beglar, 2015), parallel reading levels tests, or the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham, 2001), which, due to the typically larger numbers of items for each frequency band, can provide them with richer detail about the number of pedagogically relevant, higher frequency words that students know.

Acknowledgement

This research was supported in part by a grant from the Japan Society for the Promotion of Science (No. 24720278).

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected papers of Hirotugu Akaike* (pp. 199–213). New York, NY: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4612-1694-0_15

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27* (1), 101–118. doi:10.1177/0265532209340194

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison.

Cai, L., Thissen, D., & du Toit, S. (2011). *IRTPRO 2.1 for windows*. Chicago, IL: Scientific Software International.

Coxhead, A., Nation, P., & Sim, D. (2014). Creating and trialling six forms of the Vocabulary Size Test. *TESOLANZ Journal*, *22*, 13–26.

De Ayala, R. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing*, *30*, 253–272. doi:10.1177/0265532212459028

Karami, H. (2012). The development and validation of a bilingual version of the Vocabulary Size Test. *RELC Journal*, *43*(1), 53–67. doi:10.1177/0033688212439359

McLean, S. (2013, December). *Investigating university students' vocabulary sizes and the VST*. Paper presented at the Vocab@Vic Conference, Victoria University Wellington, New Zealand.

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the Vocabulary Size Test. *Vocabulary Learning and Instruction*, *3*(2), 47–55. doi:10.7820/vli.v03.2.mclean.et.al

McLean, S., & Kramer, B. (2014, September). *Investigating university students' vocabulary sizes and the VST*. Paper presented at 18th Annual Conference of the Japan Language Testing Association, Ritsumeikan, Japan.

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*. Advance online publication. doi:10.1177/1362168814567889

Nation, I. S. P. (2012, October 23). *The Vocabulary Size Test*. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13. Retrieved from jalt-publications.org/files/pdf/the_language_teacher/07_2007tlt.pdf

Newfields, T. (2006). Assessment literacy self-study quiz #1 [Suggested answers]. *Shiken*, *10*(2), 25–32. Retrieved from jalt.org/test/SSA1.htm

Nguyen, L. T. C., & Nation, P. (2011). A bilingual Vocabulary Size Test of English for Vietnamese learners. *RELC Journal*, *42*(1), 86–99. doi:10.1177/0033688210390264

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.* Copenhagen: Danish Institute for Educational Research.

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), 55–88. doi:10.1177/026553220101800103

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464. doi:10.1214/aos/1176344136

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*, 271–282. doi:10.1080/15434303.2014.922977

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, *44*(1), 41–61. doi:10.1006/jmps.1999.1276

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

Article

Second Language Vocabulary Assessment Studies: Validity Evidence and Future Directions

Rie Koizumi
Juntendo University
doi: http://dx.doi.org/10.7820/vli.v04.1.koizumi

Abstract

In this study, I review four papers by Stoeckel and Bennett; Shin; McDonald and Asaba; and McLean, Kramer, and Stewart. I will then summarize the validation evidence reported in each paper, in order to argue for the validity of the interpretations of the test scores as well as the uses of the tests considered in these four studies. This will help clarify areas of future research and strengthen the need for ties between specialists in the field of second language vocabulary assessment and general language assessment.

1 Introduction

The four studies I review are outstanding in terms of their research foci, designs, analyses, and interpretations for second language (L2) vocabulary assessment. However, the ultimate purposes of these studies are to understand how tests function and how test takers perform and to support or problematize test interpretations and uses. Therefore, it seems necessary to place the four studies in a wider framework of test validation. Validation frameworks help test developers/users understand the strengths and weaknesses of their tests (and research) and outline potential directions of study. As I argued in Koizumi (2015a), one strong candidate is an argument-based validation framework. This framework was presented systematically by Kane (1992, 2006) and is often employed when examining interpretations and uses based on test scores in L2 assessment studies (Chapelle & Voss, 2013). Kane's framework is rooted in Messick (1989) and American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA/APA/NCME, 1999), and test practitioners can implement validation more practically and systematically using this method (see Chapelle, Enright, & Jamieson, 2010, for the advantages of this approach over AERA/APA/NCME, 1999). An argument-based approach considers validity to be a matter of degree; therefore, tests can neither have perfect validity nor zero validity. Further, tests do not have high validity; instead, interpretations and uses based on test scores do. However, there are also stances for validation that oppose the argument-based approach (Markus & Borsboom, 2013; Newton & Shaw, 2014), and there are even variations among proponents of this approach (e.g., Bachman & Palmer, 2010; Kane, 2013; see Chapelle & Lee, 2013, for a review). In this review, I use Chapelle, Enright, and Jamieson (2008) as a

Please cite this article as: Koizumi, R. (2015). Second language vocabulary assessment studies: validity evidence and future directions. *Vocabulary Learning and Instruction*, *4* (1), 36–46. doi: 10.7820/vli.v04.1.koizumi

validation framework because their framework elaborates on the test constructs and test use (see Chapelle, Chung, Hegelheimer, Pendar, & Xu, 2010; Koizumi, Sakai, Ido, Ota, Hayama, Sato, & Nemoto, 2011; Purpura, Brown, & Schoonen, 2015, for examples).

There are two stages in the framework: (a) constructing an interpretive argument structure and making validity claims explicit, and (b) collecting evidence (i.e., backing) to support the interpretive argument structure and making a validity argument. Table 1 outlines the basic interpretive argument structure for L2 vocabulary tests. In this table, six inferences, warrants, and assumptions are used by way of an example. Each test developer/user can specify their framework. The number and descriptions of inferences, warrants, and assumptions in the structure can be modified (elaborated or simplified) across test contexts. For example, if the test is intended for educational purposes, all the inferences (from A to F) would come into play. If the test is to be used for research purposes, researchers can consider which inferences to include for their validation, based on the nature of the research. However, in the context of L2 vocabulary assessment, test scores are usually interpreted as reflecting a construct of L2 vocabulary, and examination seems needed for the Domain Definition inference to at least the Explanation inference (from A to D).

After those in charge of validation clarify the interpretive argument structure, in the second stage—that of collecting evidence to support the interpretive argument structure—they need to collect the evidence necessary to support the assumptions, which endorse a corresponding inference and warrant. When they finish collecting sufficient evidence, they develop a validity argument based on the interpretive argument structure and overall evidence, for example, "We argue that the interpretations and uses based on scores of this test are highly valid."

Ideally, researchers should obtain evidence from the Domain Definition inference and gradually progress to the higher inferences. However, this is difficult in practice, and they often find themselves collecting evidence separately for each inference in different phases and attempting to strengthen their validity argument gradually, along with more evidence. I will now briefly review and discuss each of the four abovementioned studies before summarizing the framework in the last section.

2 The New General Service List Test

Stoeckel and Bennett developed the New General Service List Test (NGSLT; Stoeckel & Bennett, n.d.) based on the NGSL (Browne, 2014) to measure written receptive knowledge and provide diagnostic profiles on the mastery of each level of the NGSL. The NGSLT has 100 multiple-choice format items (four options per item), all of which are in English. Stoeckel and Bennett administered the test to 238 students at Japanese universities and reported high reliability of the test and sections in each frequency band ($\alpha = .70$–$.93$); they also reported acceptably good quality of items in terms of discrimination and Rasch misfit analyses. They found a relatively high correlation ($r = .72$) between their test and the Test of English for International Communication (TOEIC). They further analyzed lexical profiles by showing the correct proportions for each band. They used 80% as the threshold for mastery and suggested that learners who fell in bands under 80% should review

Table 1. Inferences, Warrants, and Assumptions in the Interpretive Argument for an L2 Vocabulary Test

Inference	Warrant and assumptions (the latter numbered)
F. Utilization	Students/teachers can use test results to make decisions pertaining to their learning/teaching. *Use of the test is beneficial for learning/teaching. (1) Meanings of test scores and score reports are clearly interpretable by learners and teachers. (2) Students/teachers are willing to use test results in their learning/teaching process. Students use diagnostic results to make decisions on how to study. *Students/teachers perceive this test positively. *The use of the test provides learning/teaching opportunities by offering feedback on relevant vocabulary. (3) *Students/teachers understand the degree to which they/their students have mastered the vocabulary; they also understand their own strengths and weaknesses. *The use of the test facilitates their learning/teaching process.
E. Extrapolation	Test results are relevant to the L2 vocabulary learning context. (1) As intended, test scores are positively related to other indicators or test scores that reflect L2 skills or proficiency. (Test takers at different levels perform differently on the test, depending on their proficiency levels.)
D. Explanation	Test scores reflect the aspects of a construct of L2 vocabulary in the L2 learning context. (1) Observed test-taking processes accord with test developers' expectations. (2) A factor structure of the entire test corresponds to what would have been predicted. (3) Means of the vocabulary level/band sections correspond to what would have been predicted. (4) As intended, test scores are positively related to the scores on tests assessing similar or different types of knowledge. (5) Test scores reflect the intended construct and do not radically overestimate or underestimate the test-takers' knowledge.
C. Generalization	Test scores are consistent across test formats/test items. (1) The test includes a sufficient number of items and provides stable estimates of the test-takers' performance. (2) Test specifications are well defined so that parallel items can be created.

Table 1. (*Continued*)

Inference	Warrant and assumptions (the latter numbered)
B. Evaluation	Test results provide students/teachers/test users with accurate information. (1) Item statistical characteristics are appropriate. (2) Score reports are accurate, clear, and specific.
A. Domain definition	Test items cover relevant knowledge representative of L2 vocabulary. (1) Test items cover critical vocabulary knowledge needed for L2 use. (2) A test item format elicits knowledge relevant to and representative of the target domain.

Note. Adapted from Chapelle et al. (2008) and Chapelle, Cotos, and Lee (2015), whose terms, expressions, and visual displays are also adopted. According to Chapelle et al. (2015), the Utilization inference can be divided into the Utilization and Ramification inferences; the former pertains to the usefulness of test scores and score reports, whereas the latter is related to the washback effects of tests on learning and teaching. The warrant and assumptions pertaining to the latter are denoted by *.

lexical items in these bands. By way of future research and development, they mentioned creating parallel test forms and bilingual versions of the test and developing a large pool of items calibrated on a single scale.

The greatest strength of their research lies in its analysis, shown in Table 1 of their paper. It shows how the NGSLT can provide diagnostic profiles in frequency bands, by comparing them with TOEIC and Vocabulary Size Test (VST) scores. The analysis shows that students with the same VST estimates have different profiles: Students may have mastered high-frequency levels and gradually have less knowledge in lower frequency levels (Examinees 1, 3, and 4). Others may not have mastered high levels but had a substantial degree of knowledge in lower frequency levels, which can be called "fluctuating profiles" (as termed by Stoeckel and Bennett in their presentation; Examinees 2, 5, and 6). The patterns of learning displayed by the latter type of learner with fluctuating profiles are not usually expected. The results also show that the latter type tends to have lower TOEIC scores and, therefore, lower L2 proficiency. A similar trend was reported in Ota, Kanatani, Kosuge, and Hidai (2003). They reported that some junior high school students belonged to the latter type and tended to have lower proficiency and that one student developed their speaking ability more slowly than students not in the latter type. These may suggest that unstable knowledge without the firm basis of high-frequency vocabulary can inhibit the smooth development of L2 proficiency. An investigation into the performance and development paths of these learners would be an important future research inquiry. Recently, researchers have started to examine the types of profiles for four skills (reading, listening, writing, and speaking) and the underlying causes of uneven profiles (Ginther, Yan, & Potts, 2015; Huhta, Alderson, Nieminen, & Ullakonoja, 2015; Koizumi, 2015b). I believe that expected and unexpected patterns of vocabulary profiles could be strong predictors of uneven profiles and this type of inquiry could contribute to uncovering the nature of language proficiency.

For the NGSLT to be useful for learners and teachers, researchers may need to consider two points. First, if researchers are to claim the superiority of the

NGSLT over the VST in terms of diagnostic functions, they would need to present subscores for the VST for each frequency level so that a comparison can be made of the sensitivity of diagnosis between the two tests. If the NGSLT could provide clearer patterns or more useful information, then more people would use it. Second, learners and teachers will need a template of NGSLT score reports. Otherwise, some test users would only count the total number of correct items and not inspect the lexical profiles; this could result in misuse of the test. The effectiveness of score reports for learners and teachers is another possible research topic that can be pursued (see Doe, 2015; Jang, Dunlop, Park, & van der Boom, 2015; Sawaki & Koizumi, 2015).

3 Psychometric Properties of Word Association Tests with Regard to Adolescent EFL Learners

Shin developed and analyzed a modified version of the Word Association Test (WAT) with the aim of assessing vocabulary depth in adolescent learners of English in Korea. The original format (Read, 1993) requires test takers to select four answers (paradigmatic, syntagmatic, and/or analytic) out of eight options. Shin's format requires test takers to select one answer (paradigmatic or syntagmatic) out of six. She administered the test to 121 sixth-grade elementary school students. She reported high reliability for the entire test ($\alpha = .83$) when she deleted items with low discrimination. She also reported a moderate correlation between the WAT scores and the scores of paradigmatic and syntagmatic relationships with a reading test ($r = .55$–$.63$), and a strong correlation ($r = .83$) between two factors of paradigmatic and syntagmatic relations in a factor structure, which shows a moderate fit to the data (e.g., root mean square error of approximation $= 0.09$).

Her study is valuable because a test with a modified format—especially in a new population of test takers—requires new validation, even though some relevant evidence from previous studies can be utilized. While she obtained positive evidence in terms of internal consistency, moderate relationships with a reading measure, and a factor structure, more evidence or explanations will be necessary in terms of her rationale, for example, for changing the format and for the organization of all the test items, including why the test has more items in paradigmatic relations than syntagmatic ones (i.e., 7 vs. 17 items). Shin may also want to compare her results to those of previous studies that used the original WAT (e.g., Read, 1993; Schoonen & Verhallen, 2008). For example, Batty (2012) administered the original WAT with 145 items to 530 learners of English at a Japanese university. He reported that the best fitting model was a bifactor model wherein the primary factor underlying all the items was a vocabulary factor and wherein the two factors of synonym (paradigmatic) and collocate (syntagmatic) were uncorrelated. Shin's and Batty's differing models may suggest the effects on the factor structure of (a) a format change, (b) a change in the number of items, (c) and a change in different test-taker populations; herein lies the scope for future avenues of research.

4 ''I Don't Know'' Use and Guessing on the Bilingual Japanese VST: A Preliminary Report

McDonald and Asaba examined the effects of including the "I don't know" option on test scores and analyzed how test takers select their responses when they are unsure of the answers. They used the bilingual version of the VST with up to 14,000-word family levels, with a stem in L2 English, four options in L1 Japanese, and the extra "I don't know" option. McDonald and Asaba interviewed four of 308 Japanese university students who took the VST. They reported an increased use of the "I don't know" option in lower frequency levels. They also noticed that when test takers do not know the answers, they make both uninformed and informed guesses, the latter of which are based on "true partial knowledge," "false partial knowledge," or "test strategy use." The students used all the types of uninformed and informed guesses, except in the case of one higher proficiency student who did not use uninformed guesses or guesses based on false partial knowledge. McDonald and Asaba computed the scores based on five scenarios: (a) "scores without guesses," (b) "scores with true partial knowledge-informed guesses," (c) "scores with all partial knowledge-informed guesses," (d) "scores with all informed guesses," and (e) "scores with all guesses." They showed that the scores varied substantially from (a) to (e), especially in the case of the lower proficiency students. For example, the lowest proficiency student had a difference of 30 points, which can be translated into a difference of 3,100-word family estimates. Of the five types of scores, McDonald and Asaba argued that scores with true partial knowledge-informed guesses (i.e., b) were most relevant to the construct. Therefore, the VST's overestimation of the test-takers' vocabulary size could threaten validity, and because of this, we should interpret the results with caution.

McDonald and Asaba's qualitative investigation using the bilingual version of the VST is very beneficial in terms of understanding students' test-taking processes and the drawbacks of the multiple-choice format. Based on the abovementioned five types of scores, one interesting direction for further research would be to devise strategies for producing confidence intervals (CIs), that is, the range of vocabulary size estimates that a single test taker can obtain. When using scores without guesses (i.e., a), the lower CI limit would indicate the minimum vocabulary size estimate when the VST includes the "I don't know" option, or if the test instructions urge test takers to avoid making random guesses, or if the instructions warn students that they will be penalized for selecting an incorrect option. The upper CI limit would indicate the maximum vocabulary size estimate obtained when test takers are encouraged to make guesses (i.e., e). These CIs would include scores with true partial knowledge-informed guesses (i.e., b), which will be gained through knowledge of the target word as well as guesses based on correct partial knowledge. A two-stage testing style may be needed to derive CIs for a paper-based test format, without conducting interviews: The first stage should contain the "I don't know" option or an instruction discouraging test takers from guessing; in the second stage, test takers should be encouraged to guess. The use of computers would enable easier administration wherein, in the second stage, only the items for which test takers selected "I don't know" appear. The presentation of CIs pertaining to vocabulary size estimates will improve test-users' understanding of their

overestimation risks and discourage them to take the results at face value. However, I admit that these procedures should be undertaken after consideration of the numerous other factors that come into play, and they require careful investigation before they can be implemented. Examples of the other factors that come into play are individual differences in whether or not test takers select "I don't know," the nature of the test instructions, and the stakes of the test (see Zhang, 2013, for details). Further, since guessing is not discouraged in the original VST (Nation, 2012), the use of the abovementioned procedures might alter the originally intended construct of the VST.

5 An Empirical Examination of the Effect of Guessing on VST Scores

McLean, Kramer, and Stewart examined the degree to which test takers guess before selecting answers in the VST. They administered the 1,000- to 8,000-word family frequency levels of the original VST to 3,373 students at Japanese universities, who had a wide range of *Hensachi* or T-scores. After excluding items with low discrimination, they used the Rasch and three-parameter logistic (3PL) models to investigate the effects of guessing. They showed that even low-ability learners guessed approximately 25% of the items correctly, and they obtained 41.7–85.5% of the correct scores in the 8,000-word family frequency level by using guessing strategies. They also reported a better fit of the 3PL model compared to the Rasch model; considering this, I wonder why they did not compare the two-parameter logistic (2PL) model with the 3PL model, because a gradual increase in parameters from the Rasch to the 2PL model and from the 2PL to the 3PL model (rather than from the Rasch to the 3PL model) would indicate differences in model fit more clearly and enhance our understanding of the underlying processes of guessing.

Similar to McDonald and Asaba, McLean et al. focused on the effects of guessing on the overestimation of vocabulary size, but from a quantitative perspective. Although the two studies used different versions of the VST, respectively—bilingual and monolingual—the results provide evidence that is rather negative: the VST has consistent overestimation issues.

Similar to the case with McDonald and Asaba, a possible direction for further study would be to devise methods of obtaining CIs by taking into account guessing percentages. Another direction could be an empirical examination into how many levels beyond test-takers' existing vocabulary levels are acceptable for test takers to take the VST, using the current data. McLean et al. showed that in the 8,000-word family frequency level, even the highest proficiency group (with a *Hensachi* score of 73) used guessing strategies at as high as 41.7%. For practical purposes, they could conduct similar analyses in the 1,000- to 7,000-word levels according to test-takers' size estimate groups as well as *Hensachi* groups, possibly by using different full marks across levels although the guessing percentages derived may not be very precise because McLean et al. omitted some 6,000 and 7,000 level items due to the low discrimination. If they set a permissible percentage for guessing (e.g., 25% or 30%), they could decide whether or not to endorse

Table 2. Evidence Provided by the Four Studies in Relation to the Interpretive Argument

Inference/Test	NGSLT (Stoeckel and Bennett)	Modified WAT (Shin)	Bilingual version of the VST (McDonald and Asaba)	Original VST (McLean et al.)
F. Utilization	*	*	*	*
E. Extrapolation	1. *r* with TOEIC	*	*	1. [B] Differences between proficiency groups
D. Explanation	*	2. Factor structure; 4. *r* with a reading test	5. [N] Large effects of guessing on test processes and scores	5. [N] Large effects of guessing on test scores 2. Item dimensionality; 3. [B] Increase in difficulty level
C. Generalization	1. Reliability (2. Creating parallel forms, bilingual versions, and an item bank)	1. Reliability	*	1. [B] Reliability; measurement invariance
B. Evaluation	1. Item quality	1. Item discrimination	*	1. [B] Item quality
A. Domain Definition	1. Content representativeness supported by the use of NGWL	*	1. Content representativeness supported by the translation of VST	1. [B] Content representativeness

Note. () = Mentioned as future research in the paper. * = Urgent need for examination. [B] = Evidence from Beglar (2010). [N] = Negative evidence. This table assumes an interpretive argument structure wherein the four studies intend to use the vocabulary tests for pedagogical purposes; however, other simplified structures such as those without the Utilization inference are also possible.

Beglar's (2010) suggestion that learners should take the VST with up to two levels higher than their vocabulary size.

6 Discussion and Conclusion

The four qualitative and quantitative studies reviewed here deserve substantial credit for their contributions to our existing knowledge on L2 vocabulary assessment studies and for providing positive and negative evidence for test validation. One issue for improvement is to clearly show how their studies fill the research gaps that exist in previous ones. This would further clarify the significance of their papers and substantially enhance contributions to the field.

Table 2 summarizes evidence, as reported by each paper, as well as puts forth areas for future examination, according to each test. I have also added evidence derived from Beglar (2010) because of its comprehensive coverage of the various aspects of validity. An examination of validation involves the entire process of examining test scores, interpretations, and uses from various perspectives; consequently, this task is often beyond the scope of a single study. Therefore, we need to accumulate evidence in relation to a validation framework so that other studies can refer to the evidence in the framework and observe the progress of validation while considering their individual research contexts. While I believe that L2 vocabulary researchers will benefit from using the argument-based approach to validity as a frame of reference, they may feel overwhelmed by the numerous detailed processes involved. However, this task can be shared with language assessment experts, giving rise to mutually beneficial research and practices (Koizumi, 2015a).

Acknowledgements

I am deeply indebted to Alastair Graham-Marr of Abax Publishing for his assistance and Yo In'nami for his invaluable comments. This work was partially supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research (C) [grant number 26370737].

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA/APA/NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford, U.K.: Oxford University Press.

Batty, A. (2012). Identifying dimensions of vocabulary knowledge in the Word Associates Test. *Vocabulary Learning and Instruction*, *1*, 70–77. Retrieved from http://vli-journal.org/wp/vli-v01-1-2187-2759/

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*, 101–118. doi:10.1177/0265532209340194

Browne, C. (2014). A New General Service List: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction*, 3(2), 1–10. Retrieved from http://dx.doi.org/10.7820/vli.v03.1.browne

Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing*, *27*, 443–470. doi:10.1177/0265532210367633

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Diagnostic assessment with automated writing evaluation: A look at validity arguments for new classroom assessments. *Language Testing*, *32*, 385–405. doi:10.1177/0265532214565386

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3–13. doi:10.1111/j.1745-3992.2009.00165.x

Chapelle, C., & Lee, H.-W. (2013, July). *What is argument-based validation?* Paper presented at the 35th Language Testing Research Colloquium, KCCI (Korea Chamber of Commerce and Industry) Building, Seoul, Korea.

Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In A. Kunnan (Ed.), *The companion to language assessment* (Vol. III, pp. 1079–1097). Hoboken, NJ: Wiley-Blackwell.

Doe, C. (2015). Student interpretations of diagnostic feedback. *Language Assessment Quarterly*, *12*, *110–135*. doi:10.1080/15434303.2014.1002925

Ginther, A., Yan, X., & Potts, J. (2015, March). *The relationship between TOEFL and GPA: The case of Chinese students*. Paper presented at the 37th Language Testing Research Colloquium, Eaton Chelsea Toronto, ON, Canada.

Huhta, A., Alderson, J. C., Nieminen, L., & Ullakonoja, R. (2015, March). *Diagnostic profiling of foreign language readers and writers—Exploring the usefulness of Latent Profile Analysis in diagnostic assessment research*. Paper presented at the 37th Language Testing Research Colloquium, Eaton Chelsea Toronto, ON, Canada.

Jang, E. E., Dunlop, M., Park, G., & van der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing*, *32*, 359–383. doi:10.1177/0265532215570924

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *12*, 527–535. Retrieved from http://dx.doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. doi:10.1111/jedm.12000

Koizumi, R. (2015a). Book review: Vocabulary knowledge: Human ratings and automated measures. *Language Testing*, *32*, 124–126. doi:10.1177/0265532214541233

Koizumi, R. (2015b). Factor structure and four-skill profiles of the TOEIC® tests among Japanese university learners of English. *ARELE (Annual Review of English Language Education in Japan)*, *26*, 109–124.

Koizumi, R.,Sakai, H., Ido, T., Ota, H., Hayama, M., Sato, M., & Nemoto, A. (2011). Toward validity argument for test interpretation and use based on scores of a diagnostic grammar test for Japanese learners of English. *Japanese Journal for Research on Testing*, *7*, 99–119. Retrieved from http://www7b.biglobe.ne.jp/~koizumi/Koizumi_research.html

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: National Council on Measurement in Education/American Council on Education.

Nation, I. S. P. (2012). *The Vocabulary Size Test*. Retrieved from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Thousand Oaks, CA: Sage.

Ota, H., Kanatani, K., Kosuge, A., & Hidai, S. (2003). *Eigo ryoku wa donoyoni nobite yukuka* [How English ability is developed: Exploring English acquisition process of junior high school students]. Tokyo: Taishukan.

Purpura, J. E., Brown, J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics. *Language Learning*, *65*(Suppl. 1, Special issue: Currents in language learning series: Improving and extending quantitative reasoning in second language research), 37–75. doi:10.1111/lang.12112

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

Sawaki, Y., & Koizumi, R. (2015, March). *Japanese students' and teachers' perception and use of score reports for two large-scale EFL tests*. Paper presented at the 37th Language Testing Research Colloquium, Eaton Chelsea Toronto, ON, Canada.

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

Stoeckel, T., & Bennett, P. (n.d.). *The New General Service List Test (NGLST)*. Retrieved from http://www.newgeneralservicelist.org/ngsl-levels-test/

Zhang, X. (2013). The *I don't know* option in the Vocabulary Size Test. *TESOL Quarterly*, *47*, 790–811. doi:10.1002/tesq.98

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

Article

An Investigation of Different Text Levels on L2 Learners' Vocabulary Learning Rates in an Extensive Reading Program

Anna C-S Chang
Hsing-Wu University
doi: http://dx.doi.org/10.7820/vli.v04.1.chang

Abstract

This study investigated whether different text levels would affect L2 learners' vocabulary learning rates and further examined the correlation between the frequency of word occurrence and learning rates in two different task conditions. A group of 31 year-11 senior high-school students read five level 1 graded readers, a total of 28,796 words, and then moved on to read five level 3 graded readers, 54,676 words, during a 13-week period. One hundred and twenty-five target words were selected from the 10 graded readers, 51 words from level 1, and 74 from level 2 texts. A vocabulary test containing two test methods was administered to students: a 125-item form-meaning matching test and a 42 item contextualized translation test. Twenty-one low-frequency words from each level text were selected to examine the frequency of word occurrence and learning rates. The meaning matching test results demonstrated that the learning rate at level 1 was significantly higher than those of level 3 in the post-test. The attrition rate of level 1 texts was also higher, which led to no significant difference in learning rates in the delayed post-test. A similar learning pattern can be said for the contextualized translation tests. Positive but statistically insignificant correlations between the frequencies of occurrence of 21 target words were found in the post-tests but were stronger in the delayed post-test in both meaning matching and translation tests. Overall, the learning rates for both level 1 and level 3 texts were very high. Four reasons for the higher learning rates were explained. Pedagogical implications are discussed.

1 Background

1.1 Learning Vocabulary through Extensive Reading

The benefits of reading have been well documented in both L1 and L2 learning, one of them being gaining vocabulary knowledge. Relatively low learning rates were reported by many previous studies that examined the extent of words incidentally learned through reading a single passage or a book (e.g., Day, Omura, & Hiramatsu, 1991; Dupuy & Krashen, 1993; Horst, Cobb, & Meara, 1998; Pellicer-Sánchez & Schmitt, 2010). Learning rates, however, were found much higher through extensive reading. For example, Cho and Krashen (1994) found that the vocabulary learning gains of four L2 adults who read 8–23

Please cite this article as: Chang, A.C-S (2015). A Japanese word association database of English. *Vocabulary Learning and Instruction*, *4* (1), 47–57. doi: 10.7820/vli.v04.1.chang

books over four months ranged from 56% to 80% of tested vocabulary. Horst's (2005) study with 17 ESL learners who read over a six-week period showed that the acquisition rate for the 1,001–2,000 level words was 76% and 62% the off-list words, with an average of 69%. In an in-depth case study of one French language learner who read four graded readers over one month and was tested with 133 words, Pigada and Schmitt (2006) found 66 words (50%) were improved in one type of word knowledge, 13 (10%) in two types, and 8 (6%) in three types, with a pick-up rate of 65%. Learners of different prior knowledge before reading were found to have a different learning rate. In their two-term study with 60 Taiwanese students, Webb and Chang (2015b) found that students' learning rates were 63.18%, 44.64%, and 28.12% for the high-, intermediate-, and low-level students, respectively, and the rates were quite comparable within each group during the two-term period. A few reasons may account for the higher word learning rates through extensive reading. The first one is that unknown or partially known words may be repeatedly encountered in the texts, and the more they read, the more opportunities they have to encounter the same words (Webb & Chang, 2015b). Second, in reading multiple books, it is not possible to prevent learners from using dictionaries to look up unknown words or to ask peers or teachers the meanings of unknown words. This course of action increases the attention for involvement and observation, which, in turn, may result in higher learning rates. Third, the more one reads, the better that person's reading skills become and in turn leads to higher comprehension level and higher word learning rates (Nation & Wang, 1999). These above reasons have made extensive reading a valuable source for developing L2 vocabulary knowledge.

Despite these extensive reading studies having demonstrated impressive learning gains, the quantity in this line of research is still very limited because most studies have focused on general vocabulary knowledge improvement. Thus, more research into this area is needed to provide more empirical evidence. Moreover, the first three studies mentioned above looked at the effectiveness of L2 learners reading self-selected books on vocabulary learning gains, which limited our understanding of whether text levels had connection to the learning gains. This issue seems to have been partly solved by more recent research by Webb and Chang (2015b) who examined EFL learners studying two levels of texts (level 1 and level 2); and students' learning gains were found quite similar between level 1 and level 2. Theoretically, reading an easier text should enable a student to acquire more words than reading a more difficult one because easier texts tend to lead to higher level of comprehension, so readers could give more attention to unknown words (Pulido, 2004). To provide further evidence in this area, this study intended to compare whether a group of students who read easier texts (level 1) would gain more than they would by reading more difficult texts (level 3).

A few studies have examined the correlation between learning the form-meaning connection and the number of times each word occurred in reading a single text, and a moderate correlation has been found. For example, Saragi, Nation, and Meister (1978) found a correlation of .34 and Vidal (2011) identified a correlation of .47. Horst et al. (1998) reported a correlation of .49. Webb (2007) found a correlation of .43 for recall meaning and .23 for recognition of meaning. Not until very recently, though, had studies ever examined the effects of repetition

in multiple texts. Webb and Chang (2015a) examined how frequency and distribution of occurrence affect learning rates with 61 Taiwanese students who studied 10 level 1 books. The correlation between relative gain [(the number of words gained from learning/unknown words) × 100] in the post-test and the frequency of occurrence was found to be negatively low and non-significant ($r = -.03$, $p = .78$, $n = 100$). A possible reason could be that mixed levels of target words were randomly selected from the 10 different graded readers. This study, therefore, would further explore the correlations between these two variables but focus on low-frequency words only. To evaluate more accurately learners' receptive and productive knowledge, two test methods were used: the form-meaning matching test and the contextualized translation test. Two research questions are addressed:

RQ1: To what extent would different text levels affect L2 learners' vocabulary learning rates from reading graded readers?

RQ 2: What were the correlations between frequency of word occurrence and relative gains in different text levels and in different test methods?

2 The Study

2.1 The Participants

A group of 42 year-11 senior high-school students took part in the extensive reading program; however, only 31 of them consistently attended the program throughout. These students had engaged in extensive reading for a year before the intervention started. Their vocabulary knowledge assessed through the bilingual Vocabulary Levels Test (Schmitt, Schmitt, & Clapham, 2001) were 26/30 and 22/30 for the first and second 1,000 words.[1]

2.2 The Study Materials

Five level 1 and five level 3 graded readers were selected from the Bookworm series and the Macmillan series (see Appendix A for the details). The total word count of the level 1 text was 28,796 and 54,676 for level 3 texts. The lexical profile of the two level texts is shown as below.

Table 1. Lexical Profile of Level 1 and Level 3 Graded Readers

	Tokens (%)			
Word list	Level 1	Cumulative coverage + proper nouns	Level 3	Cumulative coverage + proper nouns
1,000	24,778 (86.05%)	90.92%	46,926 (85.83%)	89.49%
2,000	1,616 (5.61%)	96.53%	3,108 (5.61%)	95.17%
Proper nouns	1,402 (4.87%)		2,010 (3.68%)	
Total	28,796		54,676	

2.3 The Target Words and Dependent Measures

A total of 125 target words were selected to be tested, among which 51 words were from the level 1 texts and 74 words from the level 3 texts. The target words were analyzed with GSL range (Nation & Heatley, 2002) and Nation's (2012) BNC/COCA range. Approximately, 35% of the target words from level 1 texts and 55% from level 3 texts were in the lists of words beyond the 3,000 level.

The target words were measured through two test methods: a form-meaning matching test and a contextualized translation test. The former contained 125 target words divided into 10 blocks. Each block had 8–15 target words. The order of the items was rearranged in the immediate post-tests and the delayed post-tests. An example from a pretest is shown as below.

1._____ armour	a. 正義
2._____ judge	b. 請願
3._____ justice	c. 法律
4._____ petition	d. 盔甲
5._____ gang	e. 法官
6._____ creek	f. 獎賞
7._____ outlaw	g. 子彈
8._____ bullet	h. 幫派
9._____ reward	i. 歹徒
	j. 小溪

Figure 1. Form-Meaning Matching Test Example.

The contextualized translation tests involved 42 low-frequency words (21 from each text level) that were selected from the 125 target words. An example taken from *Christmas in Prague* is given below. The students had to translate the underlined word into Chinese.

> Carol is never late for rehearsals, and she knows that these concerts are important for us. (演練).

The correlations between frequency of word occurrence and learning rates: The frequency of the 21 off-list words from each text level was examined, and then the learning gains in the two test methods were computed. The frequency of the 42 words is shown in Appendix B.

2.4 Procedure

The two forms of vocabulary pretests were administered to the participants before the intervention. The translation test was taken first, then the form-meaning matching test. After the pretest, the students read one level 1 text each week, followed by an immediate post-test at week 6, and a three-month delayed post-test at week 17. The same procedure was repeated after students had read the level 3 texts. The test procedure is presented below. The level 3 texts were longer than those of level 1, so it took the participants eight weeks to finish reading the five readers.

1st week	Pretest	51 level 1 words
6th week	Post-test	51 level 1 words
6th week	Pretest	74 level 3 words
14th week	Post-test	74 level 3 words
17th week	Delayed post-test	51 level 1 words
25th week	Delayed post-test	74 level 3 words

Figure 2. Test Procedure.

To ensure that all participants did read the texts, all reading was completed in their weekly three-hour reading class. Before students started to read, the instructor usually introduced the story or the author's background. While students were reading, they also listened to the oral rendition of texts. The instructor would stop the audio CD from time to time to check students' comprehension or to allow students to ask questions. After finishing a story, the students could take notes, ask questions, or discuss the story with their classmates. All the after-reading activities were voluntary and allowed students to have some time to think about what was read, to check up unknown words in the dictionary, or to write personal feelings about the story. Before the students returned each book, a simple comprehension quiz was given.

3 Results

Table 2 shows that before the treatment, the participants were familiar with approximately a half of the target words in the level 1 texts and slightly more than one third in level 3 in the form-meaning matching test. In the contextualized translation test (see Table 3), the participants also knew about a half of the level

Table 2. Descriptive Statistics of Form-meaning Matching Test across Three Occasions

Text		Mean	SD	Minimum	Maximum
Level 1 (51 words)	Pretest	25.94 (51%)	6.20	13	39
	Post-test	47.68 (93%)	5.03	28	51
	Delayed post-test	45.39 (89%)	5.64	31	51
Level 3 (74 words)	Pretest	27.55 (37%)	9.91	10	47
	Post-test	65.48 (87%)	10.07	38	75
	Delayed post-test	62.19 (83%)	10.44	40	75

Table 3. Descriptive Statistics of Contextualized Translation Test across Three Occasions

		Mean	SD	Minimum	Maximum
Level 1 (21 words)	Pretest	10.35 (49%)	3.98	3	18
	Post-test	20.16 (96%)	1.29	17	21
	Delayed post-test	18.90 (90%)	3.03	9	21
Level 3 (21 words)	Pretest	1.81 (9%)	1.60	0	6
	Post-test	18.13 (86%)	2.67	13	21
	Delayed post-test	17.03 (81%)	3.61	10	21

Table 4. Descriptive Statistics for Different Text Levels and Learning Rates in the Form-meaning Matching Tests and Translation Tests ($n = 31$)

	Level 1		Level 2	
	Relativegain1	Relativegain2	Relativegain1	Relativegain2
Form-meaning matching tests				
Mean	87.07 (18.75)	76.74 (24.42)	79.33 (21.47)	71.43 (26.66)
Min–max	17.86 to 100	0 to 100	28.85 to 100	−14.29 to 100
Translation tests				
Mean	93.22 (10.37)	81.92 (21.69)	85.51 (13.30)	79.82 (18.10)
Min–max	71.43–100	33.33–100	60–100	44.43–100

Note: Relative gain 1: From the pretest to the post-test; relative gain 2: From the pretest to the delayed post-test.
The numbers in the parenthesis are standard deviations.

1 target words; however, they knew only 9% of the words for the level 3 texts (1.81/21). This meant that the participants would have many more words to learn from level 3 texts than from the level 1 texts.

RQ1: To what degree would different text levels affect L2 learners' vocabulary learning rates from reading graded readers?

The relative gains presented in Table 4. The results of form-meaning matching test show that the relative gains were very high for both level 1 and level 3 texts. There was a significant difference in relative gain 1 between level 1 texts and level 3 texts, $t(30) = 2.67$, $p = .012$; however, the effect size calculated with Cohen's d was small ($d = 0.38$). The relative gain 1 for the level 1 text ($M = 87.03$) was significantly higher than that of level 3 texts ($M = 79.33$). No significant difference was found in relative gain 2 between level 1 and level 3 texts, which suggested that the attrition of level 1 words learned was higher than that of level 3 (see Table 5).

Similar results can be said for the contextualized translation tests. There was a statistically significant difference in relative gain 1, $t(30) = 3.15$, $p = .004$, and the effect size was medium (Cohen's $d = 0.65$). No marked difference was found in relative gain 2, $t(30) = .49$, $p = .63$ (see Table 5). These above results provide answers to the first research question.

RQ 2: What were the correlations between frequency of word occurrence and gains in different text levels and in different test methods?

In the form-meaning matching tests, the correlations between frequency of word occurrence and relative gain in the level 1 texts were found to be positive but insignificant ($r = .04$, $p = .88$) The same results can be said for the level 3 texts ($r = .11$, $p = .65$); however, the correlations seem to be stronger in the delayed post-test than in the immediate post-test in the both level 1 ($r = .15$, $p = .51$) and level 3 texts ($r = .24$, $p = .30$; see Table 6).

In the contextualized translation tests, the correlations were positive in both level 1 ($r = .16$, $p = .49$) and level 3 ($r = .22$, $p = .35$) texts and in both the immediate post-test and the delayed post-test. The correlations were found to be comparable for the level 1 texts ($r = .20$, $p = .38$) than for the level 3 texts ($r = 18$, $p = .45$). The answer to the second research question was that there were positive but statistically insignificant correlations between frequency of word occurrence

Table 5. Paired-Sample *t*-Tests for Relative Gains of Vocabulary Learning Rates between Level 1 and Level 3 Texts on the Meaning Matching and Translation Tests

	Mean	SD	SEM	95% CID	*t*	df	*p*	Effect size (Cohen's *d*)
Form-meaning matching tests								
Relative gain 1	7.74	16.16	2.90	1.81 to 13.67	2.67	30	.012	0.38
Relative gain 2	5.31	25.98	4.67	−4.23 to 14.83	1.14	30	.27	0.21
Translation tests								
Relative gain 1	7.70	13.61	2.44	2.71 to 12.70	3.15	30	.004	0.65
Relative gain 2	2.10	23.84	4.28	−6.64 to 10.85	.49	30	.63	0.11

Note: Relative gain 1: From the pretest to the post-test; relative gain 2: From the pretest to the delayed post-test.

Table 6. Correlations between Frequency of Word Occurrence and Gains in Different Test Methods and Different Text Levels

Test methods	Text level	Relative gain 1	Relative gain 2
Form-meaning matching	1	.04	.16
	3	.11	.23
Contextualized translation	1	.16	.20
	3	.22	.18

Note: Relative gain 1: From the pretest to the post-test; relative gain 2: From the pretest to the delayed post-test.

and relative gains. The correlations were slightly stronger in the delayed post-tests in the form-meaning matching tests but were mixed in the contextualized translation tests for different level texts.

4 Discussion and Conclusion

The learning gains of the present study were slightly higher than those of previous studies. Four factors may explain the results: reading amount, motivation, comprehension check, and fondness for the selected stories. As to the amount of reading, these 31 students had been reading for a year before this study began. Their reading skills, therefore, might be competent and could be accustomed to reading immediately after the intervention. Second, most of the students were highly motivated learners. They took notes on what they learned from the graded readers although the instructor did not require them to do so. The third reason might be the comprehension check activity after reading. As per the students' request, the instructor took one of the after-reading activities from the worksheets provided by the publishers, the one most often used being *putting the events in order*. This comprehension check activity might have caused students to read more carefully and comprehend more. According to Stoeckel, Reagan, and Hann (2012), students who took simple quizzes felt slightly greater intellectual benefit from the reading task. Finally, in the students' interview reports, all reported that words learned from stories were easily remembered, and the better they liked a story, the more words they learned and the longer they retained the words. That these stories had been read by previous students and rated as very interesting could be one very important factor explaining the high learning gains.

To the correlations between the frequency of word occurrence and relative gains, the present study did not demonstrate significant correlations in different text levels nor did it in test methods. The main reason could be that exposure frequency is only one of many factors affecting the extent of word knowledge being learned from reading. Other factors, such as word context, elaboration of word processing, or types of word knowledge being assessed, may all come into play and influence the word frequency needed to learn knowledge of a word.

The present study consistently showed that the relative gains were statistically significant only in the immediate post-test. The results have an important implication on learning vocabulary through reading. Learning gains start to decay immediately after learning. Whatever level texts are studied or test methods are used would make no difference to learning gains after a three-month period. This suggests that learners need to continue reading so as to have opportunities for repeated exposure. This supports some scholars' suggestion that reading intervals between texts should be short. Nation and Wang (1999) suggest that learners should read at least one graded reader per week, and Prowse (2002) suggests more than one per week. How L2 learners acquire word knowledge from extensive reading is still under-researched. Future research in this area is certainly warranted.

Note

1. The 1st 1,000 word level was developed by Professor Paul Nation of Victoria University of Wellington, New Zealand.

References

Cho, K.-S., & Krashen, S. D. (1994). Acquisition of vocabulary from the Sweet Valley Kids series: Adult ESL acquisition. *Journal of Reading*, *37*, 662–667.

Day, R. R., Omura, C., & Hiramatsu, M. (1991). Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language*, *7*, 541–551. http://nflrc.hawaii.edu/rfl/PastIssues/rfl72day

Dupuy, B., & Krashen, S. D. (1993). Incidental vocabulary acquisition in French as a foreign language. *Applied Language Learning*, *4*, 55–63.

Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *Canadian Modern Language Review*, *61*, 355–382. doi:10.3138/cmlr.61.3.355

Horst, M., Cobb, T., & Meara, P. (1998). Beyond A Clockwork Orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*, 207–223. http://nflrc.hawaii.edu/rfl/PastIssues/rfl112horst

Nation, I. S. P. (2012). The BNC/COCA word family lists. Unpublished manuscript. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts [software]. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Nation, I. S. P., & Wang, K. M.-T. (1999). Graded readers and vocabulary. *Reading in a Foreign Language*, *12*, 355–380. http://nflrc.hawaii.edu/rfl/PastIssues/rfl122nation

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do Things Fall Apart? *Reading in a Foreign Language*, *22*, 31–55. http://nflrc.hawaii.edu/rfl/April2010/articles/pellicersanchez

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, *18*(1), 1–28. http://nflrc.hawaii.edu/rfl/April2006/pigada/pigada

Prowse, P. (2002). Top ten principles for teaching extensive reading: A response. *Reading in Foreign Language Journal*, *14*(2), 142–145. http://nflrc.hawaii.edu/rfl/October2002/discussion/prowse

Pulido, D. (2004). The relationship between text comprehension and second language incidental vocabulary acquisition: A matter of topic familiarity? *Language Learning*, *54*, 469–523. doi:10.1111/j.0023-8333.2004.00263.x

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*, *6*(2), 72–78. doi:10.1016/0346-251X(78)90027-1

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviours of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*, 55–88. doi:10.1177/026553220101800103

Stoeckel, T., Reagan, N., & Hann, F. (2012). Extensive reading quizzes and reading attitudes. *TESOL Quarterly*, *46*, 187–198. doi:10.1002/tesq.10

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, *61*, 219–258. doi:10.1111/j.1467-9922.2010.00593.x

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. doi:10.1093/applin/aml048

Webb, S., & Chang, C.-S. (2015a). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research*. Advance online publication. doi:10.1177/1362168814559800

Webb, S., & Chang, C.-S. (2015b). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *SSLA*. Advance online publication. doi:10.1017/S0272263114000606

Appendix A: Study Materials and Word Count

Text level	Word count
Level 1	
The President's Murderer	5,270
White Death	6,600
Ned Kelly	5,775
Sherlock Holmes and Duke's Son	5,800
Christmas in Prague	4,720
Level 3	
Chemical Secret	10,149
The Call of the Wild	11,045
The Hound of the Baskervilles	8,878
The Woman in Black	11,003
The Picture of Dorian Gray	12,975

Appendix B: 42 Low-Frequency Words and Their Frequency of Occurrence

	Five level 1 books	Frequency	Five level 3 books	Frequency
1	ambulance	3	certificate	7
2	armour	8	biologist	8
3	bullet	7	growl	9
4	gang	35	marsh	68
5	courtroom	12	mist	23
6	drug	16	downstream	7
7	duchess	7	portrait	26
8	guardhouse	1	seal (n)	24
9	heroin	44	sewage	5
10	hostage	12	solicitor	8
11	innocent	5	torch	9
12	jury	19	wrinkle	10
13	kidnapper	3	journalist	4
14	orchestra	7	gallery	8
15	outlaw	11	chemist	5
16	rehearsal	4	corridor	2
17	twin	2	hound	14
18	tyre	13	telescope	8
19	conductor	8	moose	6
20	harpist	6	gravestone	7
21	petition	3	cement	3

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

Article

Measuring Knowledge of Words with Multiple Meanings

Yuko Hoshino
Tokyo Fuji University
doi: http://dx.doi.org/10.7820/vli.v04.1.hoshino

Abstract

When reading or listening to English, we encounter many words, most of which are high-frequency, polysemous words. Due to their polysemous nature, not knowing one particular meaning of a high-frequency word may prevent learners from understanding the wider contexts. Therefore, it is necessary to know whether learners have knowledge of multiple meanings of one word; however, there is no universally accepted way to measure knowledge of words with multiple meanings. With this in mind, the current paper looks at the difficulties of measuring learners' knowledge of polysemous words and describes ongoing research on this topic.

1 Introduction

When reading or listening to English we encounter many words, most of which are high-frequency words. For example, Adolphs and Schmitt (2003) found that 209 word families cover more than 80% of spoken vocabulary and 1000 individual word forms cover 89.25% of the spoken British National Corpus. In addition, when we turn our attention to written texts, which generally include more low-frequency words than spoken texts, Nation (2006) summarized that the most frequent 1000 word families cover 78–81% of the total word count. Hence, high-frequency words are important for understanding English.

It is also generally understood that vocabulary knowledge correlates with comprehension of texts, as shown by Jeon and Yamashita's (2014) meta-analysis of 31 studies which yielded a correlation coefficient of $r = .79$ [.69–86]. This relationship has also been presented by investigations looking at the connection between reading comprehension and vocabulary breadth (e.g., Iso & Aizawa, 2010; Laufer & Ravenhorst-Kalovski, 2010), which refers to how many words learners know (i.e., knowing many low-frequency words means they are better learners). However, the lexical coverage of low-frequency words is much lower than that of high-frequency words. Thus, it is possible that knowing high-frequency words is equally as important as knowing low-frequency words when it comes to comprehension of texts. High-frequency words often have more than one meaning, so not knowing one particular meaning of a high-frequency word possibly prevents learners from understanding the text. In order to investigate this issue, it is necessary to know whether learners have knowledge of multiple meanings of one word, however, there is no universal way to measure knowledge of words with

Please cite this article as: Hoshino, Y. (2015). Measuring knowledge of words with multiple meanings. *Vocabulary Learning and Instruction*, *4* (1), 58–65. doi: 10.7820/vli.v04.1.hoshino

multiple meanings. With this in mind, the current paper describes ongoing research into measuring learners' knowledge of multiple meanings of one word.

2 Difficulty in Measuring Knowledge of Words with Multiple Meanings

The importance of knowing high-frequency words, which often have more than one meaning, is widely understood, but much of the existing research has focused on how effectively learners acquire low-frequent new words. Learning new words is different from learning new senses of old words. When learners first try to acquire a new word, both the word form and its meaning are new, but when they learn a new meaning of a known word, "people add to their already extensive repertoires of words for which they know multiple different senses" (Rodd, Gaskell, & Marslen-Wilson, 2002, p. 1096). Thus, learning a new meaning of a previously known word requires restructuring of one's existing mental lexicon that contains the word. This presents a number of difficulties when attempting to determine whether learners have succeeded in acquiring new meanings or not.

First is ambiguity resolution. When a word has more than one meaning, and there is a target meaning the researcher wants to measure, it is necessary to distinguish the target meaning from other meanings. One solution is to provide context. For example, Khanna and Cortese (2011), who investigated the Age of Acquisition (AoA) of ambiguous and polysemous words, presented the target word *duck* as follows:

duck

To bend down

By presenting the definition "to bend down," the researchers distinguished the target meaning from the meaning of *duck* as a bird. Degani and Tokowicz (2013) also provided minimum contexts for target words (e.g., *rose*: *red rose* and *rose above*). However, the problem remains as to whether these minimum contexts provide sufficient cues for the learners to use when their knowledge is being assessed. Also, when longer texts are presented in order to better resolve the ambiguity, the amount of information the text contains differs by context, which produces another problem.

The second difficulty pertains to grammatical knowledge. When a word's meaning changes, often its grammatical usage changes as well. For example: *time* can be used as a noun or as a verb, both relating "what is measured in minutes, hours, days." Should we differentiate *time* as a noun from *time* as a verb when we want to measure knowledge of multiple meanings of this word? A word like *present* makes this problem more complicated. *Present* has two meanings: *existing or happening now* and *gift*. *Present* meaning "now" often takes the adjectival form; however, it can also be used as a noun. On the other hand, *present* meaning "gift" is often used as a noun but can also be used as a verb. This change of word class makes it difficult to separate grammatical knowledge and knowledge about multiple meanings. Moreover, with verbs tending to be more abstract and ambiguous than nouns (Crossley, Subtirelu, & Salsbury, 2013), it is especially

important to consider this difference between word classes. There are also cases where usage differs between two or more meanings within the same word class. For example, *admit* has two meanings (*enter* and *confess*), but *admit* meaning "enter" takes a direct object after the main verb whereas *admit* meaning "confess" takes a full clause (Uçkun, 2012). Therefore it seems contextual meaning and grammatical structure help determine the intended meaning of words with multiple meanings.

The third difficulty is the relationship between meanings. Words with multiple meanings can be divided into two types: polysemous and homonymous. The meanings of polysemous words are somewhat related to each other, whereas those of homonymous words are distinct (e.g., Rodd et al., 2002). Homonymy is more prominent because the meanings are usually completely different from each other. On the other hand, polysemy does not usually stand out. For example, some English textbooks for junior high schools in Japan have separate listings for new meanings of previously learned homonymous words. Also, they include notes about these words and try to draw learners' attention to them. However, as written above, textbook writers tend not to turn their attention to polysemy probably because of the semantic relatedness between meanings. Rodd et al.'s (2012) L1 study showed that when the novel meaning and the existing meaning are closely related, the recall of novel meanings is significantly better. Yet it is not yet known whether polysemy is more effectively acquired than homonymy for L2 learners. Even within polysemy, the distance between the meanings differs. One meaning within polysemy is the prototypical meaning, and the others are extended meanings, the degree of which differs. In other words, some meanings are rather strongly related, and learners can easily infer the extended meaning from the prototypical one. However, when their relationship is too strong, we might wonder whether these meanings are two distinct meanings or not. For example, Degani and Tokowicz (2013) asked participants to rate the similarity between the meanings of one word. They used a 7-point scale, and some meanings were rated as not similar (e.g., 1.15 for *red rose* and *rose above*), but others were similar (e.g., 5.30 for *cotton dress* and *cotton thread*). The dictionary entries for *cotton* treat these two meanings as separate, but

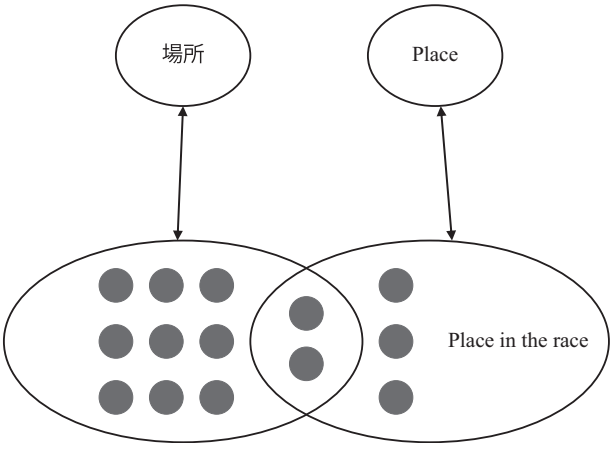

Figure 1. Sense Model (modified from Finkbeiner et al., 2004).

it is not yet known whether teachers should teach these meanings one by one or whether learners will automatically learn the extended meaning (*cotton thread*) when they learn the prototypical one (*cotton dress*).

Another problem related to meaning is cross-language differences. According to the Sense Model by Finkbeiner, Forster, Nicol, and Nakamura (2004), shown in Figure 1, several senses (small black circles in the figure) belong to one word each in the mental lexicon. Some of the senses are shared by two languages, and other senses are language-unique. For example, both *basho* in Japanese and *place* in English refer to "place as a space." Hence, *basho* and *place* are often paired and memorized in the beginning stage of learning, according to the equivalence hypothesis (Ringbom, 1987). However, pairing only one L1 word and one L2 word is insufficient because L1 and L2 concepts are not equal and some senses are peculiar to one language (Finkbeiner et al., 2004). For example, *place* has the concept of "place in the race," which has a different Japanese translation, *jun'i*. Therefore, when developing a test to measure knowledge of words with multiple meanings, it is desirable to measure the developmental change in learners' mental lexicon.

3 Summary of Hoshino (in Press)

Considering some of the abovementioned difficulties, Hoshino (in press) compared six types of test for multiple meanings of basic words to identify the best test from the viewpoints of difficulty, reliability, discriminability, and correlation coefficients. About half of the participants took tests that presented the target words in sentences, and the remaining participants took tests that presented the target words in collocations. Seven basic verbs (*break, get, have, leave, make, see,* and *take*) were the target words, and there were six types of measurement, as Table 1 shows.

In the selection test, participants chose the correct expressions of the target verbs out of eight possible answers. Table 2 shows the selection test in sentential context. The participants read the sentences, decided whether the underlined expressions were correct English expressions or not, and marked their judgment as correct (○) or incorrect (×) in the right column. The number of correct expressions was six for all the target verbs, but the participants were not notified how many correct answers there would be. The learners who were tested in collocation were presented only the underlined contexts.

Table 1. Six Types of Measurements in Hoshino (in press)

	Form (within-subject factor)	Context (between-subject factor)
1	Selection	Sentence
2	Translation	Sentence
3	Pairing	Sentence
4	Selection	Collocation
5	Translation	Collocation
6	Pairing	Collocation

Table 2. The Selection Test in Sentential Context in Hoshino (in press)

		○/×
1	He broke the secret to the public.	
2	They broke the law by not stopping at a red light.	
3	They broke the scene in their movies.	
4	He broke the end of the story.	
5	He broke the skin on his legs.	
6	She broke the news to me.	
7	The stone broke the surface of the water.	
8	I'm not going to break my promise to her.	

Table 3. The Translation Test in Sentential Context in Hoshino (in press)

		Write translation
1	He broke the secret to the public.	
2	They broke the law by not stopping at a red light.	
3	He broke the skin on his legs.	
4	She broke the news to me.	
5	The stone broke the surface of the water.	
6	I'm not going to break my promise to her.	

The translation test is presented in Table 3. The material was the same as the selection test except for the distractors (i.e., Nos. 3 and 4 in Table 2). Only the correct expressions were presented, and the participants were asked to write the translations of the underlined expressions. Similar to the selection test, only the underlined contexts were shown to the participants who were tested in the collocation condition.

The pairing test was conducted with the same target materials as the translation test. The participants were required to find three pairs of the target expressions with the same meaning. Each target verb had three pairs (i.e., Nos. 1 and 4, Nos. 2 and 6, and Nos. 3 and 5 have the same meaning in Table 3).

Results showed that of the six measurements, the translation test in sentential context had the highest reliability and discriminability. The translation test had significant correlation coefficients with the vocabulary size test and the reading test, but the selection test and the pairing test did not, also their correlation coefficients were weak or even negative. Between the two translation tests with different length of contexts (collocation or sentence), the test in sentential context had a higher correlation than the one in collocation. Considering that the construct measured by the vocabulary size test and the reading tests was at least somewhat related with the construct that the test for words with multiple meanings was designed to measure, the translation test in sentential context was the better test among the six tests in Hoshino (in press). However, questions remain. How long should the sentence be? What information should be included in it? Are there any other, better tests?

4 Other Possible Measurement Methods

Hoshino (in press) did not use the isolation test, an isolation test is one of the possible methods of measuring multiple meanings. An example is as shown below.

Write as many meanings of each word as possible.

1. present	(	)	(	)	(	)
2. take	(	)	(	)	(	)

This method is used especially when researchers want to make a list of polysemy meanings (e.g., Durkin & Manning, 1989; Tagashira, Sakata, Hoshino, & Mochizuki, 2014). Rodd et al. (2012) used a similar method. They presented the target words in isolation and asked the participants to write the properties of the novel meanings they had learned. When at least one of the properties they wrote was correct, the response was regarded as correct. However, producing as many meanings as possible without any cues is difficult even for advanced L2 users. For example, in Tagashira et al.'s study above, two English teachers produced answers to the 68 words with multiple meanings, but they answered at most five meanings per word. The average number of answers per word was only two or three, so it is difficult to grasp the development of the mental lexicon by this measure only. Furthermore, there is the problem of loan words. Japanese has a lot of loan words, and this becomes a problem when using this measure. For example, *program* has several meanings in English, and these meanings are often used in Japanese as プログラム [puroguramu]. However, プログラム has different meanings, as in a computer *program* or concert *program*. Hence, if the participants write only プログラム, we do not know which meaning of *program* they intend.

Another measure, which presents the target words in isolation, is the translation recognition test. In this test, researchers present the target word first and then its target meaning to the participants, and participants judge whether the second word is the first word's correct meaning or not as quickly and accurately as possible. The translation recognition test does not require participants to produce the translation, so it is more suitable for investigating the developmental stage of acquisition or beginning learners (Sunderman, 2014).

5 Conclusion

So far, several measurements have been introduced, but we still do not know which measurement is better than others and which should be used in which circumstances. Making a valid test for knowledge of multiple meanings will deepen our understanding of learners' mental lexicon, so the further research is necessary for this topic.

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*, 425–438. doi:10.1093/applin/24.4.425

Crossley, S. A., Subtirelu, N., & Salsbury, T. (2013). Frequency effects or context effects in second language word learning. *Studies in Second Language Acquisition*, *35*, 727–755. doi:10.1017/S0272263113000375

Degani, T., & Tokowicz, N. (2013). Cross-language influences: Translation status affects intraword sense relatedness. *Memory & Cognition*, *41*, 1046–1064. doi:10.3758/s13421-013-0322-9

Durkin, K., & Manning, J. (1989). Polysemy and the subjective lexicon: Semantic relatedness and the salience of intraword senses. *Journal of Psycholinguistic Research*, *18*, 577–612. doi:10.1007/BF01067161

Finkbeiner, M., Forster, K., Nicol, J., & Nakamura, K. (2004). The role of polysemy in masked semantic and translation priming. *Journal of Memory and Language*, *51*(1), 1–22. doi:10.1016/j.jml.2004.01.004

Hoshino, Y. (in press). Development of a test to measure knowledge regarding multiple meanings of basic words. *JLTA Journal*, *18*.

Iso, T., & Aizawa, K. (2010). Toward better interpretations of vocabulary size test scores for estimating reading proficiency. *Annual Review of English Language Education in Japan*, *21*, 141–150.

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*, 160–212. doi:10.1111/lang.12034

Khanna, M. M., & Cortese, M. J. (2011). Age of acquisition estimates for 1,208 ambiguous and polysemous words. *Behavior Research Methods*, *43*(1), 89–96. doi:10.3758/s13428-010-0027-y

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*, 15–30.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82. doi:10.3138/cmlr.63.1.59

Ringbom, H. (1987). *The role of first language in the second language learning*. Clevedon/Philadelphia; Multilingual Matters.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*, 245–266. doi:10.1006/jmla.2001.2810

Rodd, J. M., Berriman, R., Landau, M., Lee, T., Ho, C., Gaskell, M. G., & Davis, M. H. (2012). Learning new meanings for old words: Effects of semantic relatedness. *Memory & Cognition*, *40*, 1095–1108. doi:10.3758/s13421-012-0209-1

Sunderman, G. (2014). Translation recognition tasks. In J. Jegerski & B. van Patten (Eds.), *Research methods in second language psycholinguistics* (pp. 185–211). New York, NY: Routledge.

Tagashira, K., Sakata, N., Hoshino, Y., & Mochizuki, M. (2014, March). *Nihonjin EFL gakusyusha no tagigo ni taisuru honyakugo database [Translation database of polysemy by Japanese EFL learners]*. Paper presented at the Fourth Joint Conference on English Vocabulary and Lexicography Program. Chiba: Reitaku University.

Uçkun, B. (2012). Awareness of verb subcategorization probabilities with polysemous verbs: The second language situation. *System*, *40*, 360–375. doi:10.1016/j.system.2012.08.002

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

Article

Are Learners Aware of Effective Ways to Learn Second Language Vocabulary from Retrieval? Perceived Effects of Relative Spacing, Absolute Spacing, and Feedback Timing on Vocabulary Learning

Tatsuya Nakata
Faculty of Foreign Language Studies, Kansai University
doi: http://dx.doi.org/10.7820/vli.v04.1.nakata

Abstract

Although the effects of retrieval-based learning may be affected by factors such as relative spacing, absolute spacing, and feedback timing, few studies have examined learners' perceived effects of these factors on second language (L2) vocabulary learning. With this limitation in mind, the present study examined learners' perceived effects of the above three factors on L2 vocabulary learning. A questionnaire was administered to 226 Japanese college students. The results showed that the participants (1) considered relative spacing not to affect learning, (2) perceived spaced learning to be more effective than massed learning, and (3) considered immediate and delayed feedback to be equally effective, all of which are supported by earlier empirical research. The results suggest that L2 learners are able to make accurate judgments about the effectiveness of retrieval-based vocabulary learning techniques.

1 Introduction

Research shows that retrieval enhances second language (L2) lexical acquisition (e.g., Barcroft, 2007; Karpicke & Roediger, 2008). Retrieval refers to recalling information about L2 words from memory. Earlier studies also suggest that the effects of retrieval-based learning may be affected by factors such as relative spacing, absolute spacing, and feedback timing (see below for definitions). Relative spacing is defined as how retrieval attempts are distributed relative to one another (Karpicke & Bauernschmidt, 2011). Existing studies have examined the effects of two types of relative spacing schedules: equal and expanding. In the former, spacing between retrievals of a given word is held constant (e.g., 4, 4, and 4 days). In the latter, spacing between retrievals is gradually expanded (e.g., 1, 4, and 7 days). Some researchers argue that expanding spacing is more effective than equal spacing (e.g., Baddeley, 1997; Ellis, 1995; Hulstijn, 2001; Nation, 2001; Schmitt, 2007; see Roediger & Karpicke, 2010, for a review). Although some non-L2 vocabulary studies have found the advantage of expanding over equal spacing (e.g., Cull, Shaughnessy, & Zechmeister, 1996; Landauer & Bjork, 1978; Storm, Bjork, & Storm, 2010), L2 vocabulary studies have found little difference between the effects

Please cite this article as: Nakata, T. (2015). Are learners aware of effective ways to learn second language vocabulary from retrieval? Perceived effects of relative spacing, absolute spacing, and feedback timing on vocabulary learning. *Vocabulary Learning and Instruction*, *4* (1), 66–73. doi: 10.7820/vli.v04.1.nakata

of the two types of relative spacing schedules (Kang, Lindsey, Mozer, & Pashler, 2014; Karpicke & Bauernschmidt, 2011; Nakata, in press; Pyc & Rawson, 2007).

Not only relative but also absolute spacing is found to affect vocabulary learning. Absolute spacing is defined as the total amount of spacing that intervenes all retrieval opportunities of a given word (Karpicke & Bauernschmidt, 2011). For example, if a given word is retrieved five times, and each retrieval is intervened by 4 minutes, absolute spacing is 16 minutes (4 minutes $\times$ 4). There are two phenomena regarding the effects of absolute spacing: spacing effect and lag effect. First, according to the spacing effect, spaced learning, which introduces spacing between retrievals of a given word, increases learning more than a massed schedule, which does not involve any spacing (see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1989, 1996; Janiszewski, Noel, & Sawyer, 2003, for meta-analyses). Second, the lag effect refers to a phenomenon where larger absolute spacing enhances learning more than shorter absolute spacing (e.g., Bird, 2010; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Cepeda et al., 2006, 2009; Pashler, Rohrer, Cepeda, & Carpenter, 2007; Rohrer & Pashler, 2007). Although the spacing effect has been found to affect L2 vocabulary acquisition (e.g., Karpicke & Bauernschmidt, 2011; Nakata, in press), existing L2 vocabulary studies have produced inconsistent results regarding the lag effect (e.g., Crothers & Suppes, 1967; Karpicke & Bauernschmidt, 2011; Nakata, in press; Nakata & Webb, in press; Pashler, Zarow, & Triplett, 2003; Pyc & Rawson, 2007).

Another line of research suggests that the timing of feedback after retrieval attempts affects learning (e.g., Butler, Karpicke, & Roediger, 2007; Kulik & Kulik, 1988; Metcalfe, Kornell, & Finn, 2009; Nakata, 2015). For example, suppose that the learner was asked to retrieve the meaning of a Japanese word *inu* (*dog*). Which would be more effective, to provide the correct answer 3 seconds after the retrieval attempt (immediate feedback) or 1 hour after (delayed feedback)? Some researchers claim that delayed feedback increases learning more than immediate feedback because the former introduces larger spacing than the latter, which should facilitate learning according to the lag effect (e.g., Butler et al., 2007; Metcalfe et al., 2009). Although some non-L2 vocabulary studies have found the facilitative effects of delaying feedback (e.g., Butler et al., 2007; Kulik & Kulik, 1988; Metcalfe et al., 2009), L2 vocabulary research has failed to support this finding (Nakata, 2015).

The findings of earlier research are useful because they allow us to determine the optimal way to learn L2 vocabulary from retrieval. One limitation of existing research, however, is that none of them examined learners' perceived effects of relative spacing, absolute spacing, and feedback timing on L2 vocabulary learning. Cognitive psychology literature on judgments of learning (JOLs) suggests that learners are often unable to make accurate judgments about the effectiveness of retrieval-based learning techniques (e.g., Karpicke & Roediger, 2008; Kornell, 2009; Kornell & Son, 2009; Roediger & Karpicke, 2006). Examining learners' perceived effectiveness of retrieval-based learning techniques is of value because learners' perceptions may encourage or discourse the use of effective learning techniques. For instance, learners perhaps tend to employ vocabulary learning techniques that they consider more effective than those they consider less effective, as long as other variables (e.g., time or cost

required) are held constant. With the above discussion in mind, the present study examined learners' perceived effects of relative spacing, absolute spacing, and feedback timing on L2 vocabulary learning.

2 Method

A total of 226 Japanese college students participated in this study and studied English-Japanese word pairs using flashcard software (e.g., Nakata, 2011, 2013) developed by the author. One hundred and thirty-eight of them were randomly assigned to one of the four groups: massed, short, medium, and long spacing. In all four groups, the treatment involved typing the English target word corresponding to a Japanese translation provided (e.g., 幽霊 = ____?). In the massed group, 20 target English-Japanese word pairs (e.g., *apparition* – 幽霊) were practiced four times in a row without any spacing. In the short, medium, and long spacing groups, four encounters of a given target word pair were separated by approximately 1, 2, and 6 minutes on average, respectively. In the short, medium, and long spacing groups, half of the target word pairs were studied in an expanding spacing condition, and the other half were studied in an equal spacing condition. There was no equal or expanding spacing condition in the massed group because all target items were encountered four times in a row, and there was no spacing in this group. After completing the treatment and posttests, a questionnaire was administered to examine learners' perceived effects of relative and absolute spacing schedules on learning. The short, medium, and long spacing groups were asked to evaluate the usefulness of the equal and expanding schedules for learning on a 7-point scale, where 1 means *Not helpful at all* and 7 means *Very helpful*. Participants in the massed group were asked to evaluate the usefulness of the massed schedule for learning on the same 7-point scale.

The remainder of the participants studied 16 English-Japanese word pairs using flashcard software. The treatment for these participants also involved typing the English target word corresponding to a Japanese translation provided (e.g., 幽霊 = ____?). For a given participant, for half of the target word pairs, the target English-Japanese word pair (e.g., 幽霊 = *apparition*) was provided as feedback immediately after each retrieval opportunity (immediate feedback). For the other half, feedback was not given until all other items were studied (delayed feedback). After completing the treatment and posttests, the participants were asked to evaluate the effectiveness of immediate and delayed feedback on a 7-point scale, where 1 means *I learned more with immediate feedback* and 7 means *I learned more with delayed feedback*.

3 Results

Regarding the perceived effects of relative spacing schedules, the average rating (*SD*s in parentheses) on a 7-point scale (1: *Not helpful at all*; 7: *Very helpful*) was 5.16 (1.08), 5.09 (1.30), 4.91 (1.78), 4.75 (1.88), 4.55 (1.59), and 4.65 (1.40) for the short equal, short expanding, medium equal, medium expanding, long equal, and long expanding conditions, respectively. (One participant in the long spacing group did not provide responses.) When collapsed across the absolute spacing

groups, the average rating was 4.87 (1.52) for equal spacing and 4.83 (1.55) for expanding spacing. The responses were entered into a two-way mixed design 2 (relative spacing: equal/expanding) × 3 (absolute spacing: short/medium/long) ANOVA. None of the main effects or interaction was significant, main effect of relative spacing: F (1, 92) = 0.14, p = .706, $\eta_p^2 < .01$; main effect of absolute spacing: F (2, 92) = 1.07, p = .349, $\eta_p^2 = .02$; interaction between the two variables: F (2, 92) = 0.47, p = .627, $\eta_p^2 = .01$. The results are also supported by the small effect sizes ($\eta_p^2 \leq .02$). The findings indicate that (1) no statistically significant difference existed between equal and expanding spacing in the ratings, and (2) the three groups did not differ significantly from each other in their responses regarding the effects of equal and expanding spacing.

Next, let us examine learners' perceived effects of the four absolute spacing schedules (massed, short, medium, and long) on learning. In the short, medium, and long groups, the scores for equal and expanding spacing were averaged out and compared with those for the massed schedule. For instance, if a participant in the short spacing group gave 5 for equal and 6 for expanding spacing, 5.5 was used as this participant's rating for the short spacing schedule. Because there was little difference between the ratings for equal (M = 4.87, SD = 1.52) and expanding spacing (M = 4.83, SD = 1.55; see above), it was judged appropriate to use the mean score of the two relative spacing schedules to represent a given absolute spacing schedule. The average rating (SDs in parentheses) on a 7-point scale (1: *Not helpful at all*; 7: *Very helpful*) was 3.78 (1.41), 5.13 (1.09), 4.83 (1.73), and 4.60 (1.44) for the massed, short, medium, and long spacing schedules, respectively. A one-way ANOVA found a statistically significant difference among the four groups, F (3, 123) = 5.18, p = .002, η^2 = .01. According to the Bonferroni method of multiple comparisons, the scores for the short and medium spacing schedules were significantly higher than those for the massed schedule, producing medium to large effect sizes (massed vs. short: p = .002, d = 1.07; massed vs. medium: p = .025, d = 0.66). The difference between the massed and long schedules did not reach statistical significance (p = .154), but a medium effect size was observed (d = 0.59). The differences were not statistically significant for all other comparisons ($p \leq 1.000$), and no more than small effect sizes were observed ($0.15 \leq d \leq 0.42$). These results suggest that (1) the participants perceived the short and medium spacing schedules to be more effective than the massed schedule, and (2) the short, medium, and long spacing groups did not differ significantly from each other in their responses.

As for the effects of feedback timing, the average rating on a 7-point scale (1: *I learned more with immediate feedback*; 7: *I learned more with delayed feedback*) was 3.85 (SD = 2.01). The results indicate that learners tended to believe that they learned equally well from the two types of feedback.

4 Discussion and Conclusions

The present study showed that the participants (1) considered equal and expanding spacing to be equally effective, (2) perceived spaced learning to be more effective than massed learning, and (3) considered immediate and delayed feedback to be equally effective, all of which are supported by existing empirical research

(e.g., Kang et al., 2014; Karpicke & Bauernschmidt, 2011; Nakata, 2015, in press; Pyc & Rawson, 2007). The results suggest that contrary to the findings of earlier research on judgments of learning (e.g., Karpicke & Roediger, 2008; Kornell, 2009; Kornell & Son, 2009; Roediger & Karpicke, 2006), the participants in this study were able to make accurate judgments about the effectiveness of retrieval-based learning techniques. The inconsistent results were perhaps due to a difference in participants. Although most existing studies on judgments of learning have been conducted with American university students (e.g., Karpicke & Roediger, 2008; Kornell, 2009; Kornell & Son, 2009; Roediger & Karpicke, 2006), this study was conducted with Japanese college students. Because Japanese students tend to be more experienced in and proficient at rote memorisation than American students (Tinkham, 1989), the participants of the present research were perhaps more familiar with effective learning strategies than suggested by earlier studies. This may be partly the reason why the results of this study were at odds with those of the existing literature on judgments of learning. However, due to the paucity of metacognition research on Japanese students, further research is warranted.

Although the findings of this study are useful, the present study also suffers from some limitations. One limitation is that the questionnaire was given only to Japanese college students. Future research may replicate this study with participants from different backgrounds. Another limitation is that the current study only examined learners' perceived effectiveness of retrieval-based learning techniques after the treatment. In future research, it would be useful to give a questionnaire before and after the treatment.

References

Baddeley, A. D. (1997). *Human memory: Theory and practice* (Revised ed.). East Sussex: Psychology Press.

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*, 35–56. doi:10.1111/j.1467-9922.2007.00398.x

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, *31*, 635–650. doi:10.1017/S0142716410000172

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*, 273–281. doi:10.1037/1076-898X.13.4.273

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*, 236–246. doi:10.1027/1618-3169.56.4.236

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380. doi:10.1037/0033-2909.132.3.354

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102. doi:10.1111/j.1467-9280.2008.02209.x

Crothers, E., & Suppes, P. (1967). *Experiments in second language learning*. New York, NY: Academic Press.

Cull, W. L., Shaughnessy, J. J., & Zechmeister, E. B. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology: Applied*, *2*, 365–378. doi:10.1037/1076-898X.2.4.365

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review*, *1*, 309–330. doi:10.1007/BF01320097

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Human memory* (pp. 197–236). San Diego, CA: Academic Press.

Ellis, N. C. (1995). The psychology of foreign language vocabulary acquisition: Implications for CALL. *Computer Assisted Language Learning*, *8*, 103–128. doi:10.1080/0958822940080202

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal, and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge: Cambridge University Press.

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research*, *30*, 138–149. doi:10.1086/374692

Kang, S. H. K., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review*, *21*, 1544–1550. doi:10.3758/s13423-014-0636-z

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*, 1250–1257. doi:10.1037/a0023436

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968. doi:10.1126/science.1152408

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*, 1297–1317. doi:10.1002/acp.1537

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*, 493–501. doi:10.1080/09658210902832915

Kulik, J. A., & Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*, 79–97. doi:10.3102/00346543058001079

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition*, *37*, 1077–1087. doi:10.3758/MC.37.8.1077

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning*, *24*, 17–38. doi:10.1080/09588221.2010.520675

Nakata, T. (2013). Web-based lexical resources. In C. Chapelle (Ed.), The encyclopedia of applied linguistics (pp. 6166–6177). Oxford: Wiley-Blackwell.

Nakata, T. (2015). Effects of feedback timing on second language vocabulary learning: Does delaying feedback increase learning? *Language Teaching Research*, *19*, 416–434. doi:10.1177/1362168814541721

Nakata, T. (in press). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*. doi:10.1017/S0272263114000825

Nakata, T., & Webb, S. A. (in press). Does studying vocabulary in smaller sets increase learning? The effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition*. doi:10.1017/S0272263115000236

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

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin and Review*, *14*, 187–193. doi:10.3758/BF03194050

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1051–1057. doi:10.1037/0278-7393.29.6.1051

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, *35*, 1917–1927. doi:10.3758/BF03192925

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255. doi:10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & Karpicke, J. D. (2010). Intricacies of spaced retrieval: A resolution. In A. S. Benjamin. (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 23–47). New York, NY: Psychology Press.

Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, *16*, 183–186. doi:10.1111/j.1467-8721.2007.00500.x

Schmitt, N. (2007). Current trends in vocabulary learning and teaching. In J. Cummins & C. Davison (Eds.), *The international handbook of English language teaching* (pp. 827–842). Norwell, MA: Springer.

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition*, *38*, 244–253. doi:10.3758/MC.38.2.244

Tinkham, T. (1989). Rote learning, attitudes, and abilities: A comparison of Japanese and American students. *TESOL Quarterly*, *23*, 695–698. doi:10.2307/3587547

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

Article

Mastery Sentences: A Window into the Interplay between Word Knowledge Types

Andrew Gallacher
Kysushu Sango University
doi: http://dx.doi.org/10.7820/vli.v04.1.gallacher

Abstract

This study investigates students' productive knowledge of high-frequency vocabulary using a word knowledge framework. A sample of more than 300 written "mastery sentences" was analyzed and coded for errors made at the meaning level of the target word, as well at the usage level. These sentences were categorized into three groups, based on students' English proficiency, which roughly translates to low-to-mid A2, mid-to-high A2, and low B1 on the Common European Framework of Reference. Preliminary findings indicate that while word knowledge at the *meaning* level improves with proficiency, knowledge at the *usage* level appears to plateau once students reach a mid-level of proficiency. Further investigation as to why this may be the case indicates that although higher level students tend to make roughly an equal number of grammatical and collocation-type mistakes compared to their mid-level counterparts, they do so in attempts to use richer, less frequent vocabulary in their mastery sentences. These preliminary findings suggest that, in the spirit of mastery, students may be making more errors as they test the limits of newly acquired vocabulary. Future research will further investigate this hypothesis.

1 Background

Until recently, most vocabulary teaching and learning strategies have prioritized receptive skills over productive skills (Webb, 2005), mainly because productive skills have proven more difficult to elicit (De la Fuente, 2002). This is especially true in the Japanese EFL context, where students receive little-to-no opportunity for productive engagement with the language, both in and out of the classroom. By the time students reach university, they may be able to translate word-to-word meaning pretty efficiently but often still have difficulty using these words in sentences of their own creation. For this reason, studies have begun to stress the importance of incorporating productive activities into vocabulary instruction and acquisition (Laufer & Paribakht, 1998; Ryan, 1997; Schmitt, 1998; Schmitt & Meara, 1997).

One of the more popular ways of integrating productive elements into vocabulary study is to have students write sentences for new words they are learning (see Al-Murtadha, 2013; Folse, 2006; Masson, 2012; Pichette, de Serres, & Lafontaine, 2011). The argument for doing this is that by moving away from one-to-one lexical processes and focusing more on production of semantic chunks, the students will begin to engage with other "word knowledge types" (Nation,

Please cite this article as: Gallacher, A. (2015). Mastery sentences: A window into the interplay between word knowledge types. *Vocabulary Learning and Instruction*, *4* (1), 74–82. doi: 10.7820/vli.v04.1.gallacher

1990, p. 31 – see below), which in turn will help improve upon other productive skills on the way toward overall vocabulary mastery (Schmitt, 2010).

Word knowledge types (Nation, 1990):

- the meaning(s) of the word
- the written form of the word
- the spoken form of the word
- the grammatical behavior of the word
- the collocations of the word
- the register of the word
- the associations of the word
- the frequency of the word

How divergent is knowledge of definitional meanings of L2 words from knowledge of usage of these same words in original sentences? The current study attempts to examine the relationship between knowledge of word meaning, proficiency, and vocabulary usage (i.e. the other word knowledge types) through a qualitative examination of a year's worth of "mastery sentences" produced by students at three levels of L2 proficiency. This preliminary, qualitative investigation is intended as a first step toward future quantitative studies on the topic.

1.1 Mastery Sentences

The *mastery sentence* method of vocabulary study is one component of a larger integrated-learning curriculum designed to improve students' English listening, speaking, reading, and writing abilities at a university in Southern Japan (for more information about the program, please see Anderson and Wilkins (2014) and Masson (2013)). For their vocabulary study, students are required to study 400 of the 1600 most commonly used words in the British National Corpus (for more information please visit: http://www.natcorp.ox.ac.uk/) per semester, totaling 800 in a year. These words are split into 10 lists composed of 40 words each, which are then assigned as weekly homework over the course of a 10-week period. From each list, students are asked to identify the words with which they are least familiar (with a minimum of 20 being set) and use them to comprise vocabulary notebooks as recommended by (Folse, 2006).

Students are asked to include in their vocabulary notebooks the following: the correct written form of each word; its most common two meanings in Japanese; some associated words and/or collocates of each word; and an exemplary mastery sentence demonstrating one of the two meanings of the word. After completing this homework each week, students are given a quiz on 10 randomly chosen words from the previous week's list. On the quiz, students are required to write the 10 words chosen, along with a mastery sentence for each.

In essence, mastery sentences are elaborate sentences that indicate word meaning by employing as many of the other word knowledge types as possible. In this way, the target words become contextually bound to the sentence as a whole.

A successful mastery sentence thus becomes one in which the target word, if removed, could only be replaced by a direct synonym (Masson, 2012).

For example: *banana*

(1) Monkeys love to peel and eat ripe yellow *bananas*.

In previous studies measuring students' perceptions of the mastery sentence approach it was found that, despite being difficult for them to understand and put into practice, students believed the technique to be beneficial in (1) building their vocabulary knowledge and (2) improving their overall ability to compose sentences (Al-Murtadha, 2013; Masson, 2012).

2 Methodology

2.1 Participants

Three first-year Japanese private university students (age 19) enrolled in their school's Four Skills English program were chosen to participate in this study. In order to conceal their identities, each of the participants have been given a pseudonym. Their details are as follows.

2.1.1 Hank (TOEIC score = 400)

Hank (male) is majoring in International Affairs with a focus in Asian Relations. His 400 Test of English for international communication (TOEIC) score is roughly equivalent to a low-to-mid A2 level on the *Common European Framework of Reference* (CEFR). Hank had an 85% attendance rate in the Four Skills program and contributed "the minimum amount to pass" throughout the course, thus receiving an overall "C" grade. For the purposes of this study, Hank exemplifies a "low" level Four Skills student.

2.1.2 Nancy (TOEIC score = 490)

Nancy (female) is majoring in Fine Arts with a focus on design. As a design student, Nancy's overall course load tends to be more demanding than the others. As a result, time management occasionally proved problematic for Nancy, although her resilience and participation in class allowed her to pass with a "B" grade. Nancy's 490 TOEIC score puts her at a mid-to-high A2 level on the CEFR. She had an 87% attendance rate in the Four Skills program and is considered a "mid" level Four Skills student for the purposes of this research.

2.1.3 Cindy (TOEIC score = 635)

Born and raised in Okinawa (arguably a more Anglicized part of Japan), Cindy (female) entered university with a greater English fluency than most of her peers. Her 635 TOEIC score is considered a low B1 level on the CEFR. Furthermore, Cindy's major in International Affairs with a focus on English Communication, along with her continued enthusiastic participation in class, is demonstrative of her personal interest in learning English. She had a 90%

attendance rate in the Four Skills program and achieved an "A" grade overall. Cindy is considered a "high" level Four Skills student.

2.2 Procedure

This research was based on mastery sentences (MS) produced by students as a regular component of their Four Skills coursework. The participants were asked if the vocabulary-related material they produced throughout the year could be used for the current research purposes. Once permission was obtained, the participants' mastery sentence-based vocabulary quizzes were isolated, collated, and copied in preparation for qualitative assessment (see Appendix 1 for an example of the raw data).

With the tested vocabulary categorized by part of speech (noun, verb, adjective), students' mastery sentences were then scored for correct meaning (M) and usage (U) of tested words. Meaning was deemed correct when auditors (inter-rater reliability = 85%) were satisfied that a sentence provided enough context to properly convey the underlying concept or "core meaning" (see Schmitt, 2010, p. 27) of a target item. Conversely, meaning was considered incorrect when: (a) participants failed to produce a sentence for the target item altogether; (b) the sentence they produced did not adequately convey the underlying concept (i.e. the example was not a mastery sentence); or (c) the sentence conveyed a meaning of the target item that is not conventionally associated with that word.

Usage was considered correct when words were used within clauses that were both grammatically correct and contextually appropriate (i.e. displayed proper use of connotation, register, and collocation). Restricting this evaluation to clauses was done in order to avoid penalizing students that made usage errors unrelated to the target words. Usage was marked as incorrect when clauses were either (a) grammatically flawed, (b) contextually inappropriate, (c) both grammatically flawed and contextually inappropriate, or (d) non-existent.

Occasionally, there were instances in which sentences were too short or vague for the auditor to determine with a degree of confidence whether or not the student knew the meaning of the word (e.g., "There is death near animals."). These cases were classified as "insufficient data" and removed from the analysis. The remaining data were then tabulated to compare proportions of correct usage and meaning by student level and part of speech.

3 Results

Combined, participants had far more instances of correct meaning than correct usage (see Table 1), which was also true for each level of proficiency (see Table 2) and part of speech (see Table 3).

Table 1. Overall Percentages of Correct Usage and Meaning

Usage	38
Meaning	71

Table 2. Overall Percentages of Usage and Meaning by Level

	Low	Mid	High
Usage	28	41	43
Meaning	53	70	85

Table 3. Percentages of correct Usage and Meaning by Level and Part of Speech

	Low	Mid	High
Noun			
Usage	16	33	45
Meaning	47	70	89
Verb			
Usage	29	48	43
Meaning	58	74	86
Adjective			
Usage	21	34	34
Meaning	56	63	78

Note: Meaning was shown to increase with proficiency while usage appeared to plateau between the mid- and high-level learners (see Figure 1).

To investigate this further, 15 mastery sentences in which students had correct meaning but incorrect usage were isolated for both the mid- and high-level students and analyzed using Lextutor's *Vocabprofile*, found on the website: http://www.lextutor.ca/. The results of this analysis yielded profiles for both the mid- and high-level learners showing the frequency of vocabulary each student used in mastery sentences with correct meaning but incorrect usage (see Table 4).

They indicate that although the higher level student had a comparable percentage of incorrect word usages, the sentences they produced used less frequent vocabulary. Given this, it is possible that the risks they took in their sentence production led to a comparable proportion of usage errors, despite their overall higher proficiency. Table 5 illustrates this further with example mastery sentences written by the mid- and high-level participants.

In these examples, the mid-level student gains more correct usage points by relying on simpler language, as indicated by the use of higher frequency words and the first-person subjunctive. Conversely, the higher level learner is caught dropping articles and mistaking verb forms while attempting to use much less frequent vocabulary in much more difficult ways. For example she wrote *common occurrence* rather than the correct collocation in this context (***a*** *common occurrence*). However, it is commendable that she attempted to use this relatively infrequent collocation containing an abstract noun (occurrence) at all. Had the learner attempted a simpler sentence using a more concrete noun as the object (e.g., *cats are common*

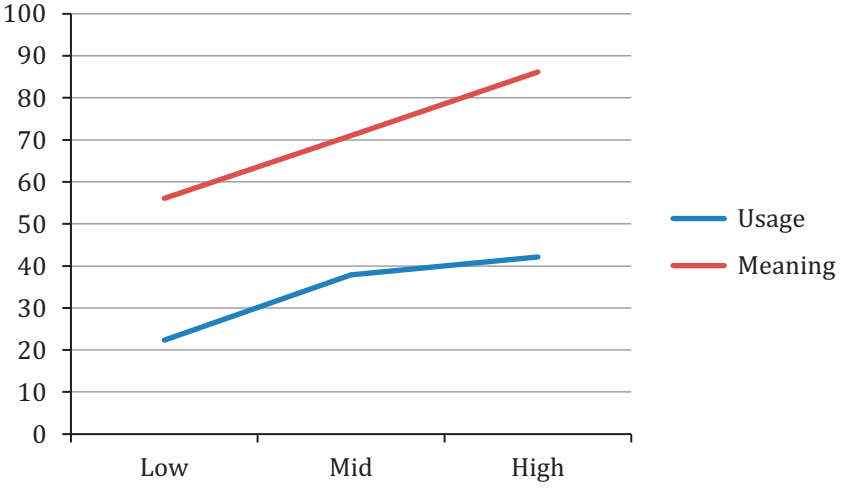

Figure 1. Usage and meaning by proficiency.

Table 4. Frequency of Vocabulary in Mastery Sentences with Correct Meaning but Incorrect Usage

Freq. level	Mid	High
K-1 words	83.54	78.31
K-2 words	9.15	10.24
K-3 words	3.66	3.01
K-4 words	–	1.81
K-5 words	–	0.6
K-6 words	1.22	0.6
Off-list:	2.44	5.42
Total	164 (100)	166 (100)

Note. Table gives percentages of tokens in each frequency band.

Table 5. Example Mastery Sentences by Mid- and High-level Learners

	Target word	Example sentences
Mid-level (correct usage and correct meaning)	first name	My first name is Nancy.
	spoon	I usually use a spoon when I eat soup.
High-level (incorrect usage and correct meaning)	common	A squall is common occurrence in the tropical regions.
	strategy	The government have to decide strategies about decrease a crime.

pets), she likely would have been marked correct for usage on this word. In this respect, it could be argued that the learner's higher proficiency acted as an impediment in regards to her ability to demonstrate correct usage.

4 Conclusions

In keeping with findings by Schmitt (1998) and Schmitt and Meara (1997), it is clear that knowledge of word usage trails that of knowledge of word meaning, regardless of proficiency. Interestingly, however, gains in knowledge of word meaning were not matched by relatively equal gains in knowledge of word usage. Instead the mid- and high-level students had roughly equal difficulty with word usage, despite the fact that the higher level student demonstrated knowledge of substantially more word meanings.

Subsequent analysis of the frequencies of words used in the mastery sentences produced by the mid- and high-level students suggested that this could possibly be a result of the higher level student attempting to use a richer range of language in her mastery sentences, which could have resulted in an equivalent proportion of usage errors. This suggests that, when writing mastery sentences, some higher level students may try to use more of the language they have learned. Thus, it may be that in the spirit of mastery, these learners may commit more errors as they test the limits of their vocabulary knowledge.

As educators, we should be encouraging of this experimentation in such cases, as the increased errors can be seen as opportunities for corrective feedback for learners that is level appropriate, and that learners are ready for. From a research perspective, it is apparent that further investigation into the nature of interaction between word knowledge types is greatly needed.

References

Al-Murtadha, M. (2013). Investigating the impact of sentence writing on second language vocabulary acquisition. *Vocabulary Education & Research Bulletin*, *2*(2), 2–3.

Anderson, C. J., & Wilkins, S. (2014). The four skills and advanced reading and writing syllabi. *Kyushu Sangyo University Language Education and Research Center Journal*, *9*, 63–74.

De la Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary. *Studies in Second Language Acquisition*, *24*(1), 81–112. doi:10.1017/S0272263102001043

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly*, *40*, 273–293. doi:10.2307/40264523

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, *48*, 365–391. doi:10.1111/0023-8333.00046

Masson, M-E. (2012). Student feedback regarding the use of "mastery sentences. *Vocabulary Education & Research Bulletin*, *1*(1), 5–6.

Masson, M-E. (2013). The development of the *Four Skills* course: An introductory course in the career English program. *Kyushu Sangyo University Language Education and Research Center Journal*, *8*, 48–65.

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

Pichette, F., de Serres, L., & Lafontaine, M. (2011). Sentence reading and writing for second language vocabulary acquisition. *Journal of Applied Linguistics*, *32*(5), 1–18. doi:10.1093/applin/amr037

Ryan, A. (1997). Learning the orthographical form of L2 vocabulary – a receptive and productive process. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 181–198). Cambridge: Cambridge University Press.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, *48*, 281–317. doi:10.1111/1467-9922.00042

Schmitt, N. (2010). *Vocabulary in language learning*. Cambridge: Cambridge University Press.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*(1), 17–36. doi:10.1017/S0272263197001022

Webb, S. (2005). Receptive and Productive Vocabulary Learning. *Studies in Second Language Acquisition*, *27*(1), 33–52. doi:10.1017/S0272263105050023

Appendix 1 Mastery Sentence – Student Quiz Example.

	Word	Mastery Sentence	Mark
1	complain	He wants to complain about his school rules.	A
2	solve	Police officers should solve all any problems.	C
3	curious	Childrens are curious about many things.	C
4	Suck (V) 吸う, しゃぶる	~~Babies are~~ I ~~drink~~ Sucked an orange juice with a straw	X
5	uncomfortable	This chair is very uncomfortable to sit for me.	A
6	pop (V) はじける	I always listening pop music. ~~A ballon is poped~~ A ballon was poped Suddenly.	X
7	Senior	We have to give a ~~a~~ seat in a train or bus when the senior person gets on ~~A~~.	A
8	Strategy	The government should make some strategy to reduce the car accident.	A
9	pretend 装う	I didn't pretend that he came to my house. I pretended ~~not to notice him~~ that I didn't notice him because I hate him.	A
10	brain (V) 非難する	blame When I failed something. I blame myself.	X

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

Article

Researching Vocabulary in the EFL Context: A Commentary on Four Studies for JALT Vocabulary SIG

Stuart Webb
University of Western Ontario
doi: http://dx.doi.org/10.7820/vli.v04.1.webb

Abstract

Four papers by Anna C-S. Chang, Yuko Hoshino, Tatsuya Nakata and Andrew Gallacher were presented in the afternoon session of the 4th Annual JALT Vocabulary SIG Vocabulary Symposium in Fukuoka, Japan on June 20, 2015. As discussant, it is my pleasure to comment upon each manuscript. The four studies investigate different issues related to vocabulary learning: learning vocabulary through extensive reading, measuring vocabulary knowledge, perceptions of vocabulary learning through flashcards, and learning vocabulary through writing. After commenting on each paper in turn, I shall present a few suggestions for future research on each of these important issues.

1 Introduction

There has been a great deal of research on vocabulary in Japan in recent years. Perhaps the reason for this is the close relationship that two leading scholars, Paul Nation and Paul Meara, have had with Japan. Paul Nation has spent a considerable amount of time in Japan based in part on his work at Temple University Japan. His frequent presentations around Japan have influenced many researchers and fueled much of the research on lexis in the Japanese EFL context. Paul Meara's work with the Swansea University MA and Ph.D. programs that focused on lexical studies attracted and established a large number of researchers based in Japan. That program has since moved to Cardiff University under Tess Fitzpatrick's guidance but continues to ignite a passion for vocabulary research and a stream of publications by students based in Japan. It should be of no surprise then that there is a lot of useful research currently going on in the Japanese EFL context, and the studies presented in this issue are an indication of this. Three of the four were conducted in Japan, and the fourth was completed in Taiwan.

The four studies investigate different issues related to vocabulary learning: learning vocabulary through extensive reading (Chang), measuring vocabulary knowledge (Hoshino), perceptions of vocabulary learning through flashcards (Nakata), and learning vocabulary through writing (Gallacher). Together they do a great job of revealing the depth and breadth of research on vocabulary that is being conducted today.

Please cite this article as: Webb, S. (2015). Researching vocabulary in the EFL context: A commentary on four studies for JALT vocabulary SIG. *Vocabulary Learning and Instruction*, *4* (1), 83–94. doi: 10.7820/vli.v04.1.webb

2 The Four Studies

2.1 An Investigation of Different Text Levels on L2 Learners' Vocabulary Learning Rates in an Extensive Reading Program, by Anna C-S. Chang

Although there has been a large amount of research investigating incidental vocabulary learning through reading a single text, there have been relatively few studies of vocabulary learning through extensive reading. The importance of vocabulary growth through reading is well established, and further studies of vocabulary learning through extensive reading are clearly warranted. The challenge of conducting studies of vocabulary learning through extensive reading is that they need to involve participants reading multiple texts over a relatively long period of time. Variables that are typically controlled for in short-term studies may then come into play. Because there are so many highly controlled studies of incidental vocabulary learning through reading a single text, there is strong support for ecologically valid studies such as this one by Anna Chang.

This paper builds on two studies that Anna Chang and I recently completed that look at how prior word knowledge (Webb & Chang, 2015) and frequency and distribution of occurrence (Webb & Chang, 2014) affect learning. It looks at text level, which is another variable that was examined in the former study, and frequency, which was examined in the latter. Because Webb and Chang's (2014) study showed an absence of a frequency effect in contrast to many earlier studies (e.g., Horst, Cobb, & Meara, 1998; Pigada & Schmitt, 2006; Waring & Takaki, 2003; Webb, 2007), further research here is useful. The paucity of research on text level makes it another useful factor to examine.

There are many interesting aspects of the study. First, although the Level 1 and Level 3 readers are expected to vary considerably in the vocabulary found in the texts, the lexical profiles of the readers are remarkably similar. The primary difference between the proportions of items in the different frequency bands in the two levels is that there are a greater proportion of proper nouns in the Level 1 texts than in the Level 3 texts. In fact, the proportion of words at the 2000 word level (5.61%) is exactly the same in the two sets of five graded readers. The similarity between the two sets of graded readers is somewhat surprising but supported by the findings of an earlier study that I conducted with John Macalister. Webb and Macalister (2013) found that there was little difference between the lexical profiles of Stage 1, 2, 3, and 4 graded readers in the Oxford Bookworms series.

The use of audio support during the extensive reading treatment also deserves mention. Research indicates that vocabulary learning gains are likely to be greater for reading while listening than reading alone (Brown, Waring, & Donkaewbua, 2008; Webb & Chang, 2012) suggesting that audio support should be included whenever possible. Audio support may account to some extent for the extremely high gains in this study in comparison to those of earlier reading studies (e.g., Horst et al., 1998; Pigada & Schmitt, 2006; Waring & Takaki, 2003).

One aspect of the methodology that could use more detail is the tests for each block of target words. The blocks included 8–15 target words, but was the number

of target words in each block controlled between the tests for Level 1 and Level 3 items? If block size was smaller for words in the Level 1 readers that may make them slightly easier than for the Level 3 readers.

It is generally useful to use multiple tests to measure vocabulary learning. In this study, two tests were administered to participants. The first test is a contextualized receptive translation test. Justification for including context in this test format is that it may help to cue recall, making it more sensitive to learning gains. This is usually a good thing. A limitation in this study, however, is that the presence of the context may contribute to learning. The results would actually suggest that learning may have occurred through completing the test because the gains were higher on this test than in the matching test that followed. Research indicates that the recall test was likely to be the more difficult of the two (Laufer & Goldstein, 2004). In the example provided in the study (Carol is never late for rehearsals, and she knows that these concerts are important for us.), even if we do not know the target word *rehearsal*, we may learn that it is semantically associated with *concerts*, and that perhaps *late* is a collocate. Completing this test may then affect the validity of the results of the second test. Ideally, the receptive translation test would have been decontextualized to ensure that there was no learning effect from taking it.

It is important to note that vocabulary learning may not have been purely incidental. There was the potential for participants to deliberately learn words in different ways, by looking up entries in dictionaries, by asking classmates, and by using newly encountered words in follow-up activities. It would be unethical to not allow participants to do these things and their use mirrors that of extensive reading programs. Thus, the lack of control may increase ecological validity but may muddy the findings (in terms of incidental vocabulary learning) to some degree.

Another possible factor that may have affected findings is the difference between the treatment times for Level 1 and Level 3 graded readers. In the former, it was 5 weeks, and in the latter it was 8 weeks. The greater interval for Level 3 graded readers may have reduced the amount of learning because knowledge of target words that were encountered in the first weeks of the treatment had greater chance to decay for this set of words.

The lack of a relationship between frequency of occurrence and vocabulary learning supported those of the earlier study by Webb and Chang (2014). However, I wondered whether looking at the data in a different way might indicate a relationship between the two variables. Rather than looking at them for the individual sets of words, perhaps combining the two sets and then looking at gains according to frequency bands (1–5 occurrences, 5–10 occurrences, 11–20 occurrences, and 20+ occurrences) might have provided slightly different results.

The results revealed extremely high relative learning gains. Gains ranged from 79.33% to 93.22% at the end of the treatment and from 71.43% to 81.92% on the delayed posttests. These gains exceed those of earlier studies and are in fact higher than those of many studies of deliberate vocabulary learning. The size of the gains might be supported by the use of audio support, as well as the possibility of dictionary use and the use of words in post reading activities. I wondered if other reasons for such large gains were learning effects from taking the tests, awareness of

the focus on vocabulary learning for the second set of target words, and because this was the second year that the participants were involved in extensive reading, perhaps an awareness of an aim within the extensive reading program to learn vocabulary. Despite these possible limitations, the results provide strong support for the use of extensive reading as a means of increasing L2 vocabulary growth.

I enjoyed reading this paper. There is a real need for longitudinal studies of vocabulary learning through extensive reading. I hope that the paper helps to show some of the issues that need to be considered in this line of research.

2.2 Measuring Knowledge of Words with Multiple Meanings, by Yuko Hoshino

There have been a large number of studies of teaching and learning vocabulary that have been conducted in recent years. However, within this research, one topic that has really lagged behind has been the development and validation of new tests of vocabulary knowledge (Webb & Sasao, 2013). Moreover, when new measures of vocabulary have been created, they have almost exclusively focused on measuring breadth of vocabulary knowledge. The focus on developing a test designed to measure a component of depth of vocabulary knowledge (knowledge of different meanings of words) in this paper is original, interesting, and of great value.

The paper provides useful discussion on the lack of research on the learning of multiple meanings of words. The issue of polysemy has been neglected in most studies of teaching and learning vocabulary. Research has provided little or no discussion of the degree to which peripheral meanings are known. Similarly, tests of vocabulary knowledge focus on measuring whether test takers are able to link L2 forms to their core meanings. The extent to which peripheral meanings are known is not measured, nor is there discussion or acknowledgment that test takers may not know these meanings when they know the core meanings.

Issues with creating a test designed to measure multiple meanings are clearly laid out, demonstrating the challenges of creating such a test. The first two issues, ambiguity resolution and grammatical knowledge, affect the degree to which test takers are able to identify target meanings. A reasonable solution is presented by the inclusion of context within the test items. The presence of context allows the tester to isolate the target meaning. The third issue relates to the distance in meaning between the prototypical meaning and the peripheral meaning. This may be the biggest challenge with creating a test that can be applied to randomly generated words. When considering words with multiple meanings, we typically think of items that have a large distance between them. *Admit*, *rose*, and *duck* are three examples. However, polysemous high-frequency words often have many meanings and often the differences between these meaning are very slight. For example, *go*, *play*, and *run* all have a large number of meanings of which several are very close. The prototypical meaning of run might be to move quickly on foot. However, someone can also *run in a race*, and *he was running for his life* conveys a meaning that is beyond simply moving quickly on foot. Providing context to help signal target meanings may again be the key to isolating target meanings. However, it may be

impossible to measure knowledge of these slight differences in meaning in a single test, because answering a question about one meaning may affect a subsequent response about a different meaning. A possible solution might be to have multiple tests. The final issue that is discussed, cross-language differences, may affect learners with different first language backgrounds. Variation in overlap between target meanings in different languages may make interpretation of test results difficult in ESL contexts where test takers may have different language learning backgrounds.

Three test formats were piloted: selection, translation, and pairing. These formats were examined in two ways, with full sentence and collocation cues. In the selection test, the test takers need to select sentences (or collocations) from a number of options that includes distracters. Two correct options are provided for each of three target meanings (six in total) and two distracters are also provided. In the translation test, test takers are required to translate the collocations into their first language. Again two examples are provided for each target meaning. The pairing test used a matching format with clusters of items. Test takers needed to match two sentences (or collocations) that included the same meanings from six different options.

Although data are not presented, we find that the most demanding of the test formats (translation), where sentences rather than only collocates were included, provided the greatest reliability and discriminability. This result might be expected because the other two formats add some degree of chance to scoring correctly; it may be possible to successfully guess the answers in the selection and pairing formats, but test takers are unlikely to translate successfully.

One question that arises when looking at the tests is why two examples of each target meaning are necessary in the selection and translation formats. It might also be fair to question whether the two examples are perceived as being identical in meaning. For example, is the meaning of *broke* in *broke the surface* and *broke the skin* identical? It might be argued that the former involves something entering, whereas the latter involves something opening. One thing that stood out in the test is that the final item is written in the present tense (break my promise) while all of the other items were presented in the past tense (e.g., broke the law). It would be best to use the same verb tense for all items.

Another question that I had about the test was whether the frequency of the words used in the sentences and collocations was controlled. I wondered whether some words such as *public*, *scene*, *law*, *surface*, and *promise* would be known. I assume that these words were included because they collocated with the target word. However, it does indicate that finding simple contexts is likely to be another challenge with creating a test format that would be appropriate for learners at varying levels.

Perhaps the biggest question about these three formats is to what extent they were measuring receptive knowledge of collocation rather than receptive knowledge of concept and referents? Because the varying meanings are always defined by collocations, it may be that it is knowledge of collocation that has the greatest impact on test performance. The proposed test formats certainly indicate that

collocation may be the key to defining peripheral meanings so perhaps the two components of lexical knowledge are intertwined.

Two more test formats, isolation and translation recognition, are also discussed. In the isolation format, test takers are required to produce as many meanings for words as possible. In essence it is a receptive recall test that involves recalling multiple meanings for each target item. This may be the ideal format as it clearly isolates the intended aspect of knowledge. However, as noted, it is also an extremely difficult format. Would you be able to produce the slight differences in meaning for the word *run* without the aid of context or collocations to cue recall? I believe that most native-speakers of English would have trouble scoring well. In the translation recognition format, test takers must indicate as quickly as possible whether meanings that are presented are correct. The focus on speed suggests that this test may be tapping into fluency or strength of knowledge.

I really enjoyed reading this paper. Measuring productive knowledge of multiple meanings is something that I had considered, but I had not gotten very far with, because there are indeed many challenges. Yuko Hoshino's investigation of this topic is a really useful step toward creating a test that would enhance our ability to measure depth of knowledge. I hope that she will continue on with this research.

2.3 Are Learners Aware of Effective Ways to Learn Second Language Vocabulary from Retrieval? Learners Perceived Effects of Relative Spacing, Absolute Spacing, and Feedback Timing on Vocabulary Learning, by Tatsuya Nakata

This paper follows up a series of carefully controlled studies that have examined how different variables affect vocabulary learning using flashcards (Nakata, 2015a; Nakata, 2015b; Nakata & Webb, 2015). This is a very useful line of research. Studies have consistently shown that learning vocabulary through a paired-associate learning approach contributes to impressive gains that are very time efficient (e.g., Webb, 1962; Webb, 2009a, 2009b). Researchers that have focused on teaching and learning vocabulary appear to overwhelmingly support a broadly inclusive approach that involves learning through both meaning-focused and form-focused conditions (e.g., Hunt & Beglar, 2005; Laufer, 2003; Nation, 2001; Schmitt, 2008; Webb, 2013; Webb & Nation, 2013). However, there may still be a perception among some researchers that decontextualized learning conditions such as flashcard learning do not deserve a place in the language learning curriculum. This may be one reason for a lack of studies in recent years on word card and flash card learning in applied linguistics and TESOL journals.

Learner perceptions of the effects of vocabulary learning methods is a useful area of research. Although there has been an abundance of studies on vocabulary and the effects of different learning conditions, there is relatively little known about which approaches to learning are perceived as being most useful. It may often take a long time for pedagogically useful research findings to reach the classroom, so investigating which activities are perceived as being effective may provide some insight into how students might learn on their own outside of the classroom.

The results of Nakata's study indicate that the Japanese students who took part in the research were able to accurately judge the effectiveness of the different learning variables; both equal and expanding spacing and immediate and delayed feedback were viewed as equally effective, while spaced learning was seen as being more effective than massed learning. These findings contrast earlier studies that have indicated that students may not accurately rate the learning effects of different conditions. Nakata attributes the differences between his findings and those of earlier findings to variation in the participant profiles.

With data limited to a relatively short questionnaire, it may be fair to question whether participants clearly understood the constructs examined, and if they did, whether their responses can be generalized to slightly different conditions. Ideally, follow-up interviews would be conducted with a sample of the participants to better explore their perceptions of the learning conditions.

One question that we might ask about the findings is whether we can generalize the participants' responses beyond the very carefully controlled spacing conditions to different types of spacing. For example, when we think of spaced learning with flashcards (or any other vocabulary learning condition), we may often consider spacing to occur in days or over multiple learning periods. However, spacing in the studies that were examined was relatively short and occurred between encounters with different items during a single learning period. Thus, it may be useful to look further at perceptions of different types of spacing. Similarly, the variable of feedback timing that was examined in the study involved manipulation of feedback within a relatively short interval in a single learning session. Although these were ecologically valid timings for feedback, there are certainly other variations of feedback manipulation both within and between sessions that would be useful to explore.

Another issue that should be considered when investigating student perception of learning techniques is how feedback on performance in the relevant learning conditions might affect responses. In Nakata's study, when the participants completed the survey, they did not know how they performed on the tests. In the classroom, we might expect learners to base their perceptions on the gains that they have made through using certain techniques. The degree to which the participants in this study knew how the different variables affected their learning performance is not clear. The results do indicate that they were making judgments that reflected their learning performance. However, investigating how knowledge of learning performance affects responses would also be useful.

Learning through electronic applications allows researchers extreme precision in the control of variables. The flashcard/paired-associate learning condition lends itself very well to this type of research. However, because it is one of many different vocabulary learning techniques, it would be useful to examine how the variables examined in Nakata's research (relative spacing, absolute spacing, feedback timing, block size) apply to other learning conditions. Examining student perceptions of different vocabulary learning techniques would be a useful follow-up to this research.

I enjoyed reading this study. There is much research still be done on learning with flashcards. This paper reveals another avenue for research in this area.

2.4 Mastery Sentences: A Window into the Interplay between Word Knowledge Types, by Andrew Gallacher

Students are often asked to do sentence production exercises, where a target word is given and they have to use that word in writing. These exercises typically result in a wide range of responses. For example, responses for the target word *banana* might include: *It is a banana. I like bananas. Bananas are a yellow fruit* and *Monkeys love to pick bunches of bananas from trees and peel and eat the sweet yellow fruit*. Evaluating responses in sentence production exercises is difficult because the sentences demonstrate different degrees of vocabulary knowledge. In the examples above, the only aspects of knowledge that are demonstrated in the first two sentences are written form and grammatical functions. In the third sentence, greater knowledge is demonstrated as knowledge of aspects of meaning are also revealed, as well as collocation. The final sentence, however, conveys a greater degree of vocabulary knowledge because knowledge of aspects of meaning are clearly conveyed and some strong collocates and associations are also included. It is only this final sentence that Gallacher might classify as a mastery sentence.

The concept of mastery sentences is an interesting one for several reasons. First, providing a concrete goal for learners to move beyond knowledge of form-meaning connection and use words correctly in sentences is a useful learning target. As Gallacher notes, many learners may be unable to use words effectively, particularly in the EFL context where opportunities for meaning-focused output are limited. Second, if mastery sentences are learned together with new target words, then perhaps the context may help to cue recall of the target words increasing retention. Moreover, if the sentence can be recalled, then it may provide the foundation necessary for creative use, helping learners to generate the target word in original sentences. Third, perhaps mastery sentences may be a useful way to reduce the effects of cross-association when words are learned in lexical sets. Paul Nation suggested that when there is no option but to learn semantically related words together, it is best to learn each word in a sentence that helps to differentiate it from other words (Nation, 2000). Mastery sentences aim to use target words in a somewhat unique way, where only direct synonyms might replace the target word. This should help to reduce the effects of interference.

The challenge with mastery sentences may be that they are difficult to generate. If we look at the examples provided in the appendix for the target words *solve* (Police officers should solve all problems) and *curious* (Children are curious about many things), although they do demonstrate productive knowledge of aspects of use, they do not meet the requirement that only direct synonyms can be used to replace the target words. We might replace *solve* with *investigate*, *consider*, *examine*, and *discuss*, whereas we could replace *curious* with *anxious*, *informed*, *told*, and *asked*. This makes me wonder whether it would be best for teachers to provide an initial mastery sentence for each target word that might serve as a model to help students generate their own sentences.

Examining the degree to which the mastery sentences demonstrated knowledge of meaning and usage was interesting. However, scoring the correct usage of target words is challenging. Exemplars that demonstrate the criteria for how to successfully score mastery sentences would be useful. One criterion that is reported

in this paper indicates that providing a correct collocation in the sentence is necessary for usage (peel and yellow were used with the target word banana; eat and soup were used with the target word spoon; occurrence was used together with common). In fact I wondered if it would be possible to create a mastery sentence without a correct collocate because collocations may be necessary to reduce the potential to substitute other words for the target word. However, the inclusion of correct collocates (occurrence used together with common) alone is insufficient because the clause that the target word appears in must be grammatically correct. One question that I had when looking at the examples was whether sentences should be incorrect if a target word is used correctly but if there are errors in the clause relating to other words. It might be argued that the omission of the article *a* before *common occurrence* had greater relevance to knowledge of the collocate *occurrence* than it did for the target word *common*. Another way to score the sentences might be to focus on the aspects of use in Nation's (2001) description of vocabulary knowledge (grammatical functions, collocations, constraints on use) that are demonstrated in a sentence. In the high-level sentences provided in the paper that were scored as incorrect usage, both sentences might be scored correctly for knowledge of grammatical functions. The first would be scored correctly for collocations (common occurrence) and the second would not.

Examining the lexical frequency profile of the sentences was useful. The greater use of lower frequency words by students at higher proficiency levels supports earlier findings (e.g., Laufer & Nation, 1995). The finding that it is the use of these lower frequency words that causes errors is quite an interesting finding. In lexical frequency profiling studies errors are often ignored. This study provides some indication that it would be useful to examine not only the use of lower frequency words in writing but also the degree to which the use of lower frequency words develops. Larger quantitative studies of mastery sentences would have probably hidden this interesting finding.

Although this study is small in scale, the results provide an interesting picture of how two aspects of vocabulary knowledge develop in writing. There is a need for qualitative studies of vocabulary (Nation & Webb, 2011). They may provide a more accurate indication of learning within a program and may also provide insight into vocabulary development that is not revealed in larger quantitative studies.

3 Suggestions for Further Research

Although the four papers discussed in this issue all focus on vocabulary, they each look at very different topics: vocabulary learning through extensive reading, measuring vocabulary knowledge, perception of the effects of vocabulary learning through flashcards, and the use of words in student writing. Each of these areas deserve further investigation. I will briefly touch on how each of these areas might be examined in further research.

There have been many studies of vocabulary learning through reading a single text. Although researchers are typically cautious in generalizing their findings to larger scale reading programs, the findings often become recognized as being typical of extensive reading. This is worrying because the results of the few studies of vocabulary learning through extensive reading have provided somewhat different

findings than those of more carefully controlled short-term studies. There is thus a need for more longitudinal studies of vocabulary learning through extensive reading. One reason for the lack of such studies is that they may be more challenging to conduct. A greater number of variables may come into play in longitudinal research so it is more difficult to attribute findings solely to the treatment. However, if we are to clearly determine the value of extensive reading to vocabulary growth, studies that involve large amounts of reading are necessary.

The development and validation of new tests of vocabulary knowledge are also needed. In recent years (and in this issue), there has been greater focus on the development of depth of vocabulary knowledge. Yet, there have been few tests designed to measure depth as a whole or individual aspects of knowledge besides form-meaning connection. With much greater focus on learning multi-word items and formulaic language, tests designed to measure collocation would be useful. Moreover, it would also be useful to measure students' skill at learning words. Yosuke Sasao's (2013) research in this area was particularly innovative and a useful model for further studies.

Despite consistent findings indicating the value of flashcard learning, learning vocabulary with flashcards has been neglected within the fields of applied linguistics and TESOL. One key question that remains to be answered is how many words can be learned with flashcards over different periods of time? Finding this answer would provide useful support for the inclusion of flashcards within the language learning curriculum. It should be noted that flashcards on their own are not the answer to vocabulary learning. However, they provide a starting point for the development of lexical knowledge. Knowledge of target words may then be expanded through further meaning-focused learning tasks. Another useful question to explore is the degree to which the presence of context within a flashcard might contribute to learning.

There is always a lot of interest in the use of words in writing. In recent years, this has been driven to some degree by new software such as RANGE (Nation & Heatley, 2002), AntWordProfiler (Anthony, 2014), and Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) that allows us to analyze large amounts of written text in great detail. TAALES (Kyle & Crossley, in press) is a new application that has evolved and improved on Coh-Metrix. It includes 135 indices of lexical sophistication and richness and may become the next key tool to research the use of words in writing. However, Gallacher's study in this issue also reveals that smaller scale studies might provide insight that is not apparent in these much larger scale studies. Large scale quantitative studies, and qualitative studies that look at the development of different aspects of vocabulary knowledge in L2 writing are still needed.

References

Anthony, L. (2014). AntWordProfiler (Version 1.4.0) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*, 136–163.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*, 193–202. doi:10.3758/BF03195564

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*, 207–223.

Hunt, A., & Beglar, D. (2005). A framework for developing EFL reading vocabulary. *Reading in a Foreign Language*, *17* (1), 23–59.

Kyle, K., & Crossley, S. A. (2014). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*. doi:10.1002/tesq.194

Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence? *Canadian Modern Language Review*, *59*, 567–587. doi:10.3138/cmlr.59.4.567

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*, 399–436. doi:10.1111/j.0023-8333.2004.00260.x

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, *16*, 307–322. doi:10.1093/applin/16.3.307

Nakata, T. (2015a). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, doi:10.1017/S0272263114000825

Nakata, T. (2015b). Effects of feedback timing on second language vocabulary learning: Does delaying feedback increase learning? *Language Teaching Research*, *19*, 416–434. doi:10.1177/1362168814541721

Nakata, T., & Webb, S. (2015). Does studying vocabulary in smaller sets increase learning? Effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition*, doi:10.1017/S0272263115000236

Nation, P. (2000). Learning vocabulary in lexical sets: Dangers and guidelines. *TESOL Journal*, *9* (2), 6–10.

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

Nation, I. S. P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts [software]. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle.

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, *18*, 1–28.

Sasao, Y. (2013). *Diagnostic Tests of English Vocabulary Learning Proficiency: Guessing from Context and Knowledge of Word Parts*. Unpublished PhD Thesis, Victoria University of Wellington, Wellington, New Zealand.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, *12* (3), 329–363.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*, 130–163.

Webb, W. B. (1962). The effects of prolonged learning on learning. *Journal of Verbal Learning and Verbal Behavior*, *1*, 173–182. doi:10.1016/S0022-5371(62)80026-7

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28* (1), 46–65. doi:10.1093/applin/aml048

Webb, S. A. (2009a). The effects of pre-learning vocabulary on reading comprehension and writing. *Canadian Modern Language Review*, *65*, 441–470. doi:10.3138/cmlr.65.3.441

Webb, S. (2009b). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal*, *40*, 360–376. doi:10.1177/0033688209343854

Webb, S. (2013). Depth of vocabulary knowledge. In C. Chappelle (Ed.), *Encyclopedia of applied linguistics* (pp. 1656–1663). Oxford, UK: Wiley-Blackwell. doi:10.1002/9781405198431.wbeal1325

Webb, S. A., & Chang, A. C.-S. (2012). Vocabulary learning through assisted and unassisted repeated reading. *Canadian Modern Language Review*, *68* (3), 267–290. doi:10.3138/cmlr.1204.1

Webb, S. A., & Chang, A. C.-S. (2014). Second language vocabulary learning through extensive reading: How does frequency and distribution of occurrence affect learning? *Language Teaching Research*. 1–20. Published online before print November 28, 2014, doi:10.1177/1362168814559800

Webb, S., & Chang, A. C.-S. (2015). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition*, doi:10.1017/S0272263114000606

Webb, S., & Macalister, J. (2013). Is text written for children useful for L2 extensive reading? *TESOL Quarterly*, *47*, 300–322. doi:10.1002/tesq.70

Webb, S., & Nation, I. S. P. (2013). Teaching vocabulary. In C. Chappelle (Ed.), *Encyclopedia of applied linguistics* (pp. 5670–5677). Oxford, UK: Wiley-Blackwell. doi:10.1002/9781405198431.wbeal1177

Webb, S. A., & Sasao, Y. (2013). New directions in vocabulary testing. *RELC Journal*, *44*, 263–277. doi:10.1177/0033688213500582