Vocabulary Learning and Instruction

Volume 3, Number 2,

September 2014

doi: http://dx.doi.org/10.7820/vli.v03.2.2187-2759

VLI Editorial Team

Editor: Raymond Stubbe

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

Reviewers: Thomas Boutorwick, Dale Brown, Peter Carter, Averil Coxhead, Frank Daulton, Yuko Hoshino, Tomoko Iishi, Kiwamu Kasahara, Masumi Kojima, Paul Meara, Kaori Nakao, Quint Oga-Baldwin, John P. Racine, John Read, Vivienne Rogers, Tim Stoeckel

Copy Editors: Alex Cameron, Mark Howarth, Tim Pritchard, 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 © 2014 *Vocabulary Learning and Instruction*, ISSN: Online 2187-2759; Print 2187-2767. All articles are copyrighted by their respective authors.

Vocabulary Learning and Instruction

Volume 3, Number 2, September 2014
doi: http://dx.doi.org/10.7820/vli.v03.2.2187-2759
Full Issue PDF (print version)

Table of Contents

Article	Page
Letter from the Editor *Raymond Stubbe*	i
A New General Service List: The Better Mousetrap We've Been Looking for? *Charles Browne*	1
Analysing the Effectiveness of Textbooks for Vocabulary Retention *Rachael Ruegg & Cherie Brown*	11
Effects of Glosses and Reviewing of Glossed Words on L2 Vocabulary Learning through Reading *Makoto Yoshii*	19
Effects of Instruction on Yes–No Responses to L2 Collocations *Junko Yamashita*	31
Vocabulary in a Second Language: Selection, Acquisition, and Testing: A Commentary on Four Studies for JALT Vocabulary SIG *Batia Laufer*	38
Estimations of Japanese University Learners' English Vocabulary Sizes Using the Vocabulary Size Test *Stuart McLean, Nicholas Hogg, & Brandon Kramer*	47
Local Item Dependence on the Vocabulary Levels Test Revisited *Tadamitsu Kamimoto*	56
Test Taking and DK Use on the Vocabulary Size Test *Dawn Lucovich*	69
Estimating Learners' Vocabulary Size under Item Response Theory *Aaron Gibson & Jeffrey Stewart*	78
A Review of Four Studies on Measuring Vocabulary Knowledge *Akiyo Hirai*	85

Letter from the Editor

Dear Readers,

Once again it is my pleasure to present to you the official proceedings for the 2014 JALT Vocab SIG Vocabulary Symposium. As in preceding years, the event was staged at Kyushu Sangyo University, on June 14th this year, and was hugely successful.

Following what has now become tradition, the morning session featured four papers concentrating on vocabulary instruction. Topics ranged from selecting vocabulary to assessing knowledge of collocations. Commenting on these fine presentations was the renowned vocabulary researcher, Dr Batia Laufer. The afternoon session focused on vocabulary assessment, with three of the four papers focusing on multiple-choice vocabulary tests such as the VST and the VLT. The discussant for this session, Dr Akiyo Hirai, addressed each paper and raised interesting themes and considerations to help facilitate further research.

On a slightly different, but relevant topic, there are now four additional articles in the 'Online First' page of our website, http://vli-journal.org/wp/online-first/, including a fascinating historical exploration by Paul Meara. More articles are at present in various stages of review.

I hope you enjoy the pages that follow and remember to keep those submissions coming. *VLI* publishes original research regarding vocabulary learning, instruction and assessment. Please check the website for submission guidelines.

Raymond Stubbe,

Editor

Vocabulary Learning and Instruction
Volume 3, Issue 2, September 2014
http://vli-journal.org

Article

A New General Service List: The Better Mousetrap We've Been Looking for?

Charles Browne
Meiji Gakuin University
doi: http://dx.doi.org/10.7820/vli.v03.2.browne

Abstract

This brief paper introduces the New General Service List (NGSL), a major update of Michael West's 1953 General Service List (GSL) of core vocabulary for second language learners. After describing the rationale behind the NGSL, the specific steps taken to create it and a discussion of the latest 1.01 version of the list, the paper moves on to comparing the text coverage offered by the NGSL against both the GSL as well as another recent GSL published by Brezina and Gablasova (referred to as Other New General Service List [ONGSL] in this paper). Results indicate that while the original GSL offers slightly better coverage for texts of classic literature (about 0.8% better than the NGSL and 4.5% more than the ONGSL), the NGSL offers 5–6% more coverage than either list for more modern corpora such as *Scientific American* or *The Economist*.

In the precomputer era of the early 20th century, Michael West and his colleagues undertook a very ambitious corpus linguistics project that gathered and analyzed more than 2.5 million words of text and culminated in the publication of what is now known as The General Service List (GSL; West, 1953), a list of about 2000 high-frequency words that were deemed important for second language learners. However, as useful and helpful as this list has been to us over the decades, it has also been criticized for being based on a corpus that is considered to be both dated (most of the texts were published before 1930), as well as too small by modern standards, and for not clearly defining what constitutes a "word."

In February 2013, on the 60th anniversary of West's publication of the GSL, my colleagues (Brent Culligan and Joseph Phillips of Aoyama Gakuin Women's Junior College), and I put up a website (www.newgeneralservicelist.org) that released a major update of West's GSL known as the New General Service List (NGSL). This list was derived from a carefully selected 273 million word subsection of the 2 billion word Cambridge English Corpus (CEC). The 1.0 version of the NGSL was then published in several journals including the July issue of the Language Teacher (Browne, 2013). Following many of the same steps that West and his colleagues did (as well as the suggestions of Professor Paul Nation, project advisor and one the leading figures in modern second language vocabulary acquisition), we did our best to combine the strong objective scientific principles of corpus and vocabulary list creation with useful pedagogic insights to create a list of approximately 2800 high-frequency words which met the following goals:

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

(1) to update and expand the size of the corpus used (273 million words) compared to the limited corpus behind the original GSL (about 2.5 million words), with the hope of increasing the generalizability and validity of the list
(2) to create a NGSL of the most important high-frequency words useful for second language learners of English which gives the highest possible coverage of English texts with the fewest words possible.
(3) to make a NGSL that is based on a clearer definition of what constitutes a word
(4) to be a starting point for discussion among interested scholars and teachers around the world, with the goal of updating and revising the list based on this input (in much the same way that West did with the original Interim version of the GSL)

Unbeknownst to us, about 6 months after we released the 1.0 version of the NGSL, another GSL was put out by Brezina and Gablasova (August, 2013), which, for the purpose of this article, I will hereafter refer to as the "Other New General Service List (ONGSL)" in order to avoid confusion. Although the ONGSL looks to be a very impressive piece of research, the purpose of their list and the way it was developed seems to be a bit different than the purpose and development process we undertook for the NGSL presented here. The authors state that they used a purely quantitative approach to try to identify high-frequency words that were common across several different corpora, two of which were hugely different in size (1 million words for the Lancaster-Oslo-Bergen Corpus (LOB), and The BE06 Corpus of British English (2006), 100 million for the British National Corpus [BNC], and 12 billion words for the En Ten Ten 12 corpora), and resulted in the identification of 2494 lemmas (according to their way of counting).

Our own NGSL project has been more directly focused on the needs of second language learners and teachers, and started with a selection of sub-corpora that were carefully balanced in size so as to avoid one corpus or type of text dominating the frequencies (which appears to be a real problem in the ONGSL) and, just as with the original GSL, our NGSL project had employed both quantitative as well as qualitative methods to attempt to identify the words that are most useful to the needs of language learners while providing the highest possible coverage. We are following many of the principles West and his colleagues used both for developing as well as for improving the list over time, and are thus referring to the latest version of the NGSL in this article as NGSL 1.01.

The purpose of this brief paper is to explain a bit more about the NGSL as well as to give some initial comparisons in text coverage between the GSL, NGSL, and ONGSL for a range of different text types.

1 The NGSL: a Word List Based on a Large, Modern Corpus

One of the obvious axioms of corpus linguistics is that any word frequency lists that are generated from a corpus will be a direct reflection of the texts in that corpus. In the case of the original GSL, there are many words on the list which, while arguably useful for second language learners of the time, seem a bit dated for the needs of today's learners. For example, the GSL contains many nautical terms

(oar, vessel, merchant, sailor, etc.), agricultural terms (plow, mill, spade, cultivator, etc.), religious terms (devil, mercy, bless, preach, grace, etc.) as well as many other terms that seem less likely to occur frequently in texts that the modern second language learner would likely use in the classroom (telegraph, chimney, coal, gaiety, shilling, etc.). As much as my colleagues and I were in awe of how much West was able to accomplish without the benefits of computers, digital text files, scanning equipment, or powerful corpus analysis software, we felt that the GSL was long overdue for an update and hoped that the application of modern technology to a more modern corpus could result in a NGSL that offered better coverage with fewer words. We were lucky enough to be given full unrestricted access to the CEC, a multibillion word corpus that contains both written and spoken text data for British and American English as well as to the Cambridge Learner Corpus, a 40 million word corpus made up of English exam responses written by English language learners, and promised by Cambridge University Press that whatever list we derived from their corpus could be made available to the public for free. We began our development of the NGSL in early 2010, using both the SketchEngine (2006) tools that Cambridge provided, as well as a wide range of other tools including publicly available tools such as Lawrence Anthony's very useful AntConc (http://www.antlab.sci.waseda.ac.jp/software.html) program as well as several specialized bits of software that we developed specifically for the purposes of this project.

The initial corpus we used was created using a subset of the CEC that was queried and analyzed using the SketchEngine Corpus query system (http://www.sketchengine.co.uk). The size of each sub-corpus that was initially included is outlined in Table 1.

The Newspaper and Academic sub-corpora were quickly eliminated for very similar reasons. First, although statistical procedures can be used to correct for minor differences in the size of sub-corpora, it was clear that the Newspaper sub-corpora at 748,391,436 tokens and the Academic sub-corpora at 260,904,352 tokens were dominating the frequencies and far too large for this kind of correction

Table 1. CEC Corpora Used for Preliminary Analysis of NGSL

Corpus	Tokens
Newspaper	748,391,436
Academic	260,904,352
Learner	38,219,480
Fiction	37,792,168
Journals	37,478,577
Magazines	37,329,846
Nonfiction	35,443,408
Radio	28,882,717
Spoken	27,934,806
Documents	19,017,236
TV	11,515,296
Total	1,282,909,322

Table 2. CEC Corpora Included in Final Analysis for NGSL

Corpus	Tokens
Learner	38,219,480
Fiction	37,792,168
Journals	37,478,577
Magazines	37,329,846
Nonfiction	35,443,408
Radio	28,882,717
Spoken	27,934,806
Documents	19,017,236
TV	11,515,296
Total	273,613,534

(a potential problem with the ONGSL since the variance between the largest and smallest corpus is 12 billion words). Second, both of these sub-corpora did not fit the profile of general English text types we were looking for with the Newspaper sub-corpus showing a marked bias toward financial terms and the Academic sub-corpus being from a specific genre not directly related to general English. As a result, both corpora were removed from the compilation.

Table 2 shows the sub-corpora that were actually used to generate the final analysis of frequencies. While smaller than the corpus described in Table 1, the corpus is still more than 100 times the size of the corpus used for the original GSL and far more balanced as a result.

The resulting word lists were then cleaned up by removing proper nouns, abbreviations, slang and other noise, and excluding certain word sets such as days of the week, months of the year and numbers (this proved to be a controversial decision and these word sets will be re-added in the 2.0 version of the list which is due out in early summer of 2014).

Then we used a sequence of computations to combine the frequencies from the various sub-corpora while adjusting for differences in their relative sizes. Specifically, we used Carroll's measure of dispersion (D_2), estimated frequency per million (U_m), and Standard Frequency Index (Carroll, Davies, & Richman, 1971) to combine the frequencies from the various sub-corpora while adjusting for differences in their relative sizes.

Finally, based on a series of meetings and discussions with Paul Nation about how to improve the list, the combined list was then compared to other important lists such as the original GSL, the BNC, and Corpus of Contemporary American English (COCA) to make sure important words were include or excluded as necessary.

2 NGSL Version 1.01

Though we were as careful and systematic as possibly in the process of developing the original NGSL, like West and his colleagues before us, we view the

release of the 1.0 version of the NGSL as no more than an interim list, representing the best research and development we could do in relative isolation, but with the next very important step being to release the NGSL publicly so that teachers and researchers around the world could begin to react to it, and give ideas and advice on how to improve it. To this end, most of 2013 was devoted to making the list and a variety of NGSL-related resources available via a dedicated website (www.newgeneraservicelist.org), publishing and presenting about the list at more than a dozen conferences around the world and creating a NGSL social media presence on websites such as on Facebook. Through these efforts and the excellent feedback and suggestions we have received from many experts, we are now releasing the 1.01 version of the NGSL both here and on the NGSL website. The net result of these changes will decrease the number of NGSL headwords by 17 from 2818 to 2801 with the following being the main changes made:

TWO WORDS ADDED:

- Insertion of TOURNAMENT, which was accidentally deleted in the initial analysis
- YEAH, which was originally counted as a derived form of YES, is now counted under its own headword

NINETEEN WORDS DELETED:

- Four numbers were deleted and moved the supplemental list
  - ZERO
  - BILLION
  - FIFTEEN
  - FIFTY
- The inflected parts of speech of pronouns were demoted and listed under their canonical objective pronoun.
  - HER was listed under SHE
  - HIM and HIS were listed under HE
  - ITS was listed under IT
  - ME and MY were listed under I
  - OUR and US were listed under WE
  - THEIR and THEM were listed under THEY
  - THESE was listed under THIS
  - THOSE was listed under THAT
  - WHOM and WHOSE were listed under WHO
  - YOUR was listed under YOU

3 What Constitutes a ''word'' in the NGSL?

There are many ways to define a word for the purpose of counting frequencies. The simplest is to look at ''types,'' where each form is counted as a different word regardless of part of speech. For example, LISTS would include both the third person singular form of the verb LIST and the plural form of the noun LIST.

The second method is to count "lexemes" where homographs are counted separately, but all the inflected forms of a word are added together. For example, the nouns LIST and LISTS would be counted together but not with the verbs LIST, LISTS, LISTED, and LISTING which would be counted separately. Inflections for nouns include the plural and the possessive. Verb inflections include the third person, the past, and the participles. Inflections for short adjectives include the comparative and the superlative.

The third method of counting words is called "word families" and was proposed by Bauer and Nation (1993). Word families include the inflected forms, and certain derived forms laid out by the generalizability and productivity of the affixes.

The NGSL uses a modified lexeme approach, where we count the headword in all its various parts of speech and include all inflected forms. Unlike the traditional definition of a lexeme, it includes all the inflected forms from the different parts of speech. For example, LIST would include LISTS, LISTED, LISTING, and LISTINGS. It does not include any of the derived forms using non-inflection suffixes. Variations such as the difference between US and UK spelling are also grouped within the same lexeme.

4 Text Coverage: Covering Your Bets with the NGSL

One of the most important goals of this project was to try to develop a NGSL that would be more efficient and useful to language learners and teachers by providing more coverage with fewer words than the original GSL. One of the problems with making a comparison between the two lists, indeed between any well-known vocabulary lists, is that the way of counting the number of words in each list needs to be done according to the same criteria. As innovative as the GSL was at the time of its creation, West's definition of what constituted a word was, by his own admission, nonsystematic and arbitrary: "no attempt has been made to be rigidly consistent in the method used for displaying the words: each word has been treated as a separate problem, and the sole aim has been clearness" (West, 1953, p. viii)

This means that for a meaningful comparison between the GSL and NGSL to be done, the words on each list need to be counted in the same way. As was mentioned in the previous section, a comparison of the number of "word families" in the GSL and NGSL reveals that there are 1964 word families in the GSL and 2368 in the NGSL (using level 6 of Bauer and Nation's 1993 word family taxonomy). Coverage within the 273 million word CEC is summarized in Table 3, showing that the 2368 word families in the NGSL provides 90.34% coverage while the 1964 word families in the original GSL provides only 84.24%. That the NGSL with approximately 400 more word families provides more coverage than the original GSL may not seem a surprising result, but when these lists are lemmatized, the usefulness of the NGSL becomes more apparent as the more than 800 fewer lemmas in the NGSL provide 6.1% more coverage than is provided by West's original GSL.

After analyzing coverage of the CEC corpus for the GSL and NGSL word lists, the next step taken was to compare coverage figures against other kinds of corpora I had at my disposal. In this round of analysis, I have also included the

Table 3. Comparison of Coverage for the CEC by the GSL and NGSL Word Lists

Vocabulary list	Number of "word families"	Number of "lemmas"	Coverage in CEC corpus (%)
GSL	1964	3623	84.24
NGSL	2368	2818	90.34

ONGSL in the analysis. All calculations were conducted using Lawrence Anthony's excellent AntWordProfiler, which easily allows for the uploading of vocabulary word lists and texts to be analyzed as long as they have been converted to .txt files. For this comparison, all word lists used were first converted to modified lemmas so that word counts would be done in the same way. A modified lemma is one that combines all possible parts of speech into one lemma. For example, the modified lemma for ROUND includes the inflections for the noun, verb, and adjective, e.g., ROUND, ROUNDS, ROUNDED, ROUNDING, ROUNDINGS, ROUNDER, and ROUNDEST.

Please note that the slight difference in number of word families and lemmas between the analysis done in early 2013 shown in Table 3 and the results given for this report in Tables 4 and 5 are due to the fact that the GSL in Table 3 was taken from the GSL/Academic Word List (AWL) version of the Range program (Heatley, Nation, & Coxhead, 2002). These lists were not specifically cited to have been developed up to Affix Level 6 (Bauer & Nation, 1993) while the lists from the BNC/COCA shown in Table 4 are. Therefore, the headwords from the GSL/AWL word lists were matched to the derived forms from the BNC/COCA lists.

The first corpus used was a 12 million word corpus of the top 100 most important classic works of English literature as rated by professors of English literature at several top Japanese universities (Browne & Culligan, 2008). All texts selected were ones that were available in the public domain for download and analysis via Project Guttenberg (2014). As a collection of classic literature texts (the newest texts available for download in Project Guttenberg are at least 50 years old), it was hypothesized that the word list which was based on the oldest corpus, the original GSL, would probably provide the highest coverage.

The second corpus was a more modern corpus of 27 million words taken from *The Economist* magazine, spanning issues from 2001 to 2010 (Culligan, 2013a). The third corpus, too, was also quite modern, a 13 million word sample taken from the *Scientific American* magazine covering issues published between 1993 and 2000 (Culligan, 2013b). Here it was hypothesized that one of the word lists based on more modern corpora (either the NGSL or the ONGSL) would provide more coverage.

As can be seen from Table 4, the GSL provided slightly better coverage (0.8%) than the NGSL for the corpus of classic literature and a more substantial 3.4% higher coverage than the ONGSL. That the GSL, which is based on a corpus with a far older collection of texts, provided the best coverage of a collection of older literary texts is perhaps an expected result, but a more surprising one was that the NGSL, which is based on a more modern corpus, was able to come close to 0.8% coverage of the GSL despite using 700 fewer lemmas.

Table 4. Comparison of GSL, NGSL, and ONGSL Coverage Figures for Three Different Genres

Word list	Number of headwords	Number of unique headwords	Number of types	Number of lemmas	Number of BNC/COCA word families	Classic	*Scientific American*	*The Economist*
GSL (Nation level 6)	1986	1927	9293	3553	2245	86.17%	65.87%	76.55%
ONGSL	2228	2189	6365	2130	1929	82.76%	68.68%	78.30%
NGSL 1.01	2801	2801	8481	2801	2483	85.35%	71.34%	81.75%
						12,377,844	13,047,726	27,337,358

Table 5. Coverage Figures for Two Well-known Novels

	Number of headwords	Number of unique headwords	Number of lemmas	Coverage of *The Count of Monte Cristo*	Coverage of *Dracula*
GSL range	1986	1927	3553	85.6	90.6
NGSL 1.1	2801	2801	2801	84.8	89.9
ONGSL	2228	2189	2130	82.2	87.4

If we narrow down the results for classic literature to look at coverage for two well-known novels within the corpus, *The Count of Monte Cristo* and *Dracula*, Table 5 shows very similar results with the GSL giving slightly better coverage than the NGSL (0.8% and 0.7% more coverage respectively), with the NGSL giving 2.5–2.6% more coverage than the ONGSL.

When looking at coverage figures for the two more modern genre-specific corpora, the efficiency of the NGSL becomes more apparent, with the NGSL giving 3.5% more coverage than the ONGSL and 5.5% more coverage than the GSL for the *Scientific American* corpus and similar figures of 3.5% and 5.2% more coverage for *the Economist* corpus.

5 Where to Find the NGSL and Associated Resources

From the very beginning, our focus has been less on simply publishing an academic paper on a new list of words than it has been on creating a list of high-frequency words that was as useful as possible for students, teachers, and researchers around the world and have thus created a dedicated website (www.newgeneralservicelist.org) to gather all associated NGSL resources in one place. Here, you can download the 1.01 (and 1.0) versions of the NGSL in lemmatized or headword form as well as all papers that have been written on the NGSL and see a list of past and upcoming conference presentations on the list. Because word lists are only useful to learners (and teachers) if there are definitions and learning tools, we have already written original definitions for all words in easy English for all NGSL words and uploaded the entire list in 50 word blocks (by frequency) to the free Quizlet vocabulary flashcard learning program (www.quizlet.com). As for analytical tools, the NGSL is already available on the free Online Graded Text Editor (OGTE) program (http://www.er-central.com/ogte/), which is part of the free extensive reading and listening website (www.er-central.com) developed by Charles Browne and Rob Waring as well as on Tom Cobb's wonderful VocabProfile tool (http://www.lextutor.ca/vp/eng/) and will soon also be available via Laurence Anthony's free AntWordProfiler Program (http://www.antlab.sci.waseda.ac.jp/antwordprofiler_index.html).

References

Bauer, L., & Nation, P. (1993). Word Families. *International Journal of Lexicography*, *6*, 253–279. doi:10.1093/ijl/6.4.253

Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*. doi:10.1093/applin/amt018

Browne, C. (2013, July). The New General Service List: Celebrating 60 years of vocabulary learning. *The Language Teacher*, *7 34*, 13–16. Retrieved from http://jalt-publications.org/tlt/issues/2013-07_37.4

Browne, C., & Culligan, B. (2008). *A collection of the top 100 works of classic English literature as rated by Japanese university professors*. Unpublished raw data.

Carroll, J. B., Davies, P., & Richman, B. (eds) (1971). *The American Heritage Word Frequency Book*. Boston, MA: Houghton Mifflin.

Culligan, B. (2013a). *A corpus of The Economist magazine articles from 2001–2010*. Unpublished raw data.

Culligan, B. (2013b). *A corpus of Scientific American articles from 1993–2000*. Unpublished raw data.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *RANGE and FREQUENCY programs*. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Project Guttenberg. (2014). Retrieved from http://www.gutenberg.org

West, M. (1953). *A General Service List of English words*. London, UK: Longman, Green & Co.

Vocabulary Learning and Instruction
Volume 3, Issue 2, September 2014
http://vli-journal.org

Article

Analyzing the Effectiveness of Textbooks for Vocabulary Retention

Rachael Ruegg and Cherie Brown
Akita International University
doi: http://dx.doi.org/10.7820/vli.v03.2.ruegg.brown

Abstract

Although many language educators are aware of factors necessary for vocabulary acquisition and retention, in many institutions around the world instructors are required to use textbooks as a basis for instruction, yet, to date, little has been done to analyze the effectiveness of textbooks to foster vocabulary acquisition and retention. Therefore, the purpose of the present study was to analyze the vocabulary content of a range of textbooks. The number of target words, frequency level of those words and length of reading texts were analyzed in reading texts and their associated activities of 20 English as a Second Language (ESL)/English as a Foreign Language (EFL) textbooks. It was found that each text had as little as no target words and on average 10, while the average text length was 639 words. In terms of the frequency level of the target vocabulary, although the textbooks claimed to be at an intermediate level or above, the frequency level of the target vocabulary was very inconsistent, with as many as 60% of target words coming from the first 1000 words or as many as 100% of target words being low frequency 'offlist' words. Furthermore, it was found that integrated skills textbooks had significantly fewer target words and significantly shorter reading texts than reading textbooks, while they also had a significantly higher percentage of words from the 1000 word list and a significantly lower percentage from the Academic Word List (AWL).

1 Background

A great number of studies have examined different factors necessary for vocabulary acquisition and retention, and many language educators are aware of these factors. For example, it is widely accepted that learners should begin their vocabulary acquisition by learning the most common words in the target language, as these provide greater coverage of written text and spoken language than other words. In the study of English vocabulary, in particular, the General Service Lists (West, 1953) of the most frequent 2000 words in the English language are the most commonly referred to for vocabulary instruction and research. After the first 2000 words have been acquired, it is generally agreed that the next focus should be the 570 academic words in the Academic Word List (AWL) for learners who hope to use the language for academic purposes (Coxhead, 2000).

However, in many institutions around the world, instructors are required by their institutions, or education ministries, to use prescribed textbooks as a basis for

Please cite this article as: Ruegg, R., & Brown, C. (2014). Analyzing the effectiveness of textbooks for vocabulary retention. *Vocabulary Learning and Instruction*, *3* (2), 11–18. doi: 10.7820/vli.v03.2.ruegg.brown

instruction. Indeed, many instructors who are not required to use textbooks also choose to do so. To date, textbooks have not been analyzed in terms of their effectiveness to foster vocabulary acquisition and retention. Therefore, the purpose of the present study was to analyze a range of textbooks which are widely available and commonly used internationally in the instruction of English as a Foreign Language (EFL), English as a Second Language (ESL) and English for Academic Purposes (EAP) with a view to discovering the extent to which they foster vocabulary acquisition and retention opportunities.

The terms 'beginner', 'elementary', 'intermediate' and 'advanced' have often been used to describe the proficiency level of learners, and these terms continue to be used in many contexts. Similarly, these terms are often used to describe the levels of textbooks. However, the terms are subjective and little effort has been made to determine what they might mean in quantifiable terms. Perhaps the closest the field of language education has to a consensus on language proficiency levels is the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001), which was created in an attempt to more objectively measure learners' language proficiency levels. However, many of the descriptors used in the CEFR are also difficult to quantify and therefore, in practice, the CEFR 'rests on the poles of teacher perceptions' (Hulstijn, 2007). More specifically, in terms of vocabulary knowledge, there is also considerable range in the amount of vocabulary known by learners based on their context of study. For example, EFL learners typically know less vocabulary than their ESL counterparts because of their relative lack of exposure to the target language.

Although the issue of how much vocabulary learners should be expected to know at a given proficiency level is somewhat opaque, textbook producers still need to decide what target vocabulary to focus on in their books. As a rough guide, graded reader series provide an estimate of how much vocabulary learners may know at various proficiency levels. The vocabulary levels of three series of graded readers (Cambridge, Macmillan and National Geographic) were used to determine a rough guide to the level of vocabulary that should be taught to learners at different levels of proficiency. From the data in Table 1, which were compiled from the information offered by the three publishers, it can be seen that beginners to pre-intermediate level learners should be focusing on the first 1000 words, elementary to intermediate level learners should be focusing on the second 1000 words and upper-intermediate to advanced level learners should be focusing on vocabulary beyond the first 2000 words. In terms of the CEFR, learners at the A1 and A2 levels should be focusing on the first 1000 words, from A2 to B2 should be focusing on the second 1000 words and from B2 upwards should be focusing on words beyond the first 2000 words.

Not only do learners need to focus deliberately on vocabulary during study, but they also need to encounter target words repeatedly, in order to remember them. According to Folse (2004), encountering a new word multiple times enables it to be moved from short-term to long-term memory and thus retained. Similarly, Cobb (1999) found that when learners encounter words in a variety of different contexts, they are better able to retain them. Consequently, not only do the target words need to be included as many times as possible in a text, but previously learned words also need to be recycled regularly (Nation, 2001).

Table 1. Graded-reader Vocabulary Levels

Publisher	Level	Words	CEFR
Cambridge	Beginner/elementary	400	A1
Cambridge	Elementary/lower-intermediate	800	A2
Cambridge	Lower-intermediate	1300	B1
Cambridge	Intermediate	1900	B1
Cambridge	Upper-intermediate	2800	B2
Cambridge	Advanced	3800	C1
Macmillan	Starter	300	
Macmillan	Beginner	600	A1
Macmillan	Elementary	1100	
Macmillan	Pre-intermediate	1400	A2–B1
Macmillan	Intermediate	1600	B1–B2
Macmillan	Upper	2200	
National Geographic	Pre-intermediate	800–1000	A2
National Geographic	Intermediate	1300–1600	B1
National Geographic	Upper-intermediate	1900–2200	B2
National Geographic	Advanced	2600–3000	C1

2 Research Method

This research analyzed the target vocabulary content of 20 textbooks which are widely available and commonly used internationally in the instruction of EFL, ESL and EAP. For the purposes of comparison, 10 reading textbooks and 10 integrated skills textbooks were selected and the fifth reading text of each book was analyzed. The textbooks used in the analyzis are listed in Table 2.

In order to ascertain the target vocabulary, one reading text and all vocabulary activities related to the text (pre-reading, post-reading, review and expansion activities) were examined. A word which appeared in the reading text and at least one activity was considered to be a target vocabulary item. The length of texts was compared, in order to determine the amount of exposure to text that learners gain from the textbook because text exposure gives learners the chance to encounter new words multiple times and also recycles previously learnt words. In the analyzis of the frequency level of the target vocabulary, the lexical profiling tool 'Web Vocabprofile' (Cobb, n.d.; Heatley, Nation, & Coxhead, 2002) was used. The tool calculates both the number of words at each of four frequency levels (1000 word level, 2000 word level, AWL and offlist) and the text coverage in the form of percentages.

3 Results

3.1 Preliminary Analyzis

The descriptive statistics for the length of texts (in words) and the number of target words can be seen in Table 3.

Table 2. Textbooks analyzed

Title	Author(s)	Publisher	Year	Level	Type
Aspire	Dummett, P., Robb Benne, R., Crossley, R.	Heinle Cengage Learning	2013	Upper-intermediate B2	Integrated skills
College Reading	Avery, J., Fellag, L.	Thomson Heinle	2006	Upper	Reading
Concepts and comments	Ackhert, P. Lee, L.	Heinle Cengage Learning	2005	Upper-intermediate	Reading
Focus on vocabulary 2	Schmitt, D. Schmitt, N.	Pearson Longman	2011	Upper-intermediate	Reading
Global Outlook	Bushell, B. Dyer, B.	McGraw Hill	2013	Intermediate	Reading
Just Right	Harmer, J. Lethaby, C.	Heinle Cengage Learning	2012	Upper-intermediate	Integrated skills
Key Concepts 2	Smith-Palinkas, B. Croghan-Ford, K.	Heinle Cengage Learning	2010	Intermediate	Reading
Mosaic 1	Wegmann, B. Knezevic, M.	McGraw Hill	2014	Intermediate – high-intermediate	Reading
New English File	Oxenden, C. Latham-Koenig, C.	Oxford	2008	Upper-intermediate B2	Integrated skills
New Headway	Soars, L. Soars, J.	Oxford	2005	Upper-intermediate B2	Integrated skills
Outcomes	Dellar, H. Walkley, A.	Heinle Cengage Learning	2010	Upper-intermediate B2	Integrated skills
Passages 1	Richards, J. Sandy, C.	Cambridge	2008	Upper-intermediate	Integrated skills
Q: Skills for Success 5	Caplan, N. Roy Douglas, S.	Oxford	2011		Reading
Quest 2	Hartmann, P.	McGraw Hill	2007	Intermediate– high-intermediate	Reading
Reading Explorer 4	MacIntyre, P.	Heinle Cengage Learning	2010	Upper-intermediate C1	Reading
Select Readings	Lee, L. Bernard, J.	Oxford	2011	Upper-intermediate B2	Reading
World Class 2	Douglas, N. Morgan, J.	Cengage Learning	2013	Advanced C1	Integrated skills
World English 3	Johannsen, K. Tarver Chase, R	Heinle Cengage Learning	2010	Intermediate B1	Integrated skills
World Link 3	Stempleski, S., Morgan, J., Douglas, N.	Heinle Cengage Learning	2011	Intermediate B2	Integrated skills

Table 2 (*Continued*)

Title	Author(s)	Publisher	Year	Level	Type
World Pass	Stempleski, S., Morgan, J., Douglas, N., Johannsen, K.	Thomson	2006	Upper-intermediate	Integrated skills

The descriptive statistics for the percentage of target words at each frequency level can be seen in Table 4.

While analyzing the text length, number and frequency level of target words, there appeared to be clear differences between the integrated skills textbooks and reading textbooks. For this reason, it was decided to conduct an additional analyzis, comparing the text length, number and frequency level of target vocabulary between the two types of textbooks.

Table 3. Descriptive Statistics for Text Length and Target Words

Variable	Mean	SD	Min	Max
Text length	639	313	243	1221
Target words	10	6	0	25

Table 4. Descriptive Statistics for Frequency Levels of Target Words

Frequency	Mean (%)	SD	Min	Max (%)
1000 words	17.17	19.64	0	60.00
2000 words	13.17	16.42	0	66.67
AWL	30.17	31.14	0	100
Offlist	34.50	27.10	0	100

3.2 Additional Analyzis

Table 5 shows the descriptive statistics for the text length (in words) and the number of target vocabulary items in both types of textbooks.

Independent samples *t*-tests were conducted to determine whether there were any significant differences between the two types of textbooks in terms of text length or number of target words. It was found that there was a significant difference (at the 0.05 level) between reading textbooks and integrated skills textbooks both in terms of the length of texts (in words), ($t = 8.86$, $df = 18$, $p = .001$) and the number or target words, ($t = 2.57$, $df = 18$, $p = .0019$).

Table 6 shows the descriptive statistics for the frequency level of the target words in both types of textbooks.

Table 5. Descriptive Statistics for Text Length and Target Words by Text Type

Variable	Text type	Mean	SD	Min	Max
Text length	Reading	860.90	289.89	432	1221
	Int. skills	417.00	116.10	243	616
Target words	Reading	12.90	6.59	6	25
	Int. skills	6.90	3.35	0	12

Table 6. Descriptive Statistics for Frequency Levels of Target Words by Text Type

Frequency	Text type	Mean	SD	Min	Max
1000 words	Reading	7.82	10.61	0	27.27
	Integrated skills	26.51	22.52	0	60
2000 words	Reading	14.65	20.80	0	66.67
	Integrated skills	11.68	11.48	0	33.33
AWL	Reading	48.22	34.17	0	100
	Integrated skills	12.12	12.51	0	40
Offlist	Reading	29.30	21.09	0	71.43
	Integrated skills	39.69	32.34	0	100

T-tests were conducted, using a Bonferroni adjustment, to compare the proportions of words from each frequency level (1000, 2000, AWL and Off-list) between the integrated skills textbooks and reading textbooks. It was found that there was a significant difference between the number of AWL words in the two types of books ($t = 3.137$, $df = 18$, $p < 0.0125$), with significantly more AWL words appearing in the reading textbooks than the integrated skills textbooks. The proportions of words at the other three frequency levels (1000, 2000 and Off-list) were not found to be significantly different between the two types of textbooks.

4 Discussion and Conclusion

Due to the lack of any clear, quantifiable proficiency levels or any guidelines regarding what vocabulary learners at different proficiency levels may be ready to learn, textbook producers are left with the difficult task of trying to determine what level of vocabulary learners at different proficiency levels should be focusing on. However, as shown in Table 1, looking at graded readers may provide a rough guide as to what levels are likely to be appropriate. It is clear that learners beyond the pre-intermediate or B1 level no longer need to focus on the first 1000 words as an object of direct study, yet 7 out of 10 integrated skills textbooks and 2 out of 10 reading textbooks focused on words from the first 1000 as target words in their textbooks. Furthermore, two of the integrated skills textbooks ('Just right' and 'Aspire'), which claimed to be at the upper-intermediate or B2 level, included 50% and 60% of target words from the first 1000 words, respectively. It seems that if learners need to focus

on these words at the highest frequency level, then they cannot be termed upper-intermediate level learners and, therefore, it is highly likely that the vocabulary level of these books is pedagogically inappropriate.

Although the claimed proficiency levels of the reading textbooks and the integrated skills textbooks were almost the same, with two of each type of textbook claiming to be at an intermediate level and the rest claiming to be at higher levels (only one book claimed to be at the advanced level and one book did not state any proficiency level), 6 of the 10 integrated skills textbooks focused predominantly on high-frequency levels (50% or more of the target vocabulary from within the first 2000 words). On the other hand, not one of the reading textbooks focused predominantly on high-frequency words (50% or more of the target words from within the first 2000 words). It seems clear that reading textbooks are expecting users to know significantly more vocabulary, although they use the same terms to describe the level of the textbook.

It may appear that reading textbooks and integrated skills textbooks have different purposes, and this argument may be used to justify the significantly higher proportion of academic vocabulary in the reading textbooks. However, textbooks may not always be used for the purposes for which they are intended. General ESL schools in English-speaking countries tend to use integrated skills textbooks rather than skill-specific texts. However, they have many learners who aim to be admitted into a formal tertiary education institution. Furthermore, even in pre-sessional EAP programs within universities in English-speaking countries, integrated skills textbooks are often used as the basis of instruction. Although it may be cheap and practical for learners to buy just one textbook and use it to learn all four skills, it seems clear in terms of vocabulary learning that institutions that choose integrated skills textbooks rather than a dedicated reading textbook may be doing a disservice to their students, as reading textbooks provide more recycling of vocabulary through longer texts, more focus on vocabulary and target vocabulary at more suitable frequency levels than integrated skills books.

What was stated by Hulstijn (2007) about the CEFR could also be said of the terms ‘beginner’, ‘elementary’, ‘intermediate’ and ‘advanced’. We already have a number of tests which can be given to individual learners to determine their vocabulary knowledge. However, this information has not been linked to proficiency levels. At this point, it seems necessary for vocabulary researchers to find a method of quantifying what learners are expected to know at each level so that textbook producers can clearly communicate the intended audience for their textbooks, enabling teachers and learners to select the material that is most appropriate to their needs.

References

Cobb, T. (1999). Breadth and depth of vocabulary acquisition with hands-on concordancing. *Computer Assisted Language Learning*, *12*, 345–360.

Cobb, T. (n.d.). *Web Vocabprofile*. Retrieved from http://www.lextutor.ca/vp/, an adaptation of Heatley, Nation, & Coxhead’s. (2002). *Range*.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

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

Folse, K. (2004). *Vocabulary myths: Applying second language research to classroom teaching*. Ann Arbor, MI: The University of Michigan Press.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). RANGE and FREQUENCY programs. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation

Hulstijn, J. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *Modern Language Journal*, *91*, 663–667. Retrieved from http://www.jstor.org/stable/4626094

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

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

Vocabulary Learning and Instruction
Volume 3, Issue 2, September 2014
http://vli-journal.org

Effects of Glosses and Reviewing of Glossed Words on L2 Vocabulary Learning through Reading

Makoto Yoshii
Prefectural University of Kumamoto
doi: http://dx.doi.org/10.7820/vli.v03.2.yoshii

Abstract

This study is an attempt to integrate incidental and intentional vocabulary learning in a reading activity without sacrificing the enjoyment of reading. The paper reports on a study which examined the effectiveness of a reading program on the web. The program contained glosses in a text and a reviewing component at the end of reading. The learners read the text for comprehension purposes on computers and were able to look up certain words by clicking on them. At the end of the reading the learners were also able to review the words they had looked up during the reading. This study examines how well learners can pick up words through this reading program. This study also examines the effectiveness of a reviewing activity by comparing the words reviewed and the words not reviewed. The study investigates if there are any differences in immediate and medium-term effects for vocabulary learning. Data from a pretest one week prior to the experiment, an immediate test right after the reading, and a delayed test were used for the analysis. Lookup behaviors of glosses and reviewing behaviors were also taken into account for analyzing the data.

1 Introduction

Integration of incidental and intentional learning has been one of the key issues in second language learning (Ellis et al., 2009; Hulstijn, 2005). It is of particular interest to second language vocabulary researchers (File & Adams, 2010; Hulstijn, 2001; Schmitt, 2008). Research indicates that L1 speakers incidentally acquire around 15% of the unknown words they encounter while reading (Swanborn & De Glopper, 1999). L2 learners can also learn vocabulary as they read for comprehension (Horst, Cobb, & Meara, 1998). L2 word learning through reading, however, often suffers from small gains, slow process, and requires repeated encounters with words (Waring & Takaki, 2003). Studies have examined ways to promote better efficiency to L2 incidental vocabulary learning. One of the prominent techniques is glossing and studies have shown its effectiveness, especially in the CALL environment (Yun, 2011). Studies have also pointed out that an intentional component is necessary in order to promote greater efficiency to incidental vocabulary learning (Laufer, 2005; Pellicer-Sanchez & Schmitt, 2010; Schmitt, 2008; Wesche & Paribakht, 2000). This study is an attempt to add an intentional component to incidental vocabulary learning in a reading activity. The intentional component was added with care so that it will not interfere with the

Please cite this article as: Yoshii, M. (2014). Effects of glosses and reviewing of glossed words on L2 vocabulary learning through reading. *Vocabulary Learning and Instruction*, *3* (2), 19–30. doi: 10.7820/vli.v03.2.yoshii

reading activity itself or reducing the pleasure of reading. The learners were foretold about the reading comprehension test but not about vocabulary test. In this sense, the notion of intentional learning in this study differs from that of other studies which regard intentional learning as learning with anticipation of some sort of test after an activity (Hulstijn, 2001). Intentional learning in this study means an additional exposure to words they encounter in the reading. The study also incorporated individualized learning by providing an opportunity to review, but not all the words in the text, but only those learners actually looked up during reading. Technology was utilized to conduct this study and the whole reading and testing procedures were implemented with an online reading program.

2 Theoretical Underpinnings of the Study

Glosses are definitions or explanations of difficult words provided in the margins of a text. Studies have reported the effectiveness of glosses for enhancing incidental L2 vocabulary learning using printed materials (Hulstijn, 1992; Hulstijn, Hollander, & Greidanus, 1996; Jacobs, Dufon, & Hong, 1994; Knight, 1994; Paribakht & Wesche, 1996; Watanabe, 1997). As technology advanced and became ubiquitous, glosses expanded gloss types to include multimedia elements such as graphics, videos, and audios. Studies have shown the effectiveness of multimedia glosses for L2 vocabulary learning (Abraham, 2007; Chun & Plass, 1996; Plass, Chun, Mayer, & Leutner, 1998; Yeh & Wang, 2003; Yoshii, 2006; Yoshii & Flaitz, 2002). In spite of a plethora of types of glosses available today, L2 learners often prefer simple textual information (Davis & Lyman-Hager, 1997; Erçetin, 2003; Laufer & Hill, 2000; Laufer & Kimmel, 1997; Lomicka, 1998; Pardo-Ballester & Rodríguez, 2010).

Are there any other ways to further enhance incidental L2 vocabulary learning besides glossing? Studies have clearly demonstrated that word-focused activities after reading promote more vocabulary gains than reading-only conditions do (Hill & Laufer, 2003; Sonbul & Schmitt, 2010; Paribakht & Wesche, 1996). According to noticing theory (Schmidt, 1990, 1992), attention is a prerequisite for learning. In order to promote the initial stage of form-meaning mapping, increased attention to words is crucial (Godfroid, Boers, & Housen, 2013; Laufer, 2005). Even though glosses could direct readers' attention to words and encourage the processing of word meanings during reading, word-focused tasks after reading are needed to bring better retention of words. Schmitt (2008) overviewed current L2 vocabulary learning research and acknowledged the importance of including both incidental learning and explicit, intentional learning. He also pointed out the importance of learners' engagement with word learning. This concurs with the Involvement Load Hypothesis (Laufer & Hulstijn, 2001) and reminds us of the importance of creating tasks which would promote more involvement of learners motivationally and cognitively.

Based on these theories, the researcher added both glossing and reviewing components to the reading program. The review part was unique in that it was an individually customized review of the words learners had looked up during reading. Each learner has a different list of the reviewed words based on his or her lookup behavior. It is hypothesized that this individualization would lead to more learner engagement and encourage more reviewing. As a result, this additional exposure and attention to words would promote better learning.

3 The Study

3.1 Purpose of the Study

The purpose of the study is to investigate the effectiveness of a reading program on the web which contains glossing and reviewing components. It is also to examine the effectiveness of customized vocabulary review in the reading program. The study examines the effectiveness of this additional review exercise for vocabulary learning by comparing the learning rates of the words reviewed and those not reviewed.

3.2 Research Questions

This study asks the following four research questions:

When learners read a text for comprehension,

(1) does the combination of glossing and reviewing promote vocabulary learning on the immediate test?
(2) does the combination of glossing and reviewing promote vocabulary retention on the delayed test?
(3) does reviewing of words they looked up during reading make a difference on the immediate test?
(4) does reviewing of words they looked up during reading make a difference on the delayed test?

3.3 Participants

The participants were 39 university students in Japan. They were freshmen majoring in English. Their average age was 19 years old and they had at least 6 years of English learning experience at junior and senior high schools.

4 Materials

A 300-word text was provided using the Moodle learning management system. This system had been used from the beginning of the term as a class management system. Therefore, the learners were familiar with the system. The text dealt with Jacques Rousseau, "One Gentle Writer Who Changed the World" as the title described. The readability of the text was 12th grade level with 41 points for the reading ease score measured by Flesch–Kincaid test. The text contained mostly words from the most frequent 1,000 word families (84%) and the second most frequent 1,000 word families (3%). This amounted to 87% of the words in the text. A small portion (5%) were from the Academic Word list (Coxhead, 2000) and the remaining words were not on the lists (8%). The text included 20 target words that were highlighted and clickable. The target words consisted of two from the first 1,000, two from the second 1,000, three from the Academic Word list, and 13 words from outside the lists. L1 glosses appeared as learners clicked on the words. Figure 1 shows the actual screenshot of the text and one word being looked up.

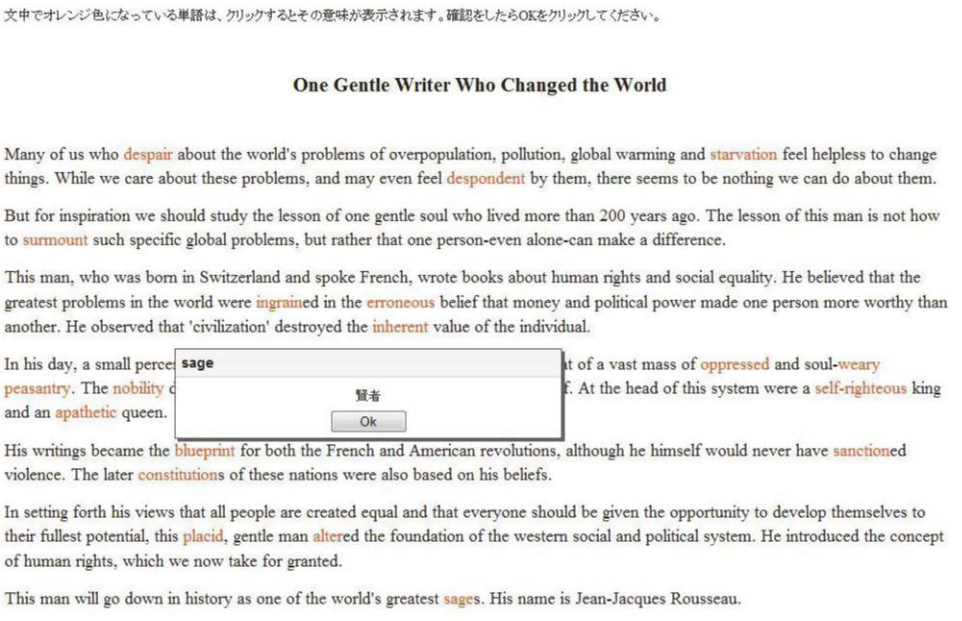

Figure 1. Actual Screenshot of the Text and a Word Being Looked Up.

The reading program stored learners' lookup behaviors and, at the end of the reading activity, provided a list of the words the learners looked up. The learners were able to review the meanings by clicking on the words as seen in Figure 2.

4.1 Procedures

All the activities were part of regular classes. They were language skill classes and particularly focused on reading and listening. A pretest was administered to the participants one week prior to the study. It was a multiple-choice test conducted on Moodle. The learners had to choose the best meaning out of four L1 meanings. The reading activity was provided at the next class a week later. The participants read the text on Moodle at the individual computers. Learners had a chance to review the words they clicked on at the end of the reading as in Figure 2. After the reading, a comprehension test was given as it had been told beforehand. Following the comprehension test, a vocabulary test (the immediate test) was unexpectedly administered. It was the same as the pretest except that the order of the questions and the choices in each question were randomly re-arranged. A delayed test was given two weeks after the immediate test. The test format was the same as the pretest and the immediate test. The order of the test items and choices were randomly re-arranged.

4.2 Data Analysis

Data from the pretest, the immediate test, and the delayed test were used for the analysis. For analyzing the data for the research questions 1 and 2, an ANOVA

Figure 2. Actual Screenshot of the Review Section.

was used to examine the differences among the tests. When the differences were significant, post hoc analysis was conducted. The unit of analysis for research questions 3 and 4 was each individual target word. For the analyses, first, the words the learners already knew prior to this study shown at the pretest were excluded. From the remaining unknown words, only those the participants actually looked up during reading were selected. In order to compare the reviewed and non-reviewed words, a chi-square test was run on these words. The test examined whether two groups of words, reviewed and non-reviewed, would differ in terms of numbers of correct and incorrect words on the immediate and the delayed tests.

5 Results

5.1 Research Questions 1 and 2: Short-term and Long-term Effects of the Reading Program

Figure 3 displays the results of the three vocabulary tests: the pretest, the immediate, and the delayed tests. The participants were able to identify almost half

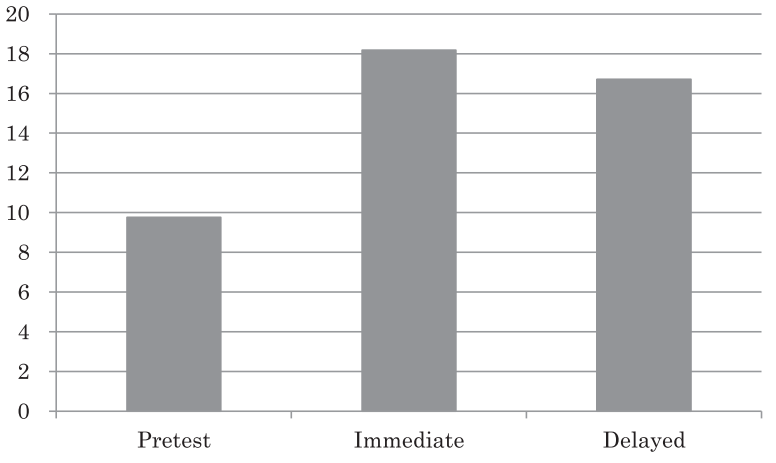

Figure 3. Scores of Three Vocabulary Tests.

of the 20 target words at the pretest ($M = 9.8$, SD $= 2.2$). The immediate score nearly doubled the score in the pretest ($M = 18.2$, SD $= 2.1$). The participants sustained the score on the delayed test with a small loss ($M = 16.7$, SD $= 2.5$). An ANOVA showed statistical significances among the tests ($F(2, 76) = 230.48$, $p < .01$). Post hoc analysis using a Bonferroni adjustment ($\alpha = .0167$) revealed that the scores of both immediate and delayed tests were significantly higher than those of the pretest ($p < .001$). The combination of glossing and reviewing promoted vocabulary learning through reading as seen on the immediate test and helped the learners retain words as seen on the delayed test.

5.2 Research Questions 3 and 4: Short-term and Long-term Effects of Reviewing

The unit of analysis for research questions 3 and 4 was each individual target word. Since the study had 20 target words for each of 39 participants, this amounts to 780 words. Figure 4 recaptures the results of the pretest. The participants already knew 381 words out of 780 words (49%) in the study, and these were excluded from the analysis.

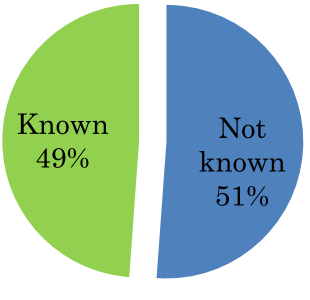

Figure 4. Overview of Pretest Results.

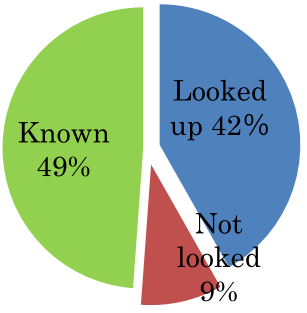

Figure 5. Overall View of Lookup Behavior during Reading.

From the remaining 399 words (51%), the words the participants actually clicked were selected. Figure 5 shows the overall results of the participants' lookup behaviors of the glossed words. Among all the 780 words in the study, 542 words were looked up (70%). This rate of lookup exceeds the rate of unknown words at the pretest (51%) and indicates that the students tried to look up even those words they got correct at the pretest. It seems that the students were trying to confirm or verify the meanings of those words. For the analysis of this study, we will use only the looked-up words among unknown words, which amounted to 326 words (42%).

Then, this group of words, which were unknown at the pretest and looked up during the reading, were further divided between those which were reviewed and those which were not. As Figure 6 indicates, 231 words (30%) were reviewed and 95 words (12%) were not reviewed.

In order to compare the reviewed ($n = 231$) and non-reviewed ($n = 95$) words, a chi-square test was run in relation with the results of the immediate and the delayed tests. The test examined whether two groups of words, reviewed and non-reviewed, would differ in terms of numbers of correct and incorrect words on the immediate and the delayed tests.

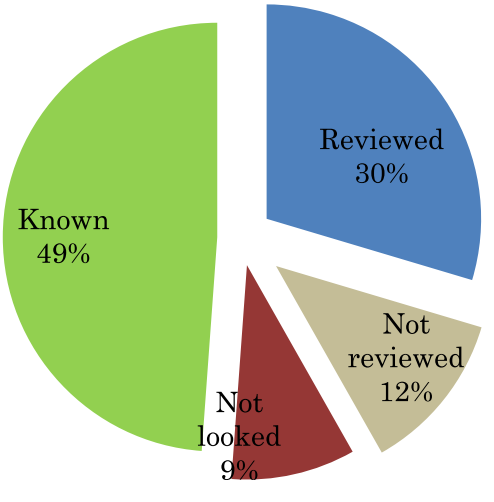

Figure 6. Overall View of the Lookup Behavior during Reviewing.

Table 1. Cross-tabulation of Group (Reviewed and Not Reviewed) and Results of Immediate Test (Correct and Incorrect)

Group	*n*	Correct	%	Incorrect	%
Reviewed	231	218	94.4	13	5.6
Not reviewed	95	89	93.7	6	6.3

Table 2. Cross-tabulation of Group (Reviewed and Not Reviewed) and Results of Delayed Test (Correct and Incorrect)

Group	*n*	Correct	%	Incorrect	%
Reviewed	231	181	78.4	50	21.6
Not reviewed	95	77	81.1	18	18.9

As in Table 1, the results of the immediate test showed that 218 reviewed words turned out to be correct answers; while 13 reviewed words were incorrect. On the other hand, among 95 words not reviewed, 89 words lead to correct answers while six words to incorrect. The ratio of the correct and incorrect answers for both reviewed and non-reviewed words were almost identical. The chi-square test did not show any significant differences between the two groups: $\chi^2(1) = .06$, $p = .81$. This indicated the high pick-up rate (94%) of word learning occurred on the immediate test regardless of reviewing.

As in Table 2, the results of the delayed test showed that 181 reviewed words (78.4%) remained as correct answers while 50 reviewed words (21.6%) became incorrect. This number of incorrect answers was a big increase from 13 words (5.6%) on the immediate test. Among the words not reviewed ($n = 95$), 77 words (81.1%) stayed as correct while 18 words (18.9%) turned out to be incorrect. The rapid increase of incorrect answers in both groups (around 20%) indicated the memory loss over two weeks between the immediate and delayed tests.

In order to compare the reviewed and non-reviewed in terms of numbers of correct and incorrect words on the delayed test, a chi-square test was run. The ratio of the correct and incorrect answers on the delayed test was also very similar between reviewed and non-reviewed words. The result did not show any significant differences between the two groups: $\chi^2(1) = .30$, $p = .59$. This indicated that, although word knowledge had already started to decline (the incorrect answers increase from around 5% on the immediate test to 20% on the delayed test), the retention rate of the words learned was still high (around 80%), regardless of reviewing.

6 Findings and Conclusions

The results of the immediate test showed significant gains from the pretest and the learners were able to sustain most of the gains on the delayed test. This was another confirmation of the effect of glossing in reading activity and an indication of the effect of the additional reviewing activity. However, the results of both the

immediate and the delayed tests did not show any significant differences between the two groups of words: reviewed words and non-reviewed words.

One reason for these results may have to do with a ceiling effect. The immediate result showed a very high learning rate (94%) and there was not much room to show any differences between reviewed and non-reviewed words.

A second reason is related to the sensitivity of the instrument. This study used the multiple choice of meaning as the instrument, and it was most likely not enough to detect the change in word knowledge. Laufer, Elder, Hill, and Congdon (2004) proposed the various levels of word knowledge in the form-meaning link and Schmitt (2010) relabeled the description as following four elements: form recall, meaning recall, form recognition, and meaning recognition. If other types or levels of word knowledge in the form-meaning linking stage were tested besides meaning recognition, the differences might have emerged.

A third reason is related to the nature of the reviewing task. The reviewing activity in this study was a simple check on a meaning of each word by clicking. This was purposely designed so that the activity would not impose on the learners a rigorous exercise after the reading activity. The log file of the participants indicated that it was just a matter of a couple of minutes to go through the reviewing. Even if they were exposed to the words again, it might not have been enough to help the participants strengthen their word learning. The reviewing might have not had enough quality to promote the depth of processing necessary for effective learning.

The reason may also be related to the fact that there may not have been a need for reinforcement. The participants had already gained enough during reading with glossing. Glossing might have had a greater impact on word learning in both tests than expected. Regardless of words reviewed or not reviewed, the correct answer rates were extremely high in both tests: 94% for the immediate and 80% for the delayed as seen in Tables 1 and 2. It could be that the glossing had already helped the learners make an initial connection between form and meaning, which is a necessary part of word learning (Laufer, 2005).

The effectiveness of reviewing needs to continue to be investigated. Many questions still remain unanswered. How can we effectively integrate incidental and intentional parts into vocabulary acquisition? What are the ways to strengthen word knowledge learners gain while reading without reducing enjoyment in reading? How can we add a reviewing part to incidental vocabulary learning effectively but not intrusively? Further studies are needed to find answers to these questions. Future research could also look at how the participants look up glossing and how they review words. Are there any strategies used by learners? If so, what are they? Qualitative study such as a think-aloud protocol may shed some light on these. Future research could also look at the effectiveness of reviewing, adapting the strict sense of intentional learning. A study could be devised which compared two groups: one group with the announcement of a vocabulary test after the reading activity and the other group without such an announcement. The announcement of the test could be delivered after the incidental stage of reading to determine the effectiveness of reviewing as an intentional component in its full sense.

Acknowledgments

This study was partially supported by Grant-in-Aid for Scientific Research (C) [grant number 23520691]. I would like to express my sincere gratitude to Professor Kita at Kumamoto University for creating this reading program, especially the individualized reviewing component. I would also like to express my appreciation to Dr. Laufer for her valuable comments on this study.

References

Abraham, L. B. (2007). Second-language reading comprehension and vocabulary learning with multimedia. *Hispania*, *90*, 98–108.

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

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal*, *80*, 183–198. doi:10.1111/j.1540-4781.1996.tb01159.x

Davis, J. N., & Lyman-Hager, M. A. (1997). Computers and L2 reading: Student performance, student attitudes. *Foreign Language Annals*, *30*(1), 58–72. doi:10.1111/j.1944-9720.1997.tb01317.x

Ellis, R., Loewen, S., Elder, C., Erlam, R., Philp, J., & Reinders, H. (2009). *Implicit and explicit knowledge in second language learning, testing and teaching*. Bristol, UK: Multilingual Matters.

Erçetin, G. (2003). Exploring ESL learners' use of hypermedia reading glosses. *CALICO Journal*, *20*, 261–283.

File, K. A., & Adams, R. (2010). Should vocabulary instruction be integrated or isolated? *TESOL Quarterly*, *44*, 222–249.

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition*, *35*, 483–517. doi:10.1017/S02722 63113000119

Hill, M., & Laufer, B. (2003). Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *IRAL – International Review of Applied Linguistics in Language Teaching*, *41*(2), 87–106. doi:10.1515/iral.2003.007

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. Retrieved from http://nflrc.hawaii.edu/rfl/PastIssues/rfl112horst.pdf

Hulstijn, J. H. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 113–125). London, UK: Macmillan.

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. 258–287). Cambridge, UK: Cambridge University Press.

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning. *Studies in Second Language Acquisition*, *27*(2), 129–140. doi:10.1017/S0272263105050084

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, *80*, 327–339. doi:10.1111/j.1540-4781.1996.tb01614.x

Jacobs, G. M., Dufon, P., & Hong, F. C. (1994). L1 and L2 vocabulary glosses in L2 reading passages: Their effectiveness for increasing comprehension and vocabulary knowledge. *Journal of Research in Reading*, *17*(1), 19–28. doi:10.1111/j.1467-9817.1994.tb00049.x

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *The Modern Language Journal*, *78*, 285–299. doi:10.1111/j.1540-4781.1994.tb02043.x

Laufer, B. (2005). Focus on form in second language vocabulary learning. *EUROSLA Yearbook*, *5*, 223–250. doi:10.1075/eurosla.5.11lau

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21*, 202–226. doi:10.1191/0265532204lt277oa

Laufer, B., & Hill, M. (2000). What lexical information do L2 learners select in a CALL dictionary and how does it affect word retention? *Language Learning & Technology*, *3*, 58–76. Retrieved from http://llt.msu.edu/vol3num2/laufer-hill/index.html

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26. doi:10.1093/applin/22.1.1

Laufer, B., & Kimmel, M. (1997). Bilingualised dictionaries: How learners really use them. *System*, *25*, 361–369. doi:10.1016/S0346-251X(97)00028-6

Lomicka, L. L. (1998). To gloss or not to gloss: An investigation of reading comprehension online. *Language Learning & Technology*, *1*, 41–50. Retrieved from http://llt.msu.edu/vol1num2/article2/default.html

Pardo-Ballester, C., & Rodríguez, J. C. (2010). Developing Spanish online readings using design-based research. *CALICO Journal*, *27*, 540–553. doi:10.11139/cj.27.3.540-553

Paribakht, T., & Wesche, M. (1996). Enhancing vocabulary acquisition through reading: A hierarchy of text-related exercise types. *The Canadian Modern Language Review*, *52*, 155–178.

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

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second language multimedia learning environment. *Journal of Educational Psychology*, *90*(1), 25–36. doi:10.1037/0022-0663.90.1.25

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*, 129–158. doi:10.1093/applin/11.2.129

Schmidt, R. (1992). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, *13*, 206–226. doi:10.1017/S0267190500002476

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, *12*, 329–363. doi:10.1177/1362168808089921

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke, UK: Palgrave Macmillan.

Sonbul, S., & Schmitt, N. (2010). Direct teaching of vocabulary after reading: Is it worth the effort? *ELT Journal*, *64*, 253–260. doi:10.1093/elt/ccp059

Swanborn, M. S. L., & De Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, *69*, 261–285. doi:10.3102/00346543069003261

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. Retrieved from http://nflrc.hawaii.edu/rfl/October2003/waring/waring.html

Watanabe, Y. (1997). Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition*, *19*, 287–307. doi:10.1017/S027226319700301X

Wesche, M. B., & Paribakht, T. S. (2000). Reading-based exercises in second language vocabulary learning: An introspective study. *The Modern Language Journal*, *84*, 196–213. doi:10.1111/0026-7902.00062

Yeh, Y., & Wang, C. W. (2003). Effects of multimedia vocabulary annotations and learning styles on vocabulary learning. *CALICO Journal*, *21*, 131–144. Retrieved from http://journals.sfu.ca/CALICO/index.php/calico/article/view/669

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology*, *10*, 85–101. Retrieved from http://llt.msu.edu/vol10num3/yoshii/default.html

Yoshii, M., & Flaitz, J. (2002). Second language incidental vocabulary retention: The effect of text and picture annotation types. *CALICO Journal*, *20*, 33–58. Retrieved from http://journals.sfu.ca/CALICO/index.php/calico/article/view/626

Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning*, *24*(1), 39–58. doi:10.1080/09588221.2010.523285

Vocabulary Learning and Instruction
Volume 3, Issue 2, September 2014
http://vli-journal.org

Article

Effects of Instruction on Yes–No Responses to L2 Collocations

Junko Yamashita
Nagoya University
doi: http://dx.doi.org/10.7820/vli.v03.2.yamashita

Abstract

The lexical decision task (LDT), in which a participant makes dichotomous judgments on target letter strings, is an established method in psycholinguistic research to investigate the mental lexicon. With the expansion of research interests from single lexeme to collocations, second language (L2) researchers have started to use a similar judgment task at a phrasal level (referred to as a phrasal decision task or PDT in this paper). However, unlike the LDT, the PDT has not yet established a standard form of prompt, and variation has been observed in previous L2 studies. Hence, the purpose of this study was to examine effects of varying instructions on PDT performance. Three instructions (*acceptable*/*commonly used*/*natural*) were tested with Japanese university students and native speakers of English, who were asked to make judgments on English word combinations. Examining responses to congruent (felicitous both in Japanese and English), incongruent (felicitous only in Japanese or English), and baseline items, the study identified some effects of instruction differences. However, these effects were not so strong as to obscure the expected cross-linguistic congruency effect. Therefore, this result has led to the conclusion that researchers have more freedom of instruction selection in the PDT, at least among the three examined in this study and to the extent that the congruency effect was measured by accuracy scores.

1 Introduction

The lexical decision task (LDT) is an established method in psycholinguistic research to gauge various research questions in order to understand the construct of the mental lexicon. In its most standard format, a participant looks at a letter string on a computer screen and makes a dichotomous judgment regarding whether it is a real word or not. Yes-items in this task (items that require an affirmative response) consist of real words, whereas No-items (those that require a negative response) comprise pseudowords (artificially made letter strings that usually adhere to spelling-sound correspondence rules). Both reaction times and error rates are measured because they provide different types of information. In brief, the error rate is an index of the degree of knowledge representation and the reaction time is an index of the efficiency of processing.

With the advancement of research from single lexeme to multiword units, second language (L2) researchers have started to use a similar judgment task at a

Please cite this article as: Yamashita, J. (2014). Effects of instruction on Yes–No responses to L2 collocations. *Vocabulary Learning and Instruction*, *3* (2), 31–37. doi: 10.7820/vli.v03.2.yamashita

phrasal level (Jiang & Nekrasova, 2007; Wolter & Gyllstad, 2013; Yamashita & Jiang, 2010). This task is called the phrasal decision task (PDT) in this paper. In the PDT, Yes-items are felicitous (real, existing, and legitimate) phrases (e.g., *as soon as, take a bath*) and No-items are syntactically and/or semantically infelicitous word combinations (e.g., *is sun the*, *hear blood*). Unlike the LDT that has a standard form of prompt (*Is the letter string a real word or not?*), variations have been observed in the instruction for the PDT or a similar task in L2 research practice. For example, Laufer (2003) embedded collocations in a sentence (e.g., *I closed the TV.*) and asked whether sentences were *correct*. Jiang and Nekrasova (2007) examined formulaic sequences (e.g., *as a result*) and inquired whether word sequences were *grammatical*. Yamashita and Jiang (2010) and Wolter and Gyllstad (2013) both researched collocations, but the former asked whether word combinations were *acceptable* and the latter asked whether they were *commonly used*. This variation may be due to this task's relatively new history in the L2 research, but instructional differences may have impacts on judgmental performances. The primary purpose of this paper is to examine effects of different instructions on L2 PDT.

To proceed, it is imperative that the target structure be clarified because there are distinct types of multiword units. This paper focuses on collocations: lexical patterns that are determined by conventions within the language rather than by a set of rigid rules. Countering idioms, collocations are looser lexical patterns in that each constituent word can co-occur naturally with many other words (e.g., *heavy stone, heavy rain, fine rain*). However, the combination is not totally free and the language imposes restrictions that are sometimes arbitrary and difficult to explain (e.g., *high mountain* vs. *tall mountain*). Thus, collocations are normally acquired through exposure to the target language rather than by applying memorized rules. In addition to the importance of input, past studies indicated the lasting influence of first language (L1) on the acquisition of L2 collocations (Wolter & Gyllstad, 2011, 2013; Yamashita & Jiang, 2010). In Yamashita and Jiang's study, for example, lower proficiency L2 learners were slower and made more errors in their responses to incongruent collocations (those that do not have L1 equivalents) than to congruent collocations (those that have equivalents in L1); higher proficiency L2 learners, on the other hand, also made more errors in incongruent collocations but they improved their reaction times and no measurable difference was observed between congruent and incongruent collocations. Thus, the L1–L2 congruency affects the acquisition of L2 collocations. This paper calls this L1 influence a congruency effect.

To investigate the effects of instruction on the L2 PDT, the current study drew on the congruency effect as a benchmark. Namely, if the PDT has successfully demonstrated the congruency effect, it is interpreted that the instruction has validly served its purpose and detected the impact of collocation congruency on the acquisition of L2 collocations. In so doing, the present study has expanded the congruency effect to infelicitous L2 word combinations that are felicitous in their translation in L1. For example, *sweet judgment* (*amai handan*) and *wide mind* (*hiroi kokoro*) represent such word combinations in the case of L1-Japanese and L2-English. When such expressions are compared with baseline word combinations, it is expected that the former would be more incorrectly identified as felicitous L2 expressions as a result of L1 influence. Although this is a likely scenario on the basis of the documentation of L2 learners' production of unnatural L2 collocations

that could be traced back to their L1 (e.g., Laufer & Waldman, 2011; Nesselhauf, 2005), the congruency effect has not been identified with the PDT by using this type of item. Hence, it was of additional interest to see the congruency effect on infelicitous L2 word combinations. Unlike mainstream psycholinguistic research, however, this study only measured accuracies. Therefore, the congruency effect is limited to accuracy measures in this paper.

The following research question guided the study: Does the form of the instruction affect the congruency effect? Three formulations of instruction were tested (*acceptable, commonly used,* and *natural*) because these were all plausible forms of instruction to investigate collocational knowledge with the PDT.

2 Method

2.1 Participants

Participants were 114 L1-Japanese university students learning English as a foreign language (EFL) and 24 native speakers of English (NS). The NS group consisted of undergraduates, postgraduates, and teachers of English at universities, who resided in Japan at the time of testing. Their self-evaluation of Japanese proficiency on a scale from 0 (no knowledge) to 10 (native speaker level) averaged 5.7 ($SD = 2.1$).

2.2 Instruments

Linguistic stimuli were adjective–noun combinations, which were either adopted from past studies (Yamashita & Jiang, 2010; Wolter & Yamashita, 2014) or created for this study. To ensure that all words would be known to the participants, the majority of constituent words were within the most frequent 3000-word level in the JACET List of 8000 Basic Words (the standard high school level in the Japanese education system). There were four categories of items with 24 stimuli in each: JE (felicitous expressions both in English and Japanese), E-only (felicitous in English but infelicitous in Japanese when they are literally translated), J-only (infelicitous in English but felicitous in Japanese when they are literally translated), and baseline (infelicitous either in English or in Japanese). Table 1 shows sample items in each category.

Items were controlled for five features: lengths, adjective frequencies, noun frequencies, collocation frequencies, and MI scores (values of the last four features were obtained from the Corpus of Contemporary American English). The entire set of 96 items was divided into three lists, each of which was combined with one of the instructions. There were no significant differences in the five item features across the lists, either. The order of instruction was counterbalanced, and items were randomized within a list. The following instructions were used. They were in English for the NS and in Japanese for the EFL learners.

Do you think the following 32 expressions are acceptable in English?
Do you think the following 32 expressions are commonly used in English?
Do you think the following 32 expressions are natural in English?

Table 1. Sample Items

Category	Sample items
E-only	bad debt, brief moment, careful attention, early bird, flat rate, slow learner, wet weather
JE	true love, cold tea, deep sleep, front door, economic growth, tragic death, high salary
J-only	dark society, frank question, light content, quiet war, straight heart, weak rain, wide mind
Baseline	abstract fruits, medical key, male table, formal ice, nervous papers, wild pen, calm eggs

The EFL learners additionally provided their confidence rating on each item and answered a posttest open-ended questionnaire on how they made judgments according to different instructions. Due to space limitations, however, only some of the questionnaire answers are discussed in this paper.

2.3 Procedure

The test was given collectively to the EFL learners in their English classes and individually to the NS.

3 Results

Descriptive statistics of test results is summarized in Table 2. After square root and log transformations were applied to the EFL and NS data, respectively, to improve distribution where necessary, 3 × 2 (instruction × item type) repeated-measures ANOVAs were run with two response types (Yes/No) and two participant groups separately. When the sphericity was violated, Greenhouse-Geiser correction was adopted.

For the real English collocations (Yes-items), ceiling effects were evident in NS's responses (all > 90%), which lead, unsurprisingly, to insignificant effects of any independent variables. The EFL learners, in contrast, displayed lower and more varied accuracies across conditions (50 ~ 81%), and three significant effects were identified: a main effect of instruction, $F(2, 214) = 89.49$, $\eta_p^2 = .44$, $p < .001$, a main effect of item type, $F(1, 113) = 375.67$, $\eta_p^2 = .77$, $p < .001$, and the interaction between instruction and item type, $F(2, 226) = 83.41$, $\eta_p^2 = .43$, $p < .001$. According to Bonferroni-adjusted post hoc multiple comparisons, the EFL learners were consistently more accurate on the JE than E-only items, demonstrating the expected congruency effect. Moreover, among the three instructions under the E-only condition, they were more accurate with the instruction containing *acceptable* (66%) than with the other two (52%, 50%).

With regard to illegitimate phrases (No-items), NS's accuracy decreased in the J-only items (53 ~ 76%), which resulted in a main effect of instruction,

Table 2. Descriptive Statistics of Accuracy

		EFL			NS		
	Item type	Mean	SD	% Mean	Mean	SD	% Mean
Yes-items							
Acceptable	E-only	5.27	1.83	.66	7.63	0.71	.95
	JE	6.01	2.02	.75	7.71	1.04	.96
Commonly used	E-only	4.18	1.56	.52	7.75	0.44	.97
	JE	6.39	1.41	.80	7.75	0.53	.97
Natural	E-only	4.00	1.85	.50	7.33	1.61	.92
	JE	6.47	1.35	.81	7.50	1.14	.94
No-items							
Acceptable	J-only	4.44	1.76	.55	5.17	2.51	.65
	Baseline	6.39	1.44	.80	7.00	1.69	.88
Commonly used	J-only	4.66	1.86	.58	6.04	1.68	.76
	Baseline	6.57	1.47	.82	7.54	0.93	.94
Natural	J-only	4.70	1.55	.59	4.21	1.84	.53
	Baseline	6.64	1.48	.83	7.42	1.47	.93

Note: The maximum achievable score is 8.

$F(2, 46) = 5.85$, $\eta_p^2 = .20$, $p = .005$, a main effect of item type, $F(1, 23) = 295.21$, $\eta_p^2 = .93$, $p < .001$, and the interaction between the two, $F(2, 46) = 5.27$, $\eta_p^2 = .19$, $p = .009$. Post hoc analysis indicated that the NS were consistently more accurate on the baseline (88 ~ 94%) than J-only items, suggesting the impact of Japanese collocations. In the J-only condition, they were significantly more accurate with the instruction containing *commonly used* (76%) than with that using *natural* (53%). Their accuracy with the instruction including *acceptable* (65%) was located between these, and manifested no significant difference from either of them. As for the EFL learners, only a main effect of item type was significant, $F(1, 113) = 532.09$, $\eta_p^2 = .83$, $p < .001$, indicating that they responded more correctly to baseline items (80 ~ 83%) than J-only items (55 ~ 59%) under all instructions.

4 Discussion

Throughout this study, the EFL learners displayed the congruency effect by both wrongly rejecting the E-only items and wrongly accepting the J-only items. The NS were, unsurprisingly, highly accurate in accepting the real English collocations, but interestingly they were also wrongly accepting J-only items more than the baseline, suggesting an effect of Japanese collocations. Although the result from J-only items aligned with the expectation for the EFL group, why did the NS say yes to non-existing English word combinations? A most likely reason is the influence from their L2 to L1 (Laufer, 2003) because all NS in this study had been exposed to the Japanese language. A piece of informal evidence is an NS participant who correctly suspected that some items were direct translations from Japanese, which suggests that, at least for this person, Japanese collocations were activated during the task in English. However, the story may not be as simple as this. In Wolter and Yamashita (2014), NS who were not exposed to the Japanese

language demonstrated the Japanese influence by reacting to J-only items more quickly than the baseline. Although the item type and the task were distinct between the studies, it may be the case that lexical patterns that exist in one language may make more sense than artificially made non-sense word combinations. This question, however, is beyond the scope of this paper and should be further explored in the future.

Instruction differences affected the magnitude of accuracy on two occasions. First, the EFL learners replied more affirmatively (thus correctly) to the E-only items with the instruction containing *acceptable* than that containing either *commonly used* or *natural*. Precise reasons are not known but the questionnaire responses aligned with this trend. The number of students who replied that their judgments did not differ across different instruction conditions was 23 (all three instructions), 2 (*acceptable* and *commonly used*), 2 (*acceptable* and *natural*), and 26 (*commonly used* and *natural*). These figures suggest that *commonly used* and *natural* tended to instigate similar judgment processes among the students. Second, the NS responded more affirmatively (thus wrongly) to the J-only items with the instruction using *natural* than that including *commonly used*. Again precise reasons are not known, but this might be because the NS had been daily exposed to English lexical patterns that the Japanese tend to mistakenly produce and may have developed such tolerance as to perceive them to be *natural*, but they may still somehow feel more strongly that such word combinations are not quite *commonly used* by native speakers.

Despite the differences caused by varying instructions, the congruency effect stayed intact in the EFL group. This result suggests that researchers could choose more freely from among these instructions when their purpose is to examine the congruency effect on the acquisition of L2 collocations, and lend support to the previous studies that used different instructions with the PDT but identified the congruency effect (Yamashita & Jiang, 2010; Wolter & Gyllstad, 2013).

However, this discussion needs to be received with caution because instruction differences showed some effects on the performance, depending on the type of items and participants. These minor differences within the realm of this study may have serious effects depending on the purposes of future studies. Moreover, the current study only investigated accuracy scores, thus it has no implication for reaction times. A similar study examining reaction times is recommended.

5 Conclusion

The purpose of this study was to examine the effects of different instructions used with the PDT, a relatively new task that has appeared in L2 psycholinguistic research. With the growing interest in the acquisition of collocations, this task may become one of the standard tasks in the field. This study indicated that, despite some differences, all three tested instructions captured the congruency effect when it was measured by accuracy scores. Hence, at least for this type of data, researchers have freedom in instruction selection for examining the congruency effect. In addition, this study demonstrated the congruency effect not only in Yes-items as did previous studies, but also in No-items, thus it has expanded the scope of linguistic items where the congruency effect makes its appearance. Evidence of the

congruency effect in infelicitous L2 word combinations is limited and further scrutiny in this area may help deepen our understanding of the acquisition of L2 lexical patterns.

References

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, *91*, 433–445. doi:10.1111/j.1540-4781.2007.00589.x

Laufer, B. (2003). The influence of L2 on L1 collocational knowledge and on L1 lexical diversity in free written expression. In V. Cook (Ed.), *Effects of the second language on the first* (pp. 19–31). Clevedon: Multilingual Matters.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, 61, 647–672. doi:10.1111/j.1467-9922.2010.00621.x

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam, the Netherlands: John Benjamins.

Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, 32, 430–449. doi:10.1093/applin/amr011

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing. *Studies in Second Language Acquisition*, 35, 451–482. doi:10.1017/S0272263113000107

Wolter, B., & Yamashita, J. (2014). Processing collocations in a second language: A case of first language activation? *Applied Psycholinguistics*. doi:10.1017/S0142716414000113

Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly*, 44, 647–668.

Vocabulary Learning and Instruction
Volume 3, Issue 2, September 2014
http://vli-journal.org

Article

Vocabulary in a Second Language: Selection, Acquisition, and Testing: A Commentary on Four Studies for JALT Vocabulary SIG

Batia Laufer
University of Haifa
doi: http://dx.doi.org/10.7820/vli.v03.2.laufer

Abstract

Four papers by Charles Browne, Rachael Ruegg & Cherie Brown, Makoto Yoshii and Junko Yamashita were presented in the morning session of the Third Annual JALT Vocabulary SIG Vocabulary Symposium in Fukuoka, Japan on June 14, 2014. As discussant, it is my pleasure to comment upon each manuscript. These lexical researchers originate from all over Japan: Tokyo, Akita, Kumamoto and Nagoya. Their lexical topics are related to three themes that are central to vocabulary research: selection, acquisition and testing. The papers are concerned with the types of words that should be selected for teaching, with the optimal conditions for vocabulary acquisition and with instruments that measure lexical proficiency, or are used in lexical research. After commenting on each paper in turn, I shall present a few suggestions for their future research.

1 Introduction

Ten years ago, the late Paul Bogaards and I edited a book *Vocabulary in a Second Language* (Bogaards & Laufer, 2004). The papers in the book covered three main themes: selection, acquisition and testing. The authors of the papers were concerned with the types of words that should be selected for teaching, with the optimal conditions for vocabulary acquisition and with instruments that measure lexical proficiency, or are used in lexical research. It seems that these three themes are still central to vocabulary research in light of more recent publications in general (e.g. Nation and Webb, 2011; Schmitt, 2010) and this symposium in particular. The four papers I will comment on address precisely the above three important themes.

Vocabulary selection is of vital importance since L2 learners do not have the time and the opportunity to learn words in the same way native speakers do, via language input. Therefore, attempts have been made by L2 learning researchers and educationists to select a manageable quantity of the most important vocabulary for learners' needs. One principle behind vocabulary selection is the cost–benefit principle, which states that learners should get the best return for their learning effort, i.e. learn words that they will meet and use often. The first two papers (Browne and Ruegg & Brown) address the question of selecting the appropriate words for instruction and practice for specific learners. Because of the paucity of input that learners have in a non-L2 environment, an important source

Please cite this article as: Laufer, B. (2014). Vocabulary in a second language: selection, acquisition, and testing: a commentary on four studies for JALT vocabulary SIG. *Vocabulary Learning and Instruction*, *3* (2), 38–46. doi: 10.7820/vli.v03.2.laufer

of classroom vocabulary acquisition is word focused instruction. Such instruction combines drawing attention to specific words, in authentic, communicative activities, as well as practicing decontextualized vocabulary. Both types of activities can occur in incidental or intentional learning conditions. The third paper (Yoshii) deals with acquisition of new words through specific contextualized and decontextualized word-focused activities. Valid tests of lexical knowledge are essential when we want to assess learners' achievement and proficiency, or evaluate research results. Though numerous measurement instruments are available for testing single words, fewer tests are available to test multiword units. The fourth paper (Yamashita) investigates a question related to testing collocations.

2 The Four Studies

2.1 A New General Service List: The Better Mousetrap We've Been Looking for? By Charles Browne

As stated above, the cost–benefit principle in vocabulary learning states that learners should get the best return for their learning effort, i.e. learn words that they will meet and use often. However, what is frequent in a language may change from time to time following changes that occur in the language. Therefore, word frequency lists need to be updated. This is what Browne has done in his study. He produced an updated General Service List that is based on millions of modern text words from the Cambridge English Corpus (CEC), excluding academic and newspaper language, but including learner language. The reason for the exclusion is that the academic and newspaper subcorpora are larger than the other corpora and might therefore have biased the general CEC word frequency towards academic and newspaper words.

I wonder whether the academic subcorpus should have been excluded. Many academic words are found among the most frequent 2000 words. For example, 33 words from the first sublist of the Academic Word List (AWL; the 60 most frequent AWL words) and 25 from the second sublist of 60 AWL words are from the first 2000 words. Maybe the final corpus from which the new GSL (NGSL) was drawn could have included a sample of the academic subcorpus that is similar in size to the other subcorpora. In this way, the size of the inclusion of the academic subcorpus would not have skewed the frequency results, but academic vocabulary would have been duly represented. My suggestion is to compare the NGSL and the AWL lists to see how many 2000 academic words are in the NGSL. It would also be interesting to run the NGSL through the academic subcorpus of CEC and see how well it covers it. The authors may find that there are enough academic words in the list to provide good coverage without including the academic subcorpus in the list.

As for the inclusion of learner language in the corpus, I wonder how different it is from samples of native speakers' language in terms of vocabulary use. Learners' selection of words is not necessarily based on word frequency. For example, speakers of Romance languages are known to overuse infrequent Latin vocabulary because of its similarity to their L1. Learners may avoid frequent words in their writing if they are uncertain about some features of their use (grammatical or semantic). It would be interesting to analyze word frequency in the learner

subcorpus and compare it to the larger corpus. If the profiles are suspiciously different, the learner subcorpus could be taken out.

The NGSL defines a word as a 'modified lexeme'. Different parts of speech are not separated if they are spelled in the same way. So 'to list' and 'list' are counted as the same modified lexeme. But different parts of speech that are spelled differently, i.e. with different morphemes, are not counted together (avoid/ avoidance). Moreover, according to this definition, if words have different meanings in different parts of speech, they are still considered to be the same modified lexeme, e.g. 'account' (v) in 'account for the differences in the results' and 'account' (n) in 'bank account'. However, the two meanings would not be in the list if one of them is not in the most frequent 2000 modified lemmas. The 'modified lemma' principle raises the question whether frequency should be the only principle for selecting when to teach which vocabulary. It seems to me that learners who learn 'avoid' would easily learn 'avoidable', 'unavoidable' and 'avoidance', but these derivatives would not be taught if they are of different frequencies and do not appear on the NGSL. The two meanings of 'account', on the other hand, are more difficult to learn even though they are one modified lemma since the lemma represents two completely different meanings. If the frequency lists are intended for learners, we may try to take into account the principle of ease and difficulty of learning as well.

The most important feature of any frequency list is its coverage and here is the main strength of the NGSL. The NGSL covers more of the CEC than the GSL (90.34% as opposed to 84.24%). One may argue that if the lists were constructed on the basis of the CEC, the most frequent words of the list would necessarily cover a large portion of the corpus. But the paper presents coverage figures of different texts as well of older novels, and more modern texts of Scientific American and Economist. For the modern non-fiction texts, NGSL provides better lemma coverage than the other lists, about 5% more than GSL and 3% more than ONGSL. (However, if we measure coverage in terms of word families, the advantage is less clear since there are about 300 more families than in the GSL.) Even if the differences in lemma coverage are small, small differences in coverage may lead to substantial results in comprehension (Laufer and Ravenhorst-Kalovski, 2010). We found that the difference in about 3% (from 87 to 90) leads to an increase in 8% of comprehension.

The paper can be related to several general questions about the construction of frequency lists, degrees of coverage, vocabulary learning and testing.

- How much overlap is there between lists of the most frequent vocabulary (BNC-based, COCA-based, GSL, NGSL, the Longman defining vocabulary, Longman Communication 3000, the JACET 8000 list)? My own experience shows that many words have different frequencies in British National Corpus (BNC) and Corpus of Contemporary American English (COCA). What is the implication of such differences for setting teaching priorities and constructing tests?
- What is the relationship between increase in coverage and increase in comprehension? An important goal of constructing the best possible lists is to improve comprehension via improving lexical coverage.
- What definition of a word (by lemmas or word families) is preferable in list construction in the context of vocabulary teaching? The list of families

includes more lemmas, but is there much more learning effort than with a lemma-based list if the additional lemmas include common derivatives?
- How can we construct vocabulary size tests with representative samples from each word frequency if the same words appear in different frequencies in different corpora and hence in different frequency lists?

2.2 Analyzing the Effectiveness of Textbooks for Vocabulary Retention by Rachael Ruegg and Cherie Brown

Vocabulary selection is also the central theme of the second paper. The authors examine a series of textbooks to see whether the vocabulary that is supposed to be important and necessary for learners at a certain proficiency level is indeed selected for practice by book writers.

The paper starts with examples showing that the widely used labels for proficiency levels – beginner, intermediate and advanced – are unclear and can mean different things in terms of vocabulary knowledge. Particularly important is the observation about the Common European Framework of Reference for Languages (CEFR) levels. As vocabulary size is not specified in the CEFR guidelines, the CEFR proficiency levels may be interpreted in different ways, as Table 1 shows (see Ruegg and Brown, p. 13, of this issue).

A sample of 20 textbooks was analyzed in terms of text length and "target words", i.e. words that appeared in the text and at least in one exercise. One can wonder whether one text sample is representative of the entire book. The authors may consider taking several (three or four) random samples per book in several (three or four) books of each type (reading and integrated skills), and show that there is no difference between the texts within each book in the number of target words and their frequencies. This may increase the validity of the sample.

I will not relate to text length results because I do not consider text length very important, unless it can be shown that the target words appear more times in the longer texts. The paper assumes this is indeed so. (This can easily be checked in a follow up study.) Since the textbook writers claim that the books are written for an intermediate level or above, the assumption is that at least the 2000 most frequent words are familiar (according to Table 1), and therefore the books should be focusing on the "beyond 2000" vocabulary. I will comment on two interesting results. The first shows how well, or badly, the books focus on the beyond 2000 words. The second result is the number of target words the books have selected to practice in a text.

The reading textbooks include 77% of beyond 2000 words among the target words and the integrated skills textbooks include 52%. If you want to make a statistical comparison between the two types of book, I would not compare them at each frequency level, but at the vocabulary level the books are supposed to teach: the 'beyond 2000' vocabulary, that is all the vocabulary that is not included within the first 2000 most frequent words. If you think that second 1000 vocabulary should be taught to the intermediate learners, then you can compare the 'beyond 1000' vocabulary. You will probably find that that reading books have a larger number of

these words. To get back to the percentage of the 'beyond 2000' words, the average statistics do not look too bad. Even if 23% of target words in the textbooks are from the first 2000 words that are familiar to the learner, the practice of such words may provide information on additional word features thus deepening the already existing word knowledge. Furthermore, focus on known words acts against forgetting them, and forgetting is a common and troublesome phenomenon in vocabulary learning. In other words, just as practicing new vocabulary is necessary for learning, practicing familiar words is beneficial for vocabulary maintenance. But the mean percentages of the 'beyond 2000' target words obscure the results of the individual books. Some of the books seem inadequate for the intended learners in terms of their target vocabulary as there seem to be too many high-frequency words.

The importance of the analysis carried out by the authors is that it demonstrates the lack of principled selection of target vocabulary in widely used textbooks and the confusion that the terms beginners, intermediate and advanced introduce. (It is unclear whether there is any other guiding principle for vocabulary selection in these textbooks.) One implication of this analysis is that each textbook should be examined in terms of the target vocabulary for suitability for particular learners before it is used in the classroom. Another implication, or rather suggestion, is not to rely on the above proficiency labels in reporting and interpreting research. In my own research, whenever I describe my participants, I state their proficiency in terms of vocabulary size on the Levels Test.

The second result, the number of target words in each text, is not discussed in the paper and I would suggest looking at it closely in a follow-up study. The authors have kindly provided me with the numbers of texts in each of the books they analyzed. The average number of texts in a reading book is 20; the average number of texts in an integrated skill book is 14. If an integrated skills book offers seven words for practice per text and only half of them are new to the learners, then about $3.5 \times 14 = 49$ new words are expected to be learnt from the entire book. A reading book includes 20 texts, each text has 13 target words and 10 of them are new to the learners (see percentages of beyond 2000 words in Table 3, Ruegg and Brown, p. 15, of this issue). In this case, 200 new words will be introduced in the book. If in a course of 15 weeks students complete one book (personal communication with the authors), the question is whether additional 49–200 new words (assuming all the target words have been learnt) can be regarded as sufficient progress in the course.

Textbook analysis of the kind the authors have performed could be an essential component of course planning and should be linked to setting vocabulary teaching goals. Since learners at different learning stages are supposed to reach specific vocabulary sizes, suitable textbooks should be chosen with these targets in mind.

2.3 Effects of Glosses and Reviewing of Glossed Words on L2 Vocabulary Learning Through Reading by Makoto Yoshii

This paper is concerned with acquisition of new words through specific word-focused activities during online reading. The first activity is looking up word

meaning in an online gloss, the second activity is revision, which is supposed to provide an additional focus on some of the looked up words that are left to the learner's choice. The results confirmed what we know from previous studies, that the use of glosses is beneficial for learning. Out of 10 words that were available for learning, 8.2 were learnt and 6.7 retained on a delayed test. The new feature of the paper is the revision activity. The author makes use of log files to examine individual patterns of looking up words and reviewing some of them.

I was impressed by the seriousness of the participants. In the look up stage, they looked up more words than were unfamiliar to them, thus verifying what they had already known. In the reviewing session, fewer words were focused on, altogether 364 words. Three hundred and twenty-six of them were unfamiliar on the pretest. Even though students were not told what the purpose of the reviewing activity was, they were serious enough to follow the teacher's instructions and review some of the words. Two hundred and thirty-one of all unfamiliar words were reviewed, i.e. 71%. Hence the degree of reviewing was not bad. My guess, which could be tested in a follow-up study, is that in other cultures, the look up and reviewing behavior might be different. (See, for example, Laufer and Yano, 2001 for a comparison of Japanese, Chinese and Israeli learners.) The results are presented not only in tables, but in grids (Figures 4–6, see Yoshii, pp. 24–25, of this issue), which is an interesting visual display I was not familiar with.

And yet the reviewing activity did not affect the acquisition scores. The relatively easy test (recognition of meaning) and the possible ceiling effect (94% were learnt regardless of reviewing) are, in my opinion, secondary reasons. The main reason could be the lack of goal specification of the reviewing activity. The learners were not asked to review the words for any specific reason, such as a task completion, or an upcoming test. An obvious reason for a learner to review a list of words is a test in which these words may be tested. In my own work (Laufer, 2006), I found an impressive increase of scores in the same learners between two stages of learning: an incidental stage and a following stage of intentional learning that involved reviewing the target words for a test. This leads us to a thorny issue of defining the differences between incidental and intentional learning. Was there an intentional learning condition in the study presented in the paper? The author intended to introduce such a condition by the reviewing task. However, according to Hulstijn (2001), intentional learning means deliberate memorization of words for a specific purpose. The most obvious one is an upcoming test, even though one can decide to commit words to memory for other reasons as well. If there is no intention to memorize vocabulary, then learning is incidental. If, however, we examine the literature on incidental and intentional learning, we will often find conflicting operationalizations of the two types of learning. In many papers, "intentional" is taken to be word focused, or practiced in non-communicative activities, e.g. writing sentences with the target words. This is probably the definition that was adopted in the paper. However, this definition is fraught with problems which are beyond the scope of this commentary. Suffice it to say that if the participants in the study had been told there was a test following the reading activity, they would probably have reviewed the words more thoroughly than they did.

I think the paper is a good preliminary study that can be followed by another study with some changes: more target words to avoid the ceiling effect, a more

challenging test or even two tests, e.g. active recall and passive recall that will yield a better spread of scores, and test announcement in order to introduce an intentional learning component. If a within-subject design is chosen, then the test announcement should come after the incidental stage. If a between-subject design is preferred, then one group could be assigned to each condition. If a mixed design is chosen (and this would be my preference), then one group would be assigned to the intentional condition only and the other to an incidental condition followed by an intentional one. This would show word gains of each condition separately and the increase that the addition of deliberate memorization would add to the incidental learning.

2.4 Effects of Instruction on Yes-No Responses to L2 Collocations by Junko Yamashita

The paper addresses an issue in research methodology – whether the form of a prompt, the precise question about phrasal expressions, can affect the answer. Participants were asked to judge phrasal expressions as 'acceptable', 'commonly used' or 'natural' in English. The slight differences in the accuracy results did not overshadow the congruency effect that was observed in other studies. Collocations that were incongruent with L1 yielded lower number of correct responses. The author concludes, therefore, that all three types of instruction can be used in studying interlingual influences of phrasal language.

I would prefer to call them 'prompt types' rather than 'types of instruction' since the phrases 'effect of instruction' and 'types of instruction' are often associated with 'instruction' in the sense of 'teaching'. Even though the three prompts are claimed to be equally legitimate, learners provided more accurate answers to 'E-only' (correct in English, but not in its Japanese word for word translation) items when the prompt was 'acceptable' rather than 'natural' or 'commonly used', and native speakers were more accurate on J-only items (correct in Japanese but not in its English word for word translation) with the 'acceptable' prompt. The author says that the reason for this difference in the effect of the prompt is unclear. My hunch is that the term 'acceptable' is a more neutral or less demanding term than the other two, as it addresses lexical knowledge only. 'Acceptable' seems to be similar to 'correct', both for learners and native speakers. For L2 learners, the terms 'commonly used' or 'natural', on the other hand, could be associated with pragmatic knowledge as well. Since lexical knowledge of a word or expression is often acquired without its pragmatic nuances by L2 learners, it is easier to answer correctly to the 'acceptable' prompt. For native speakers, 'commonly used' or 'natural' could be associated with common usage, which is not necessarily identical with correctness. Therefore, when studying L1–L2 congruency effect, the three prompts would probably yield similar results, as was the case in the paper. However, if the purpose is to study some features of the depth of knowledge that L2 learners developed, such as sensitivity to frequency or naturalness of a phrase, or to study L2 influence on L1 that native speakers start experiencing after living in L2 environment, my preference would be for 'natural' or 'commonly used' prompts.

Even though the purpose of the paper was to examine the effect of different prompts on phrasal lexical decisions, the other important finding is the congruency effect. The importance of interlingual influence in L2 learning is not as widely acknowledged and incorporated into teaching as it should be. Psycholinguistic experiments of the kind we witness in the paper corroborate the findings of error analysis, elicitation studies and experimental studies that demonstrate the pervasive role that familiar languages, particularly L1 have on learning an additional language.

3 Some Suggestions for Future Research

The four papers in the symposium are related to more general research issues in selection, acquisition and testing that can be explored in further studies.

In the domain of selecting vocabulary for language syllabi and tests, we should ask ourselves whether word frequency is indeed the only criterion for selection. From the learner's point of view, a crucial factor in L2 acquisition regardless of word frequency is word "learnability", i.e. the ease or difficulty with which a particular word can be acquired. For example, an infrequent word may be a cognate in the learner's L1, or a loan word (if the languages are genetically unrelated). This makes such words easy to learn. Similarly, some derivatives of a frequent word may be infrequent, but they too are not difficult to learn. On the other hand, a frequent word that has no semantic equivalent in L1 is difficult. Further research could explore the effect of lists based on frequency word learnability on expanding the lexical repertoire of the learners.

Learnability is an important theme of the acquisition domain. There is still work to be done on the effect of interlingual factors on vocabulary learning, by studying a particular L1 group of learners and by comparing different L1 groups, by exploring the facilitating effect of easy words and finding effective methods for teaching difficult words or chunks of words. For example, cognates and loan words are easy to learn and introducing them in large number can enrich learners' functioning in a foreign language and consequently positively affect their confidence and motivation. (If these words changed their original meaning and became 'false friends', learners should be made aware of this.) I am not aware of any empirical work that measures the effect of teaching numerous cognates/loan words to low-level learners, possibly because many EFL classes have learners with different L1s. But such a study can be carried out in monolingual classes in Japan. Intentional learning, as defined by Hulstijn (2001), has not been sufficiently explored. Most studies I know of investigate the intentional learning of *new* words. But intentional learning can also be examined as a method of consolidating the knowledge of words that have been practiced in incidental learning tasks, particularly difficult words or multi-word units like collocations. Research on collocations has received a lot of attention recently. Serious research depends on good and original tests of collocations. Unlike the other existing tests, the phrasal lexical decision can measure fluency. Though there are studies comparing the knowledge of single words and collocations, I am not familiar with research comparing the fluency of the two. Lexical Decision Task (LDT) and Phrasal Decision Task (PDT) tests can provide us with the necessary tools for this kind of research. I hope that the JALT symposium, the papers and the discussions

will contribute to the already growing body of vocabulary studies conducted by scholars who work in Japan with Japanese learners.

References

Bogaards, P., & Laufer, B. (Eds.). (2004). *Vocabulary in a second language: Selection, acquisition and testing*. Amsterdam, the Netherlands: Benjamins.

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, UK: Cambridge University Press.

Laufer, B. (2006). Comparing focus on form and focus on FormS in second language vocabulary learning. *Canadian Modern Language Review*, *63*, 149–166. doi:10.3138/cmlr.63.1.149

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. Retrieved from: http://nflrc.hawaii.edu/rfl/April2010/articles/laufer.pdf

Laufer, B., & Yano, Y. (2001). Understanding unfamiliar words in a text: Do L2 learners understand how much they don't understand. *Reading in a Foreign Language*, *13*, 549–566. Retrieved from: http://nflrc.hawaii.edu/rfl/PastIssues/rfl132laufer.pdf

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

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York, NY: Palgrave Macmillan.

Vocabulary Learning and Instruction
Volume 3, Issue 2, September 2014
http://vli-journal.org

Estimations of Japanese University Learners' English Vocabulary Sizes Using the Vocabulary Size Test

Stuart McLean[a], Nicholas Hogg[b] and Brandon Kramer[c]
[a]*Temple University, Japan;* [b]*Osaka Yuhigaoka Gakuen High School;*
[c]*Momoyama Gakuin University*
doi: http://dx.doi.org/10.7820/vli.v03.2.mclean.et.al

Abstract

Measuring students' lexica is time-consuming, as one sitting of the Vocabulary Size Test (VST) usually takes 40–60 minutes. As a result, teachers would benefit from being able to make reasonable estimates from commonly available information. This paper aims to investigate: (1) What are the mean vocabulary sizes of students at Japanese universities as a whole, and by university department (*hensachi*)? and (2) Are a university's department standardized rank scores (*hensachi*) a useful proxy for English vocabulary size? This study used a cross-sectional design where 3,449 Japanese university students were tested using Nation and Beglar's VST. The results showed an average score of 3,715.20 word families and that VST scores were significantly higher for students in higher department *hensachi* programs. This current department *hensachi* was also found to have a stronger correlation with VST scores than with other covariates when the entire sample was considered. Lastly, there appears to be a lack of consistent knowledge of the most frequent words of English, suggesting that curriculum designers at Japanese universities should focus on teaching high-frequency English words. Although the findings support the use of the VST for comparing receptive written vocabulary knowledge between learners, they perhaps do not support its use in establishing a vocabulary size to decide lexically appropriate materials.

1 Background and Aim

Measuring students' lexica is time-consuming, as one sitting of the Vocabulary Size Test (VST) (Nation & Beglar, 2007) usually takes 40–60 minutes. As a result, teachers would benefit from being able to make reasonable estimates from commonly available information.

Previous research on the vocabulary size of Japanese students is limited. Shillaw (1995) and Barrow, Nakanishi, and Nishino (1999) suggested that the vocabulary size of non-English-major Japanese university students is between 2,000 and 2,300 word families. In these studies, vocabulary knowledge was assessed over 3,000 word families, with students completing self-checking familiarity surveys. However, this approach may have simply measured the word forms students recalled being exposed to rather than measuring receptive reading vocabulary knowledge. This is important as Waring and Takaki (2003) demonstrated that the knowledge thresholds for recognizing a previously encountered word and correctly

Please cite this article as: 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

answering a receptive vocabulary knowledge item differ, and instructors are probably more interested in the latter type.

Based on the need for more recent and thorough research, the following research questions were developed:

(1) What are the mean vocabulary sizes of students at Japanese universities as a whole, and by university department rank (*hensachi*)?
(2) Are a university's department standardized rank scores (*hensachi*) a useful proxy for English vocabulary size?

2 Methodology

2.1 Instruments

To estimate students' receptive vocabulary knowledge, Nation and Beglar's (2007) VST was used. The test is based on the spoken portion of the British National Corpus (BNC), ordered by word frequency and grouped into 1,000 word bands up to the 20,000 word level. For this research, however, only the first 8,000 word families were tested, measured with 80 items, each representing knowledge of 100 words. Tests were completed in one of the two formats: traditional paper tests and online tests administered through Survey Monkey <surveymonkey.com>. Both tests were the same version of the VST and were presented with the same instructions according to the test authors (Nation & Beglar, 2007) in order to minimize variance between the two mediums. The online test was advantageous administratively in that it allowed for easy marking and statistical analysis. It was also noticed by the researchers that students who conducted the paper tests were more inclined to not answer some questions, particularly as the difficulty increased, representing a weakness in the study.

In addition to the VST, a questionnaire was given to participants to collect personal information such as year, major, study abroad experience, native language, available Test of English for International Communication (TOEIC) and Test of English as a Foreign Language (TOEFL) scores, and previous *hensachi* data. A *hensachi* is a score assigned to individual students or school departments based on student performance in a national test standardized to the national mean. 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). The *hensachi* data collected in the questionnaire refer to standardized tests the students took toward the end of high school, while the department *hensachi* scores used to address the research questions were found online through the website of Benesse, a large testing company in Japan <http://shinken.zemi.ne.jp/hensachi>. In this study, a university's department *hensachi* is the mean *hensachi* of the first-year students in a given department who took the university entrance exam and Benesse's proficiency test. It does not include those who entered the university through other means, such as recommendations. Unfortunately it was not feasible to similarly verify the sources of the students' previous *hensachi* scores obtained through the questionnaire, as each specific score varies based on the particular

exam board where it was calculated. Lastly, the TOEFL test exists in a number of formats: the internet-based test (iBT), computer-based test (CBT), and paper-based test (PBT). As a result, it was necessary to extrapolate the equivalent PBT scores from self-reported iBT and CBT scores using a conversion table provided by the makers of the TOEFL test (the Educational Testing Service, 2005).

2.2 Participants and Research Design

This study used a cross-sectional design where data from 3,449 Japanese university students were collected through a snowball sampling approach, thanks to the assistance of many university instructors who kindly agreed to administer the test in their classrooms after requests in regional newsletters, at local events, and through personal correspondences. The tests were generally completed at the beginning of term during April and May of 2012 and 2013; however, a small number of tests were completed at the start of the autumn term in September and October of 2013.

The VST scores of 22 participants were identified as outlying the sample based on the calculation of their Rasch person ability estimates using the WinSteps statistical program. Participants with scores at or beyond 3.29 standard deviations from the mean based on their ability estimates were removed from the study, in line with the recommendation of Field (2009). Closer inspection revealed that these outliers were mostly the result of participants not answering a large number of the items when taking the test. Table 1 shows a summary of the participants divided by year, major, and department *hensachi*.

Table 2 shows that the mean written receptive vocabulary size (VST score $\times$ 100) was found to be 3,715.20 word families for the remaining participants (n = 3,427). A visual inspection of the histogram of the VST scores, shown in Figure 1, indicates that the data are normally distributed. For the purpose of comparison, participants were separated into three *hensachi* groups: $\geq$61, 51–60, and $\leq$50. The correlation coefficients were computed between the participants' VST scores and available TOEIC, TOEFL, individual *hensachi*, and department *hensachi* scores.

Table 1. Summary of Participants

	Hensachi (Department)	1st year	2nd year	3rd year	4th year
English Majors	≥61	n = 1			
	51–60	n = 76	n = 46	n = 3	
	≤50	n = 101			
Science Majors	≥61	n = 320	n = 3	n = 1	n = 2
	51–60	n = 224	n = 228	n = 98	
	≤50	*n = 184*	*n = 193*	*n = 14*	
Arts Majors	≥61	n = 253	n = 23		
	51–60	n = 519	n = 337	n = 22	n = 5
	≤50	*n = 478*	*n = 210*	*n = 50*	*n = 36*

Table 2. Descriptive Statistics VST Scores

N	Min	Max	*M*	*SEM*	*SD*	Skew	*SES*	Kurt	*SEK*
3,427	500	7,400	3,715.20	21.66	1,268.15	−0.20	0.04	−0.16	0.08

SD = Standard deviation.

3 Results

Before considering the research questions, the internal consistency reliability of the VST data collected in this study was examined using the Kuder–Richardson Formula 20 test. The coefficient of reliability (KR-20 $\alpha = .91$) was high, indicating that it is highly probable that all the items measured the same construct.

As shown in Table 3, the three *hensachi* groups achieved different mean VST scores. A one-way analysis of variance (ANOVA) was conducted to evaluate the hypothesis that the participants' VST scores would significantly vary between department *hensachi* groups. The independent variable was department *hensachi* group with three levels: ≥ 61, 51–60, and ≤ 50. The dependent variable was VST scores.

The ANOVA showed a significant effect of department *Hensachi* group on VST scores, $F(2, 3424) = 1{,}383.14$, $p < .001$, $\eta^2 = 45$. The strength of the relationship between *hensachi* groups and VST scores was very strong, accounting for 45% of the variation of the dependent variable. Because the overall F test was significant, follow-up tests were conducted to evaluate pairwise differences among the means. As the variances of the three groups were heterogeneous, the Games

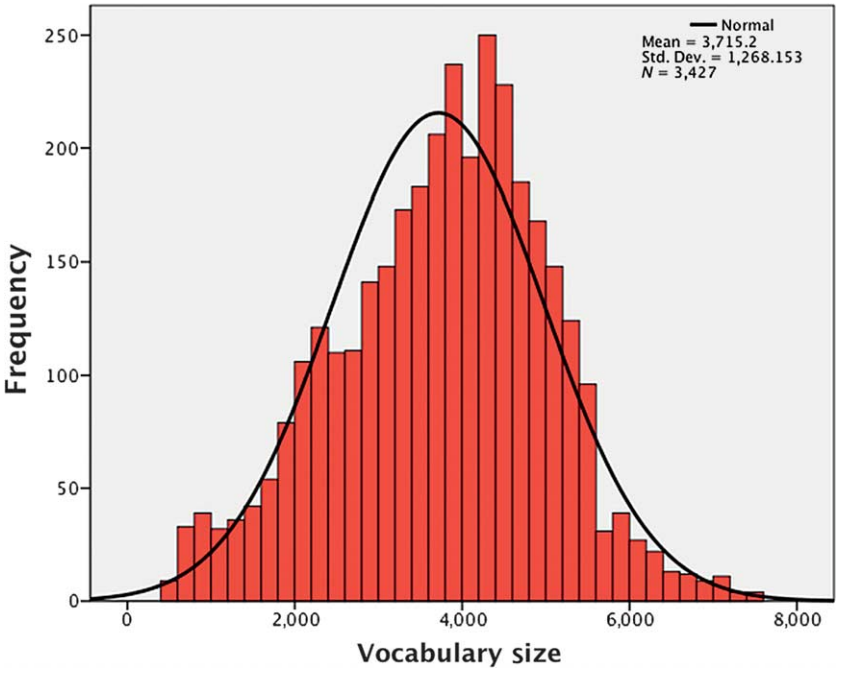

Figure 1. Histogram of grouped distribution of VST scores showing the normal distribution curve.

Table 3. Descriptive Statistics VST Score by Hensachi Group

Hensachi	*N*	Min	Max	*M*	*SEM*	*SD*	Skew	*SES*	Kurt	*SEK*
≥61	603	1,900	7,400	4,903.15	29.44	723.00	−0.19	0.10	0.77	0.20
51–60	1,558	900	7,400	4,102.89	23.18	915.15	0.31	0.06	1.39	0.12
≤50	1,266	500	7,100	2,672.27	29.87	1,062.82	0.29	0.07	0.33	0.14

SEM = Standard Error of the Mean; SES = Standard Error of Skewness; SEK = Standard Error of Kurtosis.

Howell test, which does not assume equal variance among groups, was used. To control for Type I errors, the Bonferroni approach was used, thus a p value of less than .017 (.05/3 = .017) was required for significance. The ≥61 *hensachi* group had significantly higher VST scores than the other two groups (p = <.001 for both comparisons), and that the 51–60 *hensachi* group had significantly higher VST scores than the ≤50 *hensachi* group (p = <.001).

Spearman's rho correlation coefficients were calculated between six different student scores and VST scores. To control for Type I errors, the Bonferroni approach was used, thus a p value of less than .008 (.05/6 = .008) was required for significance. As displayed in Table 4, while all six scores correlated significantly with VST scores, the highest correlation was found between department *hensachi* and VST scores at .73 ($p < .001$), indicating that students studying in departments with higher *hensachi* scores tend to score higher on the VST. Examining the correlational strength of department *hensachi* scores for only students whose TOEFL scores were available ($n = 259$), TOEFL had a higher correlation with VST scores ($r_s = .58$, $p < .001$) than department *hensachi* ($r_s = .34$, $p < .001$). Likewise, the correlation between TOEIC and VST scores ($n = 412$) was higher ($r_s = .57$, $p < .001$) than department *hensachi* ($r_s = .32$, $p < .001$) for the same students.

The mean item accuracy on the VST was calculated for the three *hensachi* groups with items grouped into 1,000-word frequency levels, as shown in Table 5. The accuracy ranged from 85% (*hensachi* ≥61 group, 1k word level) to 22% (*hensachi* ≤50 group, 6k word level). Overall, accuracy decreased as the words became less frequent as shown in Figure 2. The decline was determined to be non-linear by adding best-fit lines with the highest indices of fit. For the *hensachi* ≥61, and 51–60 groups, the lines were polynomial with coefficients of determination of $R^2 = .89$ and $R^2 = .82$, respectively. For the *hensachi* ≤50 group, the line was

Table 4. Spearman's Correlation Coefficients between Student Scores and VST Results

Factor	*n*	*p*	Correlational coefficient with VST scores	Dept *hensachi* and VST coefficient for each subpopulation
Department *hensachi*	3,427	<.001	.73*	.73*
HS English *hensachi*	826	<.001	.60*	.60*
TOEFL	259	<.001	.58*	.34*
TOEIC	412	<.001	.57*	.32*
HS 3rd subject *hensachi*	643	<.001	.50*	.64*
HS 5th subject *hensachi*	628	<.001	.49*	.61*

*$p < .008$. HS = High School.

Table 5. Percentage Item Accuracy at 1,000 Word Levels

	Word Level							
Department *hensachi*	1k	2k	3k	4k	5k	6k	7k	8k
≥61	85	71	73	68	59	42	44	48
51–60	79	54	62	56	43	35	40	42
≤50	59	34	38	36	26	22	26	25

logarithmic with a coefficient of determination of $R^2 = .84$. The trend was toward accuracy declining steeply as item difficulty increased, but then leveling off to a slower decline toward the later items.

4 Discussion

Answering research question 1, the vocabulary sizes of Japanese university students was found to be 3,715.20 word families. This estimate is greater than the 2,000–2,300 as found by Shillaw (1995) and Barrow et al. (1999). However, in these studies students were only tested on their knowledge of the most frequent 3,000 word families in contrast to the present study which tested the participants' knowledge of the most frequent 8,000 words. The mean vocabulary sizes of the ≥61, 51–60 and ≤50 *hensachi* groups are 4,903, 4,103, and 2,792 words, respectively.

Research question 2 asked: Are a university department's standardized rank scores (*hensachi*) a useful proxy for English vocabulary size? The results of the ANOVA and post hoc tests showed that VST scores were significantly higher for students in higher department *hensachi* programs. Figure 2 also shows that the

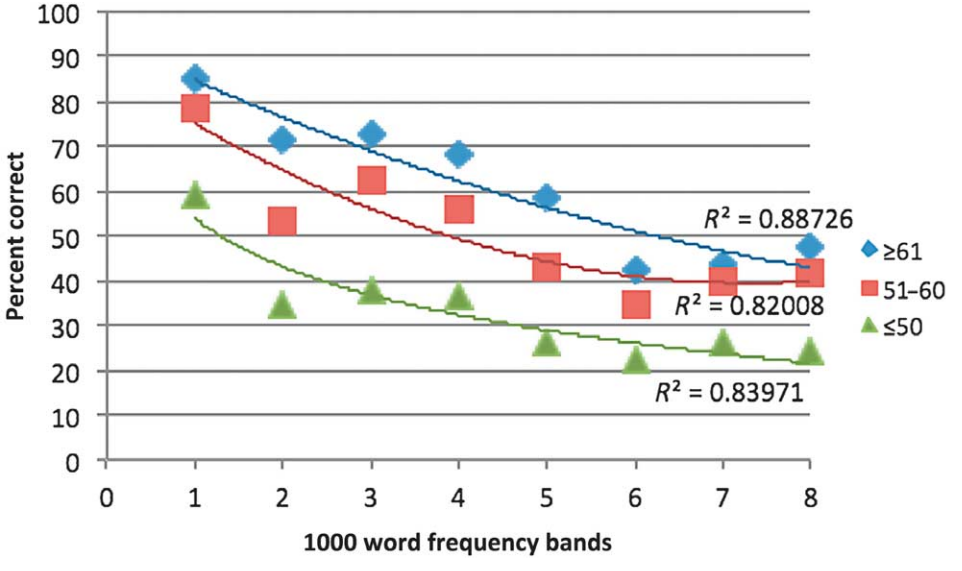

Figure 2. A scatter plot of VST item accuracy at each 1,000-word frequency level for the three *hensachi* groups with best-fit lines.

vocabulary size differences of the three *hensachi* groups are relatively consistent throughout each band of the VST. These findings are promising in that department *hensachi* scores are readily available online allowing teachers to easily gain a general understanding of their students' vocabulary size relative to other departments. However, it should be cautioned that while the mean VST scores achieved represent the approximate total written receptive vocabulary sizes of the students in the three *hensachi* ranges (4,903, 4,103, or 2,792, respectively), the results do not suggest that instructors can expect their students to have complete knowledge of the most frequent 4,903, 4,103, or 2,792 English words. The answers to the research questions, when considered along with the high internal consistency reliability, support the use of the VST for comparing written receptive vocabulary knowledge among English learners, while suggesting that it may not be appropriate for establishing specific vocabulary sizes for uses such as selecting or creating lexically appropriate materials.

In examining the correlational strength of different student scores and vocabulary size, the highest found within all participants was with their current department *hensachi*. This is explained by three factors: possibly differing sources of individual previous *hensachi* scores, the self-reported nature of all the individual scores used and inaccuracies therewith, and not the least of which the larger sample size for the comparisons between department *hensachi* and VST scores. Although the data were limited in this study, when available for an entire group of students, TOEFL scores, followed by TOEIC, appear to be better proxies for vocabulary size. When such scores are unavailable, department *hensachi* is a reasonable substitute.

For all three groups, there appears to be a lack of consistent knowledge of even the most frequent words of English. These lexical gaps were found to be even greater with lower-department *hensachi* students. This suggests that regardless of *hensachi* level or the predicted vocabulary size, vocabulary components of the curricula at Japanese universities should focus on high-frequency English words. This finding also provides negative evidence for the common assumption that texts written at the 1,000 most frequent word level are lexically appropriate for university students in Japan when conducting meaning-focused learning or fluency activities. Even if all of the words within a text were from this level, the average university student would have far less than the 98% lexical coverage recommended by Hu and Nation (2000) for unassisted comprehension. Thus some provision for dealing with unknown words should be made.

5 Conclusion

This paper reports the results of the VST taken by 3,427 students at various Japanese universities. The scores were found to be significantly higher for students in higher-department *hensachi* programs, and a strong correlation was found between the two variables, suggesting that instructors can quickly and easily gain a general understanding of students' vocabulary size simply from their department *hensachi* score. In all three groups, however, the results show a lack of consistent knowledge among the most frequent 2,000 word families found in the BNC. This emphasizes the need to focus on supporting high-frequency vocabulary acquisition regardless of university rank. Alternatively, it questions the appropriateness of

measuring Japanese university students' lexical knowledge based on word frequency within the BNC. This frequency list, for example, ranks the word "nil," at the 2,000 word level, as more frequent than "quiz," at the 5,000 word level, while containing more loanwords than may be proportionate to the BNC as a whole. The results support the use of the VST for separating learners by vocabulary knowledge, but suggest that it may not be appropriate for establishing students' vocabulary size for deciding and creating appropriate materials.

5.1 Limitations and Future Research

Preliminary in-depth interviews conducted with participants after completing the test suggest that, as is to be expected on a multiple choice test, guessing can inflate VST scores, and especially so for test takers with limited lexical knowledge. The issue of student guessing within these data is being addressed for future research, and while the analyses described in this paper indicate that this does not hurt the test's ability to separate the students according to ability, it may cast doubt on the test's ability to accurately estimate the latent vocabulary size. This is in line with the results discussed by Stewart (2014), who also expresses concern that student guessing causes an overestimate of vocabulary size using the VST.

Another limitation is the difficulty of determining the degree to which participants took the test seriously. Although it has yet to be explored in depth, test fatigue appears to have affected the $\leq$50 *hensachi* group to a greater extent based on the number of unanswered items. It is possible that more participants could and should have been removed had it come to light that they had given up or guessed wildly.

In addition to the results presented in this paper, there are a number of directions that would be interesting to pursue and would shed further light onto both Japanese students' vocabulary sizes as well as the VST in general. First, it would be ideal if the participants included more English majors and 3rd and 4th year students. Second, the self-reported *hensachi* data collected for participants' individual English, three-subject, and five-subject *hensachi* scores were of limited value as it was not ascertained which examination boards' tests were taken. The correlations between these scores and the VST may have been higher had this information been accounted for. The issue of loanwords in the test possibly inflating Japanese students' scores should also be explored. While the use of loanwords in the test is not itself a problem, it can become problematic if they make up a greater proportion of the test questions than the corpus it represents.

References

Barrow, J., Nakanishi, Y., & Nishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System*, *27* (2), 223–247. doi:10.1016/S0346-251X(99)00018-4

Educational Testing Service. (2005). *TOEFL iBT scores: Better information about the ability to communicate in an academic setting*. Retrieved from hhl.de/fileadmin/texte/_relaunch/Conversion_Table_TOEFL_(PBT,CBT,iBT).pdf

Field, A. (2009). *Discovering statistics using SPSS*. London, UK: Sage.

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13* (1), 403–430. Retrieved from nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehchao.pdf

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31* (7), 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

Shillaw, J. (1995). Using a word list as a focus for vocabulary learning. *The Language Teacher*, *19* (2), 58–59. Retrieved from jalt-publications.org/tlt/issues/1995-02_19.2

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11* (3), 271–282. doi:10.1080/15434303.2014.922977

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* (2), 130–163. Retrieved from http://nflrc.hawaii.edu/rfl/October2003/waring/waring.html

Vocabulary Learning and Instruction
Volume 3, Issue 2, September 2014
http://vli-journal.org

Article

Local Item Dependence on the Vocabulary Levels Test Revisited

Tadamitsu Kamimoto
Kumamoto Gakuen University
doi: http://dx.doi.org/10.7820/vli.v03.2.kamimoto

Abstract

The purpose of this study was to address the question of local item dependence (LID) on the Vocabulary Levels Test (VLT). The test format adopts a matching format consisting of six words and three definitions. A review of the literature suggested that items presented in such a format cannot be considered independent. However, Schmitt, Schmitt, and Clapham reported that according to Rasch analysis, most of the items performed independently. The present study examined effects of LID from a non-Rasch approach. A set of three clusters was collapsed into one large cluster of 18 words and 9 definitions on the assumption that such a treatment would make items practically independent. Both the collapsed VLT and the intact VLT were given to 114 Japanese English as a Foreign Language (EFL) students at an interval of one week. Results showed that scores were 15% higher on the original VLT than on the collapsed form. Furthermore, an LID index based on the correct/wrong response types between the two tests indicated that scores on the original VLT were about 19% inflated. Implications were drawn and discussed.

1 Background

Nation's (1990) Vocabulary Levels Test (VLT) and its revised forms (Nation, 2001; Schmitt, 2000; Schmitt, Schmitt, & Clapham, 2001) are still in use (e.g., Akbarian, 2010; Li & MacGregor, 2010; Rashidi & Piran, 2011; Staehr, 2008; Webb, 2007). Validations of the VLT have been carried out by Read (1988), Beglar and Hunt (1999) and Schmitt et al. (2001). They all indicated that the test is a good measurement of vocabulary knowledge. However, there is one issue which has not been adequately addressed: local item dependence (LID). It is assumed that a response to an item should be independent of a response to another item (Hambleton, Swaminathan, & Roger, 1991; Lee, 2004; Zenisky, Hambleton, & Sireci, 2002). However, the VLT seems to violate this assumption (e.g., Nation, 2007).

Where does the possible violation of item independence come from? It comes from the test format – the matching format, as shown in Figure 1. This format, interacting with test takers' knowledge of words, is responsible for triggering LID to various degrees. The most serious effect of LID is to increase the chance of guessing correctly (Mochizuki, 1998; Shillaw, 1999; Stewart, & White, 2011).

Please cite this article as: Kamimoto, T. (2014). Local item dependence on the vocabulary levels test revisited. *Vocabulary Learning and Instruction*, *3* (2), 56–68. doi: 10.7820/vli.v03.2.kamimoto

1	business	
2	clock	_____ part of a house
3	horse	_____ animal with four legs
4	pencil	_____ something used for writing
5	shoe	
6	wall	

Figure 1. Example of the VLT Format.

In the validation studies, Beglar and Hunt (1999) and Schmitt et al. (2001) found some dependent items on the basis of Rasch analysis. However, they interpreted that most items were not dependent and caused no particular adverse effects to their following analyses. Then, they shifted items to make parallel forms [see Xing & Fulcher (2007) for criticism]. This paper shall revisit the LID issue and examine it from a non-Rasch approach.

2 Method

2.1 Material

This study used the original VLT (Nation, 1990). Among the five word levels available, four levels were used. One level consisted of six clusters, each of which had six words and three definitions, as shown in Figure 1. The original test was referred to as "6 3 form" (see Appendix 1). On the basis of this material, an experimental matching format was constructed by collapsing a set of three clusters into one large cluster consisting of 18 words and 9 definitions. This collapsed form was referred to as "18 9 form" (see Appendix 2).

2.2 Participants and Procedure

Three groups of EFL Japanese university students took the 18 9 form first and one week later the 6 3 form. Due to absence on either of the two occasions, several students were dropped from the analysis, leaving a total of 114 participants.

3 Results and Discussion

3.1 Descriptive Statistics

Table 1 shows participants' overall performance on the two forms. There are two points to be noted. First, the mean scores for the 6 3 form were larger across all the four word-frequency levels than the mean scores for the 18 9 form. Second, as the word-frequency levels became lower, the mean scores for the two forms became smaller but differences in scores gradually increased. Paired *t*-tests show that all the differences were statistically significant at the 0.01 level ($df = 113$, t values $=$ 9.16 for 2K, $r = .65$; 14.90 for 3K, $r = .82$; 16.63 for 5K, $r = .84$; and 16.75 for AWL (Academic Word List), $r = .84$, two-tailed).

Table 1. Means and Standard Deviations of Scores for the Two Forms

	Mean (%)			SD (%)	
Word level	6 3 form	18 9 form	Differences (%)	6 3 form	18 9 form
2K	15.96 (88.67)	14.52 (80.67)	1.44 (8.00)	1.66 (9.22)	2.29 (12.72)
3K	13.63 (75.72)	10.89 (60.50)	2.75 (15.22)	2.46 (13.67)	2.92 (16.22)
5K	10.54 (58.56)	7.27 (40.39)	3.26 (18.13)	2.42 (13.72)	2.07 (11.50)
AWL	9.34 (51.89)	5.72 (31.78)	3.62 (20.11)	2.91 (16.17)	2.97 (16.50)
Sum[a]	49.47 (68.71)	38.40 (53.33)	11.07 (15.38)	7.39 (10.26)	8.49 (11.76)

The maximum score per word level is 18.
[a]Figures in this row do not always add up to a sum of four figures in each column due to rounding.

3.2 Item Facility

Item facility (IF) indices were employed to identify LID items. We subtracted IF values of items for the 18 9 form from corresponding IF values of the same items for the 6 3 form. Results showed that the IF differences ranged from −0.009 to +0.421, with a mean of 0.154 (see Appendix 3 for details). All IF differences fell above zero except one item,[1] and three items had the same score. With regard to the remaining 68 items, participants performed better on the 6 3 form than on the 18 9 form.

Figure 2 is a graphical representation of the entire IF data. Each vertical line has two IF values at its ends; a triangle indicates an IF value for the 6 3 form and a cross indicates an IF value for the 18 9 form. Its length represents how widely the IF values of the same item are different. Although several items with no or very short length were found here and there, most of the items had various degrees of length. In short, the number of items with some length outnumbers by far the number of items with no or very short length, suggesting strongly that the 6 3 form inflated the scores of most items.

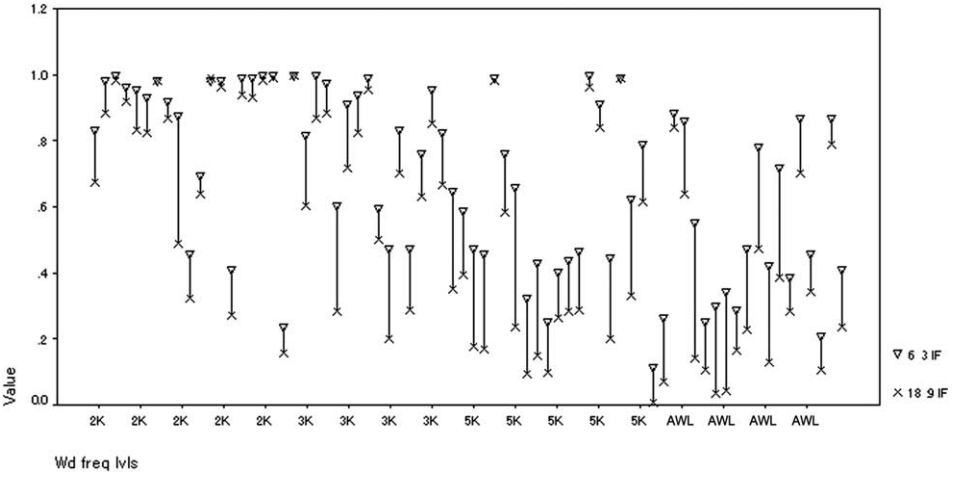

Figure 2. Drop-line Plot of IF Values between the Two Forms.

Table 2. A Frequency Table of Items by Word Levels and IF Differences

	2K	3K	5K	AWL	Total (%)
IF diff ranges					
$0.50 \leqq \sim < 0.60$					
$0.40 \leqq \sim < 0.50$			1	1	2 (2.78)
$0.30 \leqq \sim < 0.40$	1	1		2	4 (5.56)
$0.20 \leqq \sim < 0.30$		3	6	5	14 (19.44)
$0.10 \leqq \sim < 0.20$	5	9	7	8	29 (38.89)
$0.00 \leqq \sim < 0.10$	11	5	4	2	22 (26.39)
$-0.10 \leqq \sim < 0.00$	1				1 (4.17)
Total	18	18	18	18	72 (100)

In the hope of sketching out the general features of the data on item difficulty, the frequency counts were tabulated according to both the IF difference range and the word levels in Table 2. For identification of LID items, two cut-off points can be adopted, depending on the level of severity. If an IF difference of 0.10 is adopted on an assumption that measurement error can happen within an IF difference range of less than 0.10, LID items account for 49 (68.06%) of 72 items. They include 6 2K items, 13 3K items, 14 5K items and 16 AWL items. If a lenient criterion of 0.20 is adopted, LID items account for 20 (27.78%) of all the items: one 2K item, four 3K items, seven 5K items and eight AWL items. It should also be noted that the number of the LID items increased as word difficulty increased in the order of word frequency. In other words, it appears that LID tends to become more marked as words become less frequent.

3.3 LID Items

IF indices are a handy but rough measure for the identification of LID items because IF values on the 6 3 form represent the end product of a series of test-taking behaviours within a cluster. For more accurate identification, it is preferable to see whether the participants' responses are consistent or not between the 6 3 and 18 9 forms. Figure 3 illustrates four possible response types on items between the two forms. When participants matched an item correctly to a definition

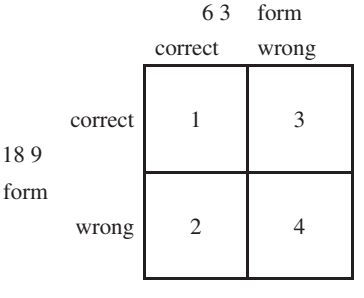

Figure 3. A Taxonomy of Response Types between the Two Forms.

on both the 6 3 and 18 9 forms, their response can be considered as LID-free. Participants would answer right irrespective of other words within a cluster. This response type corresponds to the first cell. On the other hand, if a response was correct on the 6 3 form but wrong on the 18 9 form, it suggests that the participants did not actually know an item and they unduly gained a score on the 6 3 format. This response type is where guessing effects by LID are found. The second cell is the case in point. The opposite case where a response is correct on the 18 9 form but wrong on the 6 3 form is expected to be rare. This corresponds to the third cell. Finally, a wrong response on both forms suggests that the participants did not know the item. This is a response type of the fourth cell. Our chief concern is with response Type 2.

Table 3 shows frequencies of the following three types of data: (1) the number correct on the two forms, (2) an overlap between them corresponding to the first cell in Figure 3 and (3) deviations on the 6 3 and 18 9 forms from an overlap corresponding, respectively, to the second and third cells. Items are displayed in the order of appearance on the 6 3 form. A glance at the table indicates that the frequency for response Type 2 ranged from 0 to 53 (46.49%), indicating that guessing effects of LID varied considerably from one item to another.

This frequency table is also very useful in comprehending how items in a cluster affect one another. As an illustrative example, let us take up the case of "spoil" from the 2K word level, which had the second highest LID index of 46 (40.35%) among all the items. As the table shows, 56 participants were correct on the 18 9 form, and one week later the number correct for the item on the 6 3 form increased to 100. How did such a large increase take place? The other two target items with which "spoil" shared the same cluster were "invite" and "hide". They had an overlap of 110 and 99, respectively, and the LID indices for them were 2 and 6. In short, these two items were extremely easy when compared with "spoil". After matching them correctly to definitions, the participants apparently tackled with "spoil". In the end, it appears that 46 participants ended up gaining a score, while the remaining number of the participants chose one of the three distractors of "blame," "hit" and "pour" which were relatively easy.

Table 4, summing up the data above, displays descriptive statistics of the guessing rates according to word frequency level. Means ranged from 9.89% for the 2K word level to 24.61% for the AWL level with a mean of 18.84%. The guessing rates increased as word frequency decreased, suggesting that the less frequent the word levels may be, the more dependent items tend to be. It is interesting to note that the guessing rates are higher than the corresponding IF differences shown in Table 1. This evidence, adequately supporting Stewart & White (2011), showed that successful guessing may be more prevalent than their theoretical estimation of guessing even when items are less frequent.

4 Conclusion

These findings have shown that the original matching format of the VLT has a mean guessing rate of 18.84%. This rate may vary according to test takers' lexical proficiency and the number of word levels involved. It seems reasonable to

Table 3. A Frequency Table of Response Types

		Number correct		Response		
				Type 1	Type 2	Type 3
Word ID	Item	6 3	18 9	Overlap	6 3 (%)	18 9 (%)
2K01	original	95	77	76	19 (16.67)	1 (0.88)
2K02	private	112	101	100	12 (10.53)	1 (0.88)
2K03	total	114	112	112	2 (1.75)	0 (0.00)
2K04	manufacture	110	105	104	6 (5.26)	1 (0.88)
2K05	elect	109	95	91	18 (15.79)	4 (3.51)
2K06	melt	106	94	93	13 (11.40)	1 (0.88)
2K07	invite	112	112	110	2 (1.75)	2 (1.75)
2K08	hide	105	99	99	6 (5.26)	0 (0.00)
2K09	spoil	100	56	54	46 (40.35)	2 (1.75)
2K10	roar	52	37	27	25 (21.93)	10 (8.77)
2K11	debt	79	73	63	16 (14.04)	10 (8.77)
2K12	pride	112	113	111	1 (0.88)	2 (1.75)
2K13	temperature	112	110	109	3 (2.63)	1 (0.88)
2K14	flesh	47	31	30	17 (14.91)	1 (0.88)
2K15	salary	113	107	107	6 (5.26)	0 (0.00)
2K16	sport	113	106	105	8 (7.02)	1 (0.88)
2K17	victory	114	112	112	2 (1.75)	0 (0.00)
2K18	birth	114	113	113	1 (0.88)	0 (0.00)
3K01	herd	27	18	13	14 (12.28)	5 (4.39)
3K02	angel	114	114	114	0 (0.00)	0 (0.00)
3K03	administration	93	69	59	34 (29.83)	10 (8.77)
3K04	bench	114	99	99	15 (13.16)	0 (0.00)
3K05	charity	111	101	101	10 (8.77)	0 (0.00)
3K06	province	69	32	29	40 (35.09)	3 (2.63)
3K07	echo	104	82	78	26 (22.81)	4 (3.51)
3K08	darling	107	94	93	14 (12.28)	1 (0.88)
3K09	slice	113	109	109	4 (3.51)	0 (0.00)
3K10	palm	68	57	46	22 (19.30)	11 (9.65)
3K11	scheme	54	23	22	32 (28.07)	1 (0.88)
3K12	thrill	95	80	78	17 (14.91)	2 (1.75)
3K13	illustrate	54	33	23	31 (27.19)	10 (8.77)
3K14	encounter	87	72	66	21 (18.42)	6 (5.26)
3K15	toss	109	97	96	13 (11.40)	1 (0.88)
3K16	annual	94	76	75	19 (16.67)	1 (0.88)
3K17	savage	74	40	33	41 (35.97)	7 (6.14)
3K18	definite	67	45	37	30 (26.32)	8 (7.02)
5K01	phase	54	20	14	40 (35.09)	6 (5.26)
5K02	mess	52	19	16	36 (31.58)	3 (2.63)
5K03	apron	113	112	111	2 (1.75)	1 (0.88)
5K04	trumpet	87	66	61	26 (22.81)	5 (4.39)
5K05	stool	75	27	22	53 (46.49)	5 (4.39)
5K06	sermon	37	11	9	28 (24.56)	2 (1.75)
5K07	compliment	49	17	13	36 (31.58)	4 (3.51)

Table 3 (*Continued*)

		Number correct		Response		
				Type 1	Type 2	Type 3
Word ID	Item	6 3	18 9	Overlap	6 3 (%)	18 9 (%)
5K08	apparatus	29	11	6	23 (20.18)	5 (4.39)
5K09	revenue	46	30	15	31 (27.19)	15 (13.16)
5K10	precede	50	32	21	29 (25.44)	11 (9.65)
5K11	collapse	53	33	26	27 (23.68)	7 (6.14)
5K12	skip	114	110	110	4 (3.51)	0 (0.00)
5K13	blend	104	96	91	13 (11.40)	5 (4.39)
5K14	devise	51	23	16	35 (30.70)	7 (6.14)
5K15	hug	113	113	113	0 (0.00)	0 (0.00)
5K16	gloomy	71	38	34	37 (32.46)	4 (3.51)
5K17	fragrant	90	70	68	22 (19.30)	2 (1.75)
5K18	wholesome	13	1	0	13 (11.40)	1 (0.88)
AWL01	affluence	30	8	4	26 (22.81)	4 (3.51)
AWL02	episode	101	96	89	12 (10.53)	7 (6.14)
AWL03	innovation	96	73	68	28 (24.56)	5 (4.39)
AWL04	deficiency	63	16	15	48 (42.11)	1 (0.88)
AWL05	prestige	29	12	5	24 (21.05)	7 (6.14)
AWL06	oscillation	34	4	2	32 (28.07)	2 (1.75)
AWL07	configuration	39	5	2	37 (32.46)	3 (2.63)
AWL08	discourse	33	19	9	24 (21.05)	10 (8.77)
AWL09	hypothesis	54	26	20	34 (29.83)	6 (5.26)
AWL10	supplement	89	54	52	37 (32.46)	2 (1.75)
AWL11	expel	48	15	11	37 (32.46)	4 (3.51)
AWL12	deprive	82	44	40	42 (36.84)	4 (3.51)
AWL13	restrict	44	32	22	22 (19.30)	10 (8.77)
AWL14	transform	99	80	76	23 (20.18)	4 (3.51)
AWL15	assume	52	39	36	16 (14.04)	3 (2.63)
AWL16	indigenous	24	12	2	22 (19.30)	10 (8.77)
AWL17	minimum	99	90	86	13 (11.40)	4 (3.51)
AWL18	anonymous	47	27	19	28 (24.56)	8 (7.02)

conclude, however, that this study indicates that items in clusters tend to be more dependent than the previous Rasch-based studies suggested. This may lead us to wonder if they have met the item response theory (IRT) requirement of item independence between responses to any pair of items.

Table 4. Descriptive Statistics of Guessing Rates by Word-frequency Level

	Mean (%)	SD (%)
2K	11.28 (9.89)	11.35 (9.96)
3K	21.28 (18.67)	11.70 (10.26)
5K	25.28 (22.17)	14.26 (12.51)
AWL	28.06 (24.61)	9.85 (8.64)
Four levels	21.47 (18.84)	13.29 (11.65)

Note

1. The item is "pride" from the 2K word frequency level. This was the only item with a minus IF difference from among a total of 72 items. Its value (−0.009) was extremely small and equivalent to one participant. This irregularity may be caused by a participant's carelessness. It can be dismissed as negligible.

References

Akbarian, I. (2010). The relationship between vocabulary size and depth for ESP/EAP learners. *System*, *38*, 391–401. doi:10.1016/j.system.2010.06.013

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level. *Language Testing*, *16*, 131–162. doi:10.1177/026553229901600202

Hambleton, R. K., Swaminathan, H., & Roger, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). London, UK: Sage.

Lee, Y.-W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q_3 statistics in an EFL reading comprehension test. *Language Testing*, *21*, 74–100. doi:10.1191/0265532204lt260oa

Li, L., & MacGregor, L. J. (2010). Investigating the receptive vocabulary size of university-level Chinese learners of English: How suitable is the Vocabulary Levels Test? *Language and Education*, *24*, 239–249. doi:10.1080/09500781003642478

Mochizuki, M. (1998). Nihonjin Eigo Gakushusha no tameno goi saizu tesuto [A vocabulary size test for Japanese learners of English]. *The IRLT (Institute for Research in Language Teaching) Bulletin*, *12*, 27–53.

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

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

Nation, I.S.P. (2007). Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 35–43). Cambridge, UK: Cambridge University Press.

Rashidi, N., & Piran, M. (2011). The effect of extensive and intensive reading on Iranian EFL learners' vocabulary size and depth. *Journal of Language Teaching and Research*, *2*, 471–482. doi:10.4304/jltr.2.2.471-482

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal*, *19*(2), 12–25. doi:10.1177/003368828801900202

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge, UK: Cambridge University Press.

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

Shillaw, J. (1999). *The application of the Rasch model to yes/no Vocabulary Tests* (Unpublished Ph.D. thesis). University of Wales Swansea, Swansea, Wales.

Staehr, 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., & White, D. A. (2011). Estimating guessing effects on the Vocabulary Levels Test for differing degrees of word knowledge. *TESOL Quarterly*, *45*, 370–380.

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

Xing, P., & Fulcher, G. (2007). Reliability assessment for two versions of Vocabulary Levels Tests. *System*, *35*, 182–191. doi:10.1016/j.system.2006.12.009

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, *39*, 291–309. doi:10.1111/j.1745-3984.2002.tb01144.x

Appendix 1

6 3 Form (Sample)

3K word frequency level

1 administration
2 angel _____ 群れ
3 frost _____ 天使
4 herd _____ 運営
5 fort
6 pond

1 bench
2 charity _____ 長椅子
3 mate _____ 慈善(じぜん)
4 jar _____ 地方
5 mirror
6 province

1 coach
2 darling _____ こだま
3 echo _____ 最愛の人
4 interior _____ 薄切り 1 枚
5 opera
6 slice

1 marble
2 palm _____ 手のひら
3 ridge _____ 計画
4 scheme _____ わくわくする感じ
5 statue
6 thrill

1 discharge
2 encounter _____ 例証する
3 illustrate _____ 出合う
4 prevail _____ を投げる
5 restore
6 toss

1 annual
2 previous _____ 年 1 回の
3 mental _____ 野生の
4 concealed _____ はっきりと確定した
5 definite
6 savage

Appendix 2

18 9 Form (Sample)

3K word frequency level

(A)

1 administration
2 angel _____ 薄切り 1 枚
3 bench _____ 運営
4 charity _____ こだま
5 coach _____ 最愛の人
6 darling _____ 慈善
7 echo _____ 地方
8 fort _____ 天使
9 frost _____ 長椅子
10 herd _____ 群れ
11 interior
12 jar
13 mate
14 mirror
15 opera
16 pond
17 province
18 slice

(B)

1 annual
2 concealed _____ 計画
3 definite _____ 出合う
4 discharge _____ 手のひら
5 encounter _____ 年 1 回の
6 illustrate _____ はっきりと確定した
7 marble _____ 野生の
8 mental _____ 例証する
9 palm _____ わくわくする感じ
10 prevail _____ を投げる
11 previous
12 restore
13 ridge
14 savage
15 scheme
16 statue
17 thrill
18 toss

Appendix 3

All the 72 Items in the Increasing Order of IF Differences

No	Item ID*	Items	6 3 IF	18 9 IF	Differences
1	2K12	pride	0.982	0.991	−0.009
2	2K07	invite	0.982	0.982	0.000
3	3K02	angel	1.000	1.000	0.000
4	5K15	hug	0.991	0.991	0.000
5	2K18	birth	1.000	0.991	0.009
6	5K03	apron	0.991	0.982	0.009
7	2K13	temperature	0.982	0.965	0.018
8	2K03	total	1.000	0.982	0.018
9	2K17	victory	1.000	0.982	0.018
10	3K09	slice	0.991	0.956	0.035
11	5K12	skip	1.000	0.965	0.035
12	2K04	manufacture	0.965	0.921	0.044
13	AWL02	episode	0.886	0.842	0.044
14	2K08	hide	0.921	0.868	0.053
15	2K15	salary	0.991	0.939	0.053
16	2K11	debt	0.693	0.640	0.053
17	2K16	sport	0.991	0.930	0.061
18	5K13	blend	0.912	0.842	0.070
19	3K02	herd	0.237	0.158	0.079
20	AWL17	minimum	0.868	0.789	0.079
21	3K05	charity	0.974	0.886	0.088
22	2K02	private	0.982	0.886	0.096
23	3K10	palm	0.596	0.500	0.096
24	AWL13	restrict	0.386	0.283	0.103
25	2K06	melt	0.930	0.825	0.105
26	3K15	toss	0.956	0.851	0.105
27	5K18	wholesome	0.114	0.009	0.105
28	AWL16	indigenous	0.211	0.105	0.105
29	AWL15	assume	0.456	0.342	0.114
30	3K08	darling	0.939	0.825	0.114
31	2K05	elect	0.956	0.833	0.123
32	AWL08	discourse	0.289	0.167	0.123
33	2K10	roar	0.456	0.325	0.132
34	3K04	bench	1.000	0.868	0.132
35	3K12	thrill	0.833	0.702	0.132
36	3K14	encounter	0.763	0.632	0.132
37	5K09	revenue	0.404	0.265	0.138
38	2K14	flesh	0.412	0.272	0.140
39	AWL05	prestige	0.254	0.105	0.149
40	5K10	precede	0.439	0.283	0.155
41	5K08	apparatus	0.254	0.097	0.157
42	2K01	original	0.833	0.675	0.158
43	3K16	annual	0.825	0.667	0.158

Appendix 3 (*Continued*)

No	Item ID*	Items	6 3 IF	18 9 IF	Differences
44	AWL14	transform	0.868	0.702	0.167
45	5K11	collapse	0.465	0.289	0.175
46	5K17	fragrant	0.789	0.614	0.175
47	AWL18	anonymous	0.412	0.237	0.175
48	5K04	trumpet	0.763	0.584	0.179
49	3K13	illustrate	0.474	0.289	0.184
50	3K07	echo	0.912	0.719	0.193
51	AWL01	affluence	0.263	0.070	0.193
52	3K18	definite	0.588	0.395	0.193
53	3K03	administration	0.816	0.605	0.211
54	AWL03	innovation	0.860	0.640	0.219
55	5K06	sermon	0.325	0.096	0.228
56	AWL09	hypothesis	0.474	0.228	0.246
57	5K14	devise	0.447	0.202	0.246
58	AWL06	oscillation	0.298	0.035	0.263
59	3K11	scheme	0.474	0.202	0.272
60	5K07	compliment	0.430	0.149	0.281
61	5K02	mess	0.456	0.168	0.288
62	AWL11	expel	0.421	0.132	0.289
63	5K16	gloomy	0.623	0.333	0.289
64	5K01	phase	0.474	0.177	0.297
65	3K17	savage	0.649	0.351	0.298
66	AWL07	configuration	0.342	0.044	0.298
67	AWL10	supplement	0.781	0.474	0.307
68	3K06	province	0.605	0.283	0.322
69	AWL12	deprive	0.719	0.386	0.333
70	2K09	spoil	0.877	0.491	0.386
71	AWL04	deficiency	0.553	0.140	0.412
72	5K05	stool	0.658	0.237	0.421

Note: Item ID* consists of a combination of a word frequency level and a two-digit number (from 01 to 18) which corresponds to the appearing order of a definition within a word-frequency level in the original VLT.

Vocabulary Learning and Instruction
Volume 3, Issue 2, September 2014
http://vli-journal.org

Article

Test Taking and DK Use on the Vocabulary Size Test

Dawn Lucovich
Teachers College Columbia University, The Writing Center
doi: http://dx.doi.org/10.7820/vli.v03.2.lucovich

Abstract

This study continued an investigation of test strategy and usage of an "I don't know" (DK) option on the 20,000-word family, 100-item Vocabulary Size Test (VST; Nation & Beglar, 2007). Unlike previous studies, which reported on an L1 American English user (Lucovich, 2014), the participants in this study were two L1 Greek–L2 American English users. Each participant took the original and DK-added versions of the VST. A structured qualitative interview, based on Nagy, Herman, and Anderson's (1985) word interview protocol and piloted in previous studies, asked the participants about their answers on the VST. This study aimed to discover how non-Japanese L1/non-English L1 test takers determined their answers, how they qualitatively perceived and used the DK option, and whether this differed from the L1 American English and L1 Japanese users of English. As in previous studies, the participants used the DK option only on unknown items where partial knowledge or test taking strategies were unusable.

1 Introduction

Although vocabulary learning and instruction is considered integral to academic curricula and to reading comprehension, knowledge of learners' vocabulary size tends to be overlooked. Having an estimate of vocabulary size is a crucial part of determining reading level, which in turn affects curriculum development, textbook selection, learning objectives, and classroom goals. Pedagogically, vocabulary size can be used to gauge prior knowledge, learner progress, effectiveness of instruction, and to guide vocabulary instruction and assessment.

The Vocabulary Size Test (VST; Nation & Beglar, 2007) has been established as a valid and reliable measure of written receptive vocabulary size (Beglar, 2010). In addition, it is easy to administer, take, score, and interpret, and is useful to both test takers and administrators. However, guessing on multiple-choice VSTs and potential vocabulary size inflation is a concern (Stewart, 2014; Stewart & White, 2011). Although the VST typically lacks an "opt-out" response, one method that has been shown to counteract guessing on multiple-choice vocabulary tests is the addition of an "I don't know" (DK) option (Bennett & Stoeckel, 2012; Zhang, 2013). However, as the collection of both quantitative and qualitative data on the process of vocabulary test taking and on the VST has been encouraged (Nation, 2012a, 2013, personal communication), more intensive qualitative research is

Please cite this article as: 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

needed to complement the existing quantitative data on vocabulary testing. As in this study, Nation (2012a) has also employed a "labour-intensive method of testing" that used one-on-one interviews in order to "make sure that the results meant something" (p. 10).

This study aimed to discover: (1) how L1 Greek–L2 American English (i.e. non-Japanese L1/non-English L1) test takers determined their answers on the VST; (2) how they qualitatively perceived and used the DK option; (3) whether this differed from an L1 American English and L1 Japanese users of English.

1.1 The Issue of Guessing in Vocabulary Testing

As with most multiple-choice vocabulary tests, the VST lacks an "opt-out" response. Since the VST is a written multiple-choice test of receptive vocabulary size, partial knowledge that leads to a correct answer is encouraged; hence, the test does not include an opt-out response (Nation, 2012b). Instead of answering items, however, test takers might leave items unanswered (no response or NR) or guess randomly. For the test administrator, items coded as NR do not provide information about an item or the test taker's vocabulary knowledge. NR cannot differentiate between human error and (lack of) knowledge: The item could have been accidentally missed (human error); intentionally skipped (possibly due to lack of vocabulary knowledge) and then forgotten (human error); or intentionally skipped (lack of vocabulary knowledge).

From a test taker's point of view, there is little advantage to leaving an item unanswered as it will automatically be marked wrong. Without an opt-out response, penalty directions, or with unknown consequences, there is a much greater advantage, score-wise, to guessing, as test takers have a chance of getting the answer correct. Without an opt-out response, a test administrator cannot differentiate between a correct answer (vocabulary knowledge), an informed guess (partial vocabulary knowledge or test taking strategy), and an uninformed guess (lack of vocabulary knowledge or test taking strategy).

1.2 The DK Option in Vocabulary Testing

One method that could potentially deter guessing is the inclusion of a DK option among the choices. The DK option on its own has been found to slightly discourage guessing and to significantly discourage guessing when coupled with a scoring penalty (Zhang, 2013). However, whether a DK option discourages informed guessing or only uninformed guessing, and how and when, is something that can be better explored through qualitative means.

The addition of a DK option could have a variety of effects. Test takers could potentially be using a DK option when: they have never studied or heard a word before, they cannot guess based on partial knowledge, they cannot apply a test strategy like process of elimination, they cannot definitively decide between two choices, they feel test fatigue, they do not know or are not sure of one or more answers sequentially, they do not care about the outcome of a test, or when the test results do not matter or the effort of test taking outweighs the impact of the results.

How test takers perceive and use the DK option is important to answer as the inclusion of a DK option could lead to score overestimation or underestimation. Similarly, these perceptions and usage may differ between different test takers or types of test takers for reasons such as education level, L1, or English proficiency, which are explored in this study.

2 Methodology

2.1 Instruments

The VST (Nation & Beglar, 2007) can test up to the 20,000-word family level (Nation, 2012b) for users of English, whether they are L1, EFL, or ESL users. The 100-item monolingual (i.e. English) VST was used as an instrument for this study. Since two versions of the VST have been tested and deemed equivalent versions (Beglar, 2010), one version was utilized in its original format, while a fifth choice (*I don't know*) was added to the second, equivalent VST. A sample item from each of the VSTs follows:

Original VST:

1. drive: He <drives> fast.

a swims
b learns
c throws balls
d uses a car

DK-added VST:

1. see: They <saw it>.

a closed it tightly
b waited for it
c looked at it
d started it up
e I don't know

Directions are not provided on the VST, and should be added by the test administrator according to the level of test strictness required. Unlike Zhang (2013), which included three testing conditions (including one with a guessing penalty), this study included test directions with high ecological and face validity; that is, they were most likely to resemble program or classroom-level administrations of the VST. The directions used in this study follow:

Original VST:

> For each item, read the word and the example sentence. Circle the letter of the answer (a, b, c, or d) that most closely matches the word.

DK-added VST:

> For each item, read the word and the example sentence. Circle the letter of the answer (a, b, c, or d) that most closely matches the word. If you do not know the answer, circle e.

2.2 Procedure

The participants were two Greek American males in their mid-thirties. Participant A holds a B.A. Economics and Participant B holds a B.S. Psychology, respectively. Both participants spoke Greek at home until approximately kindergarten; both attended elementary school in Greece for 2–3 years before returning to the USA. Participant A returned to Greece in his early thirties for several years. However, Participant A volunteered that "my English is better than my Greek now" and Participant B agreed that "now I definitely speak English a lot more fluently". Compared to previous research, which reported the results of a high-proficiency L1 English user and intermediate-proficiency L1 Japanese users of English, the L1 Greek users can be said to occupy an intervening position in English-language proficiency: they are L2 users of English, but proficient enough they that now consider their L2 to be their dominant language.

The participants were met on two separate occasions and individually briefed about the study. Both the original and DK-added VSTs were taken by each participant. Each test taking period was between 20 and 60 minutes long. Each interview was conducted one-on-one and lasted an average of 20 minutes. A short debriefing was then conducted. Each session lasted no longer than 2 hours in total.

The qualitative interview utilized a structured item interview based on the word interview protocol by Nagy, Herman, and Anderson (1985) with elements of a retrospective think-aloud protocol. Each interview was conducted immediately after the test was administered. The interviews were transcribed and coded by the interviewer using provisional coding (Miles & Huberman, 1994) developed in previous research (Lucovich, 2014). The questions included:

1. Can you tell me what the word means?
2. Why did you choose that answer?
3. How did you arrive at that answer?

3 Results

Below are the results of the dichotomously scored (correct–incorrect) answers on the two versions of the VST (Figures 1 and 2), including the frequency range of the incorrect and DK answers.

> *Research Question 1*. How do L1 Greek–L2 American English test takers determine their answers on the VST?

First, the results of the participants' structured interviews fell into several broad categories. When asked how they knew a word and if they could demonstrate

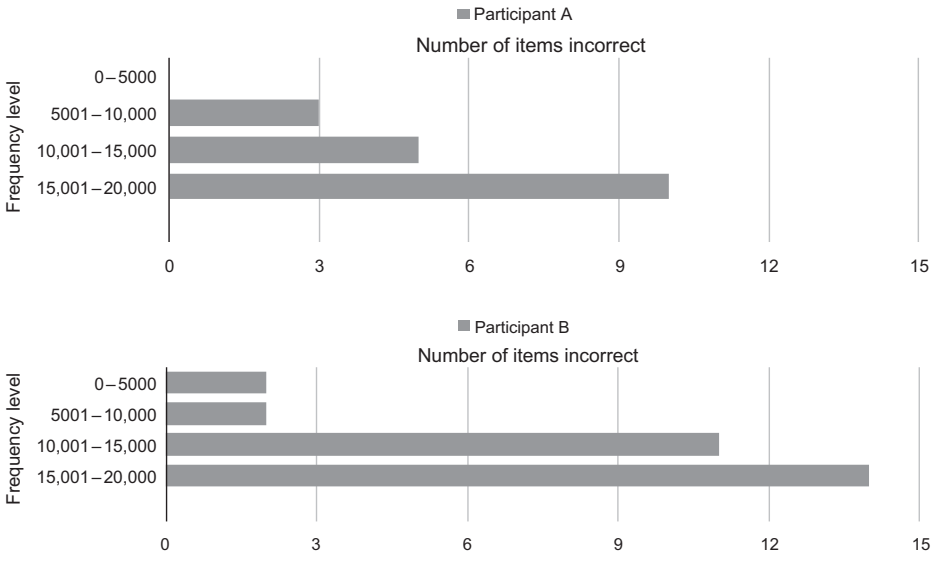

Figure 1. Number of incorrect items on the original VST according to frequency level by participant.

knowledge, they consistently identified or were classified into the following conditions:

1. *Knowledge-informed answer*: They were familiar with the target word.
2. *Partial knowledge-informed answer*: They had some knowledge of the target word, but this knowledge was not sufficient enough to satisfy condition #1.

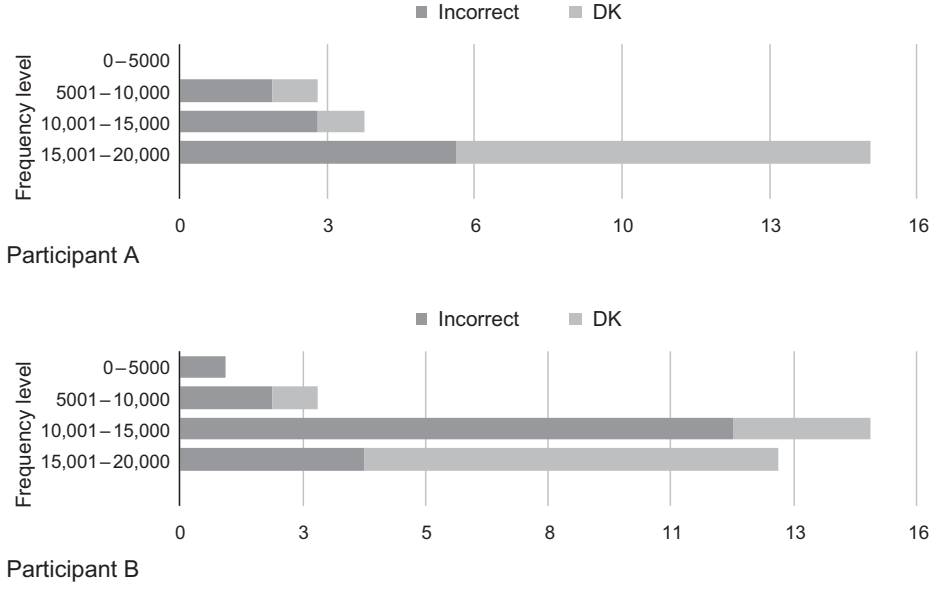

Figure 2. Number of incorrect items vs. DK usage on the DK-added VST according to frequency level by participant.

3. *Test strategy-informed answer*: They had some knowledge of a non-target word, a distractor, non-distractor, or item stem.
4. *Informed guess*: They had some knowledge or strategy, but it was insufficient to satisfy conditions #1, 2, or 3.
5. *Uninformed guess*: They lacked any knowledge or were unable to provide any additional information on the item or target word.

The participants' coded answers confirmed the provisional coding developed in previous research (Lucovich, 2014). In pilot studies, an L1 American English user and L1 Japanese users of English identified or were classified into identical conditions.

The participants cited the following reasons for their answers on the 100-item VST (in decreasing order of use, not differentiating between correct/incorrect answers):

Original VST:

Knowledge-informed answer (70%)
Partial knowledge-informed answer (16%)
Test strategy-informed answer (6%)
Uninformed guess (5%)
Informed guess (4%)

DK-added VST:

Knowledge-informed answer (61%)
Partial knowledge-informed answer (17%)
DK usage (12%)
Test strategy-informed answer (7%)
Uninformed guess (2%)
Informed guess (2%)

However, unlike previous studies, the participants also mentioned their use of cognates (or perceived cognates) as partial knowledge or as a test strategy. For #75 "cenotaph", Participant B said: "I don't know if there's any relation, but there's a Greek word which means 'coffin'. So I think that's why I picked ['memorial for people buried elsewhere' for] 'cenotaph.'" Unlike the L1 English users or the L1 Japanese users of English, the L1 Greek test takers could rely on their L1 in order to inform guesses at lower frequency words – that is, words at a frequency level that they would not typically be able to answer correctly.

Research Question 2. How do test takers qualitatively perceive and use the DK option?

Both test takers used the DK option only when they had no other choice, either due to lack of knowledge, lack of partial knowledge, or inability to apply test strategy to the item. The most common reason cited for using the DK option was "never hearing the word before".

Participants arrived at a DK option in one of two ways. First, participants used DK when they had no initial knowledge of a VST item. They cited reasons that illustrated a clear lack of knowledge: "I just didn't know what the answer was" or "Because I really had no idea…". Second, they used DK when they had cycled through their partial knowledge and potential test strategies, but still lacked sufficient knowledge to make an informed guess. In this condition, they skipped making an uninformed guess and instead chose the DK option. Participant A stated, "If I saw two or three [choices] that I didn't know or really had no idea, then I put 'I don't know.'" The following excerpt from Participant B's retrospective think-aloud during the interview additionally illustrates this test taking process:

35. ubiquitous: Many unwanted plants <are ubiquitous>.

a are difficult to get rid of
b have long, strong roots
c are found everywhere
d die away in the winter
e I don't know

> I was pretty sure it wasn't B. But when you see many unwanted plants everywhere, it's probably A or C. I think that would make sense 'cause it says "unwanted plants" . . . if that was associated back to "unwanted". So many unwanted plants are difficult to get rid of, that makes sense. Many unwanted plants are found everywhere. That also kind of makes sense, but I didn't know, and I got frustrated so I just picked E 'cause. . . it could be D, it could be C, it could be A.

Here, Participant B is clearly using partial knowledge of the target word "ubiquitous", followed by the test strategy of elimination and stem usage. Since he was unable to narrow the choices any further, he chose the DK option.

At the end of the testing period, both participants were asked if they would have used an "I don't know" option on the Original VST. Participant B volunteered: "Definitely . . . there was definitely many times when. . . it could be A, B, C, or even D."

The DK option is working as intended: it is not being overutilized, as it is chosen only when the test taker lacks sufficient knowledge; nor is it being underutilized, as it is chosen (in preference to make an uninformed guess) when the test taker lacks any knowledge of the correct answer.

Research Question 3. Do these perceptions and DK usage differ from the L1 American English and L1 Japanese users of English?

Finally, compared to the results of the pilot studies with an L1 American English user and intermediate-proficiency Japanese users of English, both L1 Greek–L2 English participants guessed more often; accordingly, compared to the L1 Japanese user of English, they used DK less often.

Participants A and B were similar to the L1 American English user (Lucovich, in press), who only used the DK options under the following conditions: They "did not know the meaning at all", had "heard the word before but [didn't] know the meaning", or it "[was] a word [they'd] never heard before".

4 Discussion

No penalty for guessing was included in the test taking directions, nor applied when scoring the test. As such, the directions replicate typical test taking at the program or classroom level. Since the directions do not disclose the scoring procedure, they neither overly encourage guessing, nor discourage it (as the penalty for guessing incorrectly is the same as an incorrect answer or NR).

Future administration of the VST at the program or classroom level should consider how much test strictness is necessary and how accurate of a vocabulary size estimate is needed in order to ascertain whether a penalty should be included. The testing conditions for this study result in a more accurate vocabulary size estimate as participants neither overused the DK option (i.e. using it when they could actually demonstrate knowledge of the word), nor underused it (i.e. they used it when they encountered an unknown word and could not demonstrate knowledge or apply a test strategy).

As this study investigated only two L1 Greek–L2 American English users, the findings cannot be broadly generalized to other populations or contexts. However, the findings are important as they are aligned with previous findings on the same topic and lend credence that test taking and DK usage on the VST are consistent among English users with higher education, regardless of L1 and despite varying English proficiency levels.

5 Conclusion

Based on the findings of this study in conjunction with previous studies, an interview code is being developed for use in future VST and multiple-choice test research. Utilizing this interview code, future research will examine whether these findings are consistent across larger sample sizes of populations and with lower-proficiency English users. Future work on the VST itself and other vocabulary tests could also focus on increasing their appeal, accessibility, and meaningfulness to language programs, classroom instructors, and language learners. Finally, further research should examine the effects of adding a DK option in other language learning and testing contexts, and its subsequent effect on guessing, test strategy, and score.

References

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

Bennett, P., & Stoeckel, T. (2012). Variations in format and willingness to skip items in a multiple-choice vocabulary test. *VERB*, *1*(2), 2–3.

Lucovich, D. (2014). Adding "I don't know" to the Vocabulary Size Test. In H. Brown & N. Sonda (Eds.), *JALT 2013 Conference Proceedings*. Tokyo: JALT.

Miles, M. B., & Huberman, A. M. (1994). Qualitative analysis: An expanded sourcebook (2nd ed.). Thousand Oaks, CA: Sage.

Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly*, *20*, 233–253.

Nation, I. S. P. (2012a, August). Measuring vocabulary size in an uncommonly taught language. Paper presented at the International Conference on Language Proficiency Testing in the Less Commonly Taught Languages, Bangkok, Thailand. Retrieved from http://www.sti.chula.ac.th/files/conference%20file/doc/paul%20nation.pdf

Nation, I. S. P. (2012b, 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*(7), 9–13.

Stewart, J., & White, D. (2011). Estimating guessing effects on the Vocabulary Levels Test for differing degrees of word knowledge. *TESOL Quarterly*, *45*, 370–380.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*(3), 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 3, Issue 2, September 2014
http://vli-journal.org

Article

Estimating Learners' Vocabulary Size under Item Response Theory

Aaron Gibson and Jeffrey Stewart
Kyushu Sangyo University
doi: http://dx.doi.org/10.7820/vli.v03.2.gibson.stewart

Abstract

Perhaps the most qualitatively interpretable vocabulary test score is an estimate of the total number of words the learner knows in the tested domain, such as a frequency word list, or vocabulary taught as part of a course curriculum. In cases where it is not possible to test the entire domain word-for-word, vocabulary tests such as the vocabulary levels test (Nation, 1990) and vocabulary size test (Beglar, 2010; Nation & Beglar, 2007) typically employ a polling method, in which total vocabulary size is inferred from a sample of tested words. A drawback of this method is that it assumes the tested words are randomly sampled from and therefore representative of the tested domain, which can affect test reliability in cases where there are many words in the domain that are far below or above learners' ability. This paper outlines an alternate method for estimating vocabulary size from a test score using item response theory, which allows estimation of total vocabulary size from a nonrandom sample of words well matched to learners' ability, resulting in tests of practical length with high reliability that can be used to estimate the total number of words a learner knows. Such a test scoring method, currently in use at a private university in southern Japan, is used as an example.

1 Background

Perhaps the most qualitatively interpretable vocabulary test score is one that estimates the total number of words the learner knows in the tested domain, such as a frequency word list, or vocabulary taught as part of a course curriculum. In cases where the tested domain includes very large numbers of words; however, it quickly becomes impractical to test learners on every word. Checklists, which simply ask learners to report their knowledge of words can be used, but such self-report measures may not be appropriate in cases where learners are assigned a grade, and there are concerns regarding the tested construct, which could be recognition of words rather than definitional knowledge (see Beeckmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001; Mochida & Harrington, 2006). Consequently, many popular vocabulary tests employ multiple-choice formats, or other formats that require the learner to provide the definition of the word, limiting the total number that can be tested within a typical class period.

In such cases, tests that report scores as total vocabulary size typically employ a polling method, in which total size is inferred from a sample of tested words. A

Please cite this article as: Gibson, A., & Stewart, J. (2014). Estimating learners' vocabulary size under item response theory. *Vocabulary Learning and Instruction*, *3* (2), 78–84. doi: 10.7820/vli.v03.2.gibson.stewart

drawback of this method is that it assumes the words are randomly sampled from the domain, which can affect test reliability in cases where there are many words in the domain that are far above or far below learners' ability. Under both classical test theory (Brown, 2005) and most common item response models (Parchev, 2004), tests have the highest reliability when the learner has a 50% probability of a correct guess. This may not be the case for tests of domains that include many high- or many low-frequency words however, leading to a "ceiling" or "basement" effects in scores.

Alternately, vocabulary size can be estimated under item response theory (IRT), which allows vocabulary size to be estimated on a common scale even in cases where learners take tests customized toward their respective levels of proficiency. This allows test makers to create tests with optimal internal reliability, even if the words used are not representative of the entire pool of words that their vocabulary size is estimated for. For example, the number of the first 5,000 most frequent words known by a low-level learner can be estimated by testing them on the items only from the first 2,000. Conversely, a high-level learner need only take a test of the 5,000 band to obtain comparable estimates, as so many higher frequency words are known by them that testing these items provides little information. This paper will outline this method, and describe how it was operationalized for test scoring at private university in southern Japan. It is hoped this paper can help curriculum organizers use similar scoring systems at their own institutions.

2 Context

As detailed in Fryer, Stewart, Anderson, Bovee, and Gibson (2011), curricula for first- and second-year English conversation classes at Kyushu Sangyo University ($n = 4{,}000$) include weekly lists of vocabulary selected from the 2,000 most common words in English as determined by analysis of the British National Corpus (BNC, 2007). The university's Language and Education Research Center (LERC) provides pre- and post-tests to assess student achievement on words taught each semester. Under the current curricula, first- and second-year students are tested on productive knowledge (Stewart, 2012a) of 150 words from the BNC 2,000 band each semester, for a total of 600 over the course of two years. Due to practical concerns however, only 30 of these words are tested as part of a larger vocabulary test at the end of each semester.

Semester tests are equated under the Rasch model, and until 2014, posttest scores were reported to students using a "scale score" based on Rasch ability estimates, with a score of 500 signifying the school average (for further details, see Stewart, 2012b). Although this method is convenient for test equating, a drawback of the logit-based scale score approach is diminished test score interpretability for both students and their teachers. Aside from relation to the school mean, what qualitative interpretation does a score, of, say, 600 signify for students, and in what way can it be interpreted by their teachers? Ultimately, what teachers desire is a precise estimate of the number of words on each semester word list known by students both before and after a semester of instruction rather than a statistical abstraction, regardless of how theoretically

sound the equating of scores is across test forms and student proficiency levels. The center resolved to use the following method to estimate the total number of words students learned each semester using the results of the relatively short tests.

3 Size Estimation under IRT

One of the simplest and most common IRT models, the Rasch model, is written as:

$$P_{ij} = \frac{e(\theta_i - b_j)}{1 + e(\theta_i - b_j)}$$

in which θ represents the ability level of the student, b represents the difficulty of the item on the same scale, and e is the natural log (2.171). Assuming the model fits, the formula can be used to determine the probability that student i will answer item j correctly.

Suppose a student with a theta ability level of 1 encounters the following words (corresponding hypothetical difficulties listed in parentheses):

union (1.25), *proper* (1.49), *mark* (−0.98), *separate* (1.32), *agreement* (1.1)

If the student's ability level is, for example, 1 logit, the differences between her ability and these (hypothetical) word difficulties are as follows:

$$-0.25,\ -0.49,\ +1.98,\ -0.32,\ -0.1$$

The logit differences can then be converted to probability of a correct answer to each test item using the above formula. Using a conventional Rasch model with a slope of 1, the probabilities of correctly providing a definition for each of the above words are as follows:

$$0.44,\ 0.38,\ 0.88,\ 0.42,\ 0.48$$

These probabilities can then be added together to estimate the expected value, i.e., the total number of these words that the student would be expected to know given their ability level, which in this case is approximately 2.59, or about half.

A useful feature of this model is that it is not necessary to test students on every word on a semester word list; with both the ability of a given student and the difficulty of each word on the semester list estimated on a common scale, it then becomes possible to calculate the probability the student knows each word on the list, including words that the individual student was not actually tested on (for an example, see Stewart & Gibson, 2010). Estimation of the total number of words known on a given list can be performed by determining the item pool score (Dorans, 2000), analogous to a computer adaptive test that estimates students' probability of success on hundreds of test items contained in an item bank after administering a subset of those test questions. To estimate the difficulties of words contained in semester word lists, multiple test forms are distributed to students, the sum of which test all the words contained on the list.

All of these test forms and their unique items are statistically equated, and logit difficulty estimates are calculated on a common scale for every word in the curriculum.

4 Method

Doing the above, however, requires giving multiple tests to multiple groups of students, in order to calculate the difficulties of all the respective words. As the class time available for tests must be kept to a minimum, it was necessary for word tests to be divided in such a way that each individual student wrote a subset of no more than 30 test questions. Therefore, items testing all 600 "breadth" words taught as part of the institution's first- and second-year English conversation classes were piloted using a total of 20 separate pretest forms over the course of two academic years and four semesters. Ten pretests for first-year students and ten pretests for second-year students (five for each semester) were created, each with an average of 15 items unique to the form and the remaining as common anchor items added for equating purposes.

For each semester of both years, five tests were first shuffled and then randomly distributed to participating students. This ensured each test form was administered to roughly equal sample sizes (an average of approximately 283 per form), and of roughly equal ability level (the school mean). After two years of data collection, semester test forms were then equated under IRT.

5 Analysis and Results

Data analysis was performed using the *ltm* package for R (Rizopoulos, 2006). Three item response models were tested for fit to the data: a 2-parameter logistic (2PL) model, in which item slopes are estimated separately; a constrained Rasch model, in which the item discrimination (slope) is presumed to be 1 for all items; and finally an "unconstrained" Rasch model, in which a single slope is empirically estimated for all items (typically close to the average slope of the same items under the 2PL model). This final model can not only offer superior model fit in cases where item slopes are relatively uniform, but also substantially higher or lower than 1.

Model fit comparisons conducted in *ltm* using the Akaike information criterion (Akaike, 1973) and Bayesian information criterion (Schwarz, 1978) indicated the unconstrained Rasch model had the best fit. The model's slope was estimated as 1.27, indicating that the tested items have relatively high discrimination.

After selecting the model, ability estimates were computed for the sample. Mean pretest item pool scores were calculated for low-, mid-, and high-level English conversation classes, in order to estimate the number of curriculum words known by students prior to instruction by class level. This was done using an excel spreadsheet (see Kim, 2004, for a comprehensive how-to guide). The spreadsheet calculates probabilities of knowing words by ability level, and sums totals to

			-3	-2.5	-2	-1.5	-1	-0.5	0
Enter Item Parameters Here			**5**	**8**	**14**	**22**	**34**	**47**	**63**
a = Item discrimination b = difficulty			**Probability of correct guess by ability level**						
			θ (ability level of student)						
	a	**b**	-3	-2.5	-2	-1.5	-1	-0.5	0
AC1.1.association	1.27	0.57	**0.01**	**0.02**	**0.04**	**0.07**	**0.12**	**0.20**	**0.33**
AC1.10.management	1.27	1.52	**0.00**	**0.01**	**0.01**	**0.02**	**0.04**	**0.07**	**0.13**
AC1.11.whose	1.27	-1.25	**0.10**	**0.17**	**0.28**	**0.42**	**0.58**	**0.72**	**0.83**
AC1.12.address	1.27	-0.83	**0.06**	**0.11**	**0.18**	**0.30**	**0.45**	**0.60**	**0.74**
AC1.121.structure	1.27	1.25	**0.00**	**0.01**	**0.02**	**0.03**	**0.05**	**0.10**	**0.17**
AC1.122.divide	1.27	1.49	**0.00**	**0.01**	**0.01**	**0.02**	**0.04**	**0.07**	**0.13**
AC1.123.clock	1.27	-0.98	**0.07**	**0.13**	**0.22**	**0.34**	**0.49**	**0.65**	**0.78**
AC1.124.relation	1.27	1.32	**0.00**	**0.01**	**0.01**	**0.03**	**0.05**	**0.09**	**0.16**
AC1.125.encourage	1.27	1.10	**0.01**	**0.01**	**0.02**	**0.04**	**0.06**	**0.12**	**0.20**
AC1.126.go	1.27	2.45	**0.00**	**0.00**	**0.00**	**0.01**	**0.01**	**0.02**	**0.04**

Figure 1. Spreadsheet to calculate the expected value of the ability level of students from the difficulty of the tested items.

Table 1. Mean Logit Ability and Estimated Number of Words Known Prior to Instruction for First-year First Semester Low-, Mid-, and High-level Classes

Level	Mean logit ability	Estimated number of words known
Low	−0.165	57.69
Mid	0.161	68.19
High	0.609	82.83

produce expected values. Figure 1 provides an example of such a file, abbreviated for brevity.

Next, mean student ability was calculated for each class level. This information was used to determine how many of the 150 words students knew by level before beginning the semester (see Table 1).

The results indicate that on average, first-year students have productive knowledge of 63 of these 150 words taught in the first semester prior to instruction. By level, low-level students know approximately 58 words; mid-level students know approximately 68; and high-level students know approximately 83. These estimates will be used as ipsative benchmarks when posttest results are assessed in July 2014; progress on the word list will be measured in reference to student knowledge of the words before entering the course. Our goal is to design the posttest section in such a manner that an estimate of words known that does not differ from pretest estimates will result in a low score, and only students estimated as knowing nearly all the words on the list will be able to attain a perfect score. By doing this, we will be able to report scores on this section of the final test to students and their teachers in a manner that is highly interpretable and has clear relevance to content mastery.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Academiai Kiado.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the yes/no vocabulary test: Some methodological issues in theory and practice. *Language Testing*, *18*(3), 235–274. doi:10.1177/026553220101800301

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

The British National Corpus. (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium (version 3) [BNC XML Edition]. Retrieved from http://www.natcorp.ox.ac.uk/

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. Upper Saddle River, NJ: Prentice Hall Regents.

Dorans, N. J. (2000). Scaling and equating. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 135–157). Mahwah, NJ: Lawrence Erlbaum Associates.

Fryer, L. K., Stewart, J., Anderson, C. J., Bovee, H. N., & Gibson, A. (2011). *Coordinating a vocabulary curriculum: Exploration, pilot, trial and future directions.* JALT2010 Conference Proceedings, Tokyo, Japan.

Kim, J. (2004). *An excel manual for item response theory.* Retrieved from http://coeweb.gsu.edu/coshima/EPRS8410/Sarah_Project1%2012%203%202004.pdf

Mochida, A., & Harrington, M. (2006). The yes/no test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*(1), 73–98. doi:10.1191/0265532206lt321oa

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

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

Parchev, I. (2004). *A visual guide to item response theory.* Retrieved from http://www.metheval.uni-jena.de/irt/VisualIRT.pdf

Rizopoulos, D. (2006). *ltm:* An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. Retrieved from http://www.jstatsoft.org/v17/i05

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

Stewart, J. (2012a). A multiple-choice test of active vocabulary knowledge. *Vocabulary Learning and Instruction*, *1*(1), 53–59. doi:10.7820/vli.v01.1.stewart

Stewart, J. (2012b). The LERC vocabulary program: Score gains for first and second year students. *Kyushu Sangyo University Language Education and Research Center Journal*, *7*, 87–93.

Stewart, J., & Gibson, A. (2010). Equating classroom pre and post tests under item response theory. *SHIKEN*, *14*(2), 11–18. Retrieved from http://jalt.org/test/ste_gib1.htm

Vocabulary Learning and Instruction
Volume 3, Issue 2, September 2014
http://vli-journal.org

Article

A Review of Four Studies on Measuring Vocabulary Knowledge

Akiyo Hirai
University of Tsukuba
doi: http://dx.doi.org/10.7820/vli.v03.2.hirai

Abstract

The purpose of this paper is to review each of the following four vocabulary studies: (1) Estimations of Japanese University Learners' English Vocabulary Sizes Using the Vocabulary Size Test (VST), by Stuart McLean, Nicholas Hogg, and Brandon Kramer; (2) Local Item Dependence on the Vocabulary Levels Test (VLT) Revisited, by Tadamitsu Kamimoto; (3) Test Taking and DK Use on the VST, by Dawn Lucovich; (4) Estimating Learners' Vocabulary Size under Item Response Theory (IRT), by Aaron Gibson.

1 Estimations of Japanese University Learners' English Vocabulary Sizes Using the Vocabulary Size Test

McLean, Hogg, and Kramer investigated university students' vocabulary sizes using the VST (Nation & Beglar, 2007). They found that the VST seriously overestimated vocabulary size. They showed that even the low-level students maintained a score of approximately 20% correct at the 5,000 vocabulary level and higher, instead of decreasing linearly as the difficulty of the vocabulary level went up. A major cause for this result is the VST's four-option multiple-choice (MC) format, which allows for a high probability of guessing correctly. Another cause is students' familiarity with English loanwords in Japanese. Although many loanwords are relatively low frequency words, they can be easier for Japanese students because they are already familiar with the pronunciation and meanings of the words from Japanese.

Besides these major findings, the authors raised the interesting question of whether university departments' *hensachi*, or *T*-score, can be a useful proxy for English vocabulary size. A *hensachi* score is a standard score used to rank which percentile students or schools are in. *T*-scores always have an average of 50 and are calculated using the following formulas:

$$T\text{-score} = 10\ (z\text{-score}) + 50$$

$$[z\text{-score} = (\text{a raw score - average}) / \text{standard deviation}]$$

To answer this research question (RQ), the authors took the correlation between departments' *hensachi* scores, which are widely available on the Internet, and 3,427 students' vocabulary sizes. They found a relatively high correlation of .73, and

Please cite this article as: Hirai, A. (2014). A Review of Four Studies on Measuring Vocabulary Knowledge. *Vocabulary Learning and Instruction*, *3* (2), 85–92. doi: 10.7820/vli.v03.2.hirai

therefore concluded that the departments' *hensachi* are a fair indicator of Japanese university students' relative vocabulary size.

Such a high correlation is reasonably expected because the departments' *hensachi* represents students' general proficiency levels. Vocabulary knowledge is widely regarded as a major factor in English proficiency (e.g., Milton, 2009, pp. 175–176; Stæhr, 2008). In particular, the relationship between vocabulary size and reading ability, which is the main skill tested by most Japanese university entrance examinations, is the strongest among the four language skills. According to Stæhr (2008), as much as 72% of the variance in the ability to score an average mark or above on the reading test can be explained by vocabulary size. Thus, simply proving the correlation coefficient is not enough to show the significance of this study. Besides, their rich data deserve further explanation.

One idea to make the study more significant and informative is to draw a best-fit regression line from *hensachi* to vocabulary size and create a conversion table. It would be helpful in grasping students' vocabulary size if the conversion table provided vocabulary sizes that are equivalent to *hensachi* scores, starting from the lowest score, i.e., around 40, and progressing upward in small steps, until the highest, around 80. A teacher might want to know, for example, the mean vocabulary size that is equivalent to a departmental *hensachi* score of 60 when she wants to decide on a textbook to choose before the class starts. At present, *hensachi* scores are used for classifying only three proficiency groups, which seems to be a bit too vague and uninformative. A group with *hensachi* below 50 is categorized as the Low group, and approximately one-third of students fall into that category. This category should be further divided to identify lower-level students' vocabulary sizes. Since a large sample of data were collected from all over the country, a relatively accurate regression line for the conversion table may be obtained.

In relation to the validity of vocabulary size estimation derived from *hensachi* scores, the inclusion of VST scores for first-year or at most second-year university students alone may produce a more precise relationship between *hensachi* and vocabulary size. It is possible that third- and fourth-year students' vocabulary knowledge has changed since leaving high school. Another related issue concerns the correlation coefficients between the VST scores and other test scores, such as the *Test of English as a Foreign Language* (TOEFL) and *Test of English for International Communication* (TOEIC) tests. Because the sample sizes for these correlations are small, it would have been better to also provide the *hensachi* level and vocabulary size of the students.

Lastly, McLean, Hogg, and Kramer's study deserves credit for collecting such a large sample of students and revealing that Japanese university students' English vocabulary sizes are wide-ranging. They have a mean of 3,715 words, and the growth of their vocabulary knowledge is not parallel to the frequency lists. Thus, the selection of classroom material must be made with care, depending on their majors and needs, in order to provide appropriate instruction.

As in the case of the VST, the Vocabulary Levels Test (VLT; Nation, 1990) seems to have a similar tendency to overestimate test takers' vocabulary knowledge, which is pointed out by the next study.

2 Local Item Dependence on the Vocabulary Levels Test Revisited

The second study was done by Kamimoto, who claims that LID arises as a result of the matching format of the VLT. To examine the LID, he compared the original test of three 6 3 (i.e., 6 words and 3 definitions translated in Japanese) matching forms and the combined 18 9 (i.e., 18 words and 9 definitions) form.

LID should be avoided since it violates local independence, one of the assumptions of both classical and item response testing theories. Local independence means that "when the abilities influencing test performance are held constant, examinees' responses to any pair of items are statistically independent" (Hambleton, Swaminathan, & Roger, 1991). In this regard, the format of the VLT seems to be a problem since after one or two prompts have been answered from the six options, and the number of options for the last definition is reduced. In other words, these definitions and options influence each other. Thus, to reduce the local dependence, Kamimoto increased the options and definitions by three times as many as were on the original form.

However, if the matching format is the source of the LID, the 18 9 form is still regarded as a matching format and, strictly speaking, the LID still exists. Thus, as shown in Figure 1, it would be worth comparing the original 6 3 form with 6 1 or 4 1 form (i.e., MC formats). For example, as shown in Figure 1, one set of the 6 3 form can be divided into three sets of 6 1 forms for the same target words, and each set can be presented to different groups. Thus, six sets of the 6 3 form can be broken down as follows:

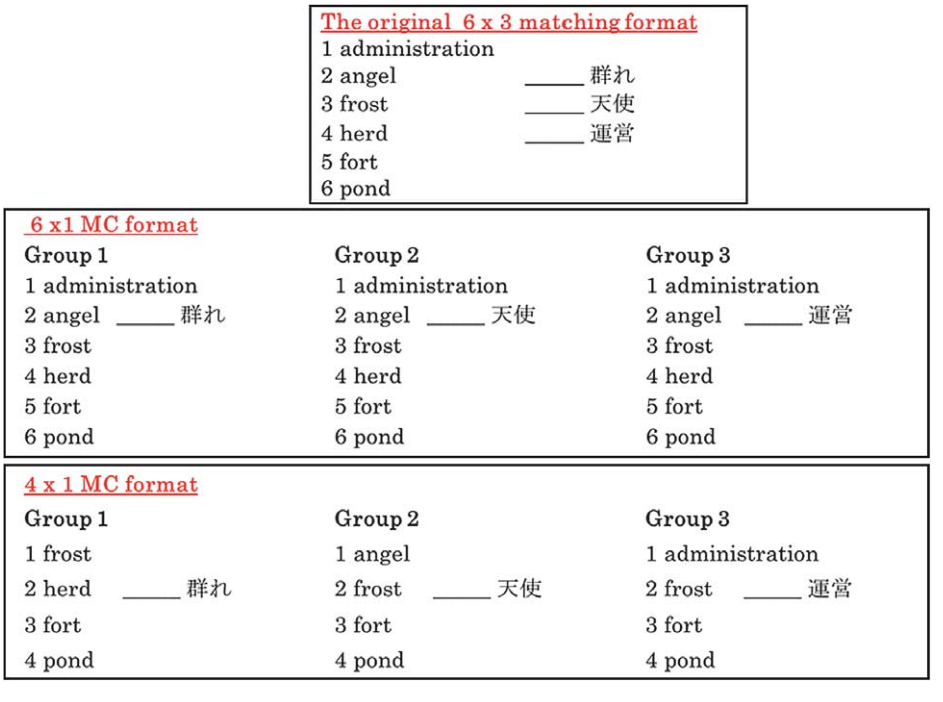

Figure 1. Alternative designs for comparing the original matching form with MC forms.

Group 1 takes six sets of the 6 1 form (the six options and the first prompt of Forms A and B); Group 2 takes six sets of the 6 1 form (the six options and the second prompt of Forms A and B); and Group 3 takes six sets of the 6 1 form (the six options and the third prompt of Forms A and B). A couple of weeks later, all the groups take the original 6 3 form (Forms A and B). By comparing the original form with the MC responses, guessing produced by the LID could be more clearly shown.

As for the Item Facility (IF) analyses of the 6 3 form and 18 9 form, the Type 2 IF index, which represents responses answered correctly on the 6 3 form but incorrectly on the 18 9, and the Type 3 IF index, which indicates responses answered incorrectly on the 6 3 form but correctly on the 18 9 form, are very informative. Through these analyses, the author successfully detected high guessing items, such as the words "spoil," "stool," and "deficiency," and pinpointed the causes of the high guessing of these items by examining the other options and definitions in the sets. Thus, the Types 2 and 3 IF analyses can be used diagnostically to replace problematic distractors in the original VLT test.

Following the detailed individual item analyses, the author summarized the responses categorized in the Type 2 IF index and found an average guessing rate of 19%. However, the responses categorized in Type 3 should also be summarized because these responses seem to come from purely random guessing. By comparing the guessing rates for these two types, the author could show more clearly how much more the 6 3 form allows for guessing than the 18 9 form does.

With such a high guessing rate (i.e., the mean rate of Type 2 responses) in the original VLT, Kamimoto concluded that items in clusters tend to be more dependent than the previous Rasch-based studies indicate (e.g., Schmitt, Schmitt, & Clapham, 2001). However, it is worth noting that the VLT the authors used for this study was the Japanese version of the original VLT, while Schmitt et al. (2001) used revised versions. Such a difference may have led to somewhat different results, besides the difference in the use of Rasch analysis. Although it is difficult to make clear whether the high guessing rate of 19% is due to LID or simply the difference in the probability of guessing from 6 options or 18 options, this study is significant in that it shows clearly a high level of guessing that is hard to ignore.

As the above two studies pointed out, test format is a major influence on the item difficulty of a trait measured. Both the MC format and the matching format were found to allow test takers to answer correctly using various types of guesswork, such as informed guessing and uninformed (i.e., random) guessing. From the viewpoint of the washback of a test, a good item is one that encourages test takers to interact with it actively, making them try hard to retrieve their target knowledge. In this regard, the act of informed guessing, the process in which test takers use their partial knowledge while guessing, is preferable to uninformed guessing. The next study deals with reducing only the latter type of guessing.

3 A Qualitative Analysis of the Options on the VST

Lucovich has suggested the addition of a "don't know" (DK) option to the current four-option items of the VST to reduce test takers' uninformed guessing.

To investigate the effect of the DK option, she addressed the following RQs: (1) How did two Greek advanced learners of English (as examples of non-Japanese/non-English test takers) determine their answers on the VST? (2) How did they qualitatively perceive and use the DK option? (3) Did the two Greek participants differ in the choice of the DK option from the first language (L1) American English and L1 Japanese users of English?

Regarding RQs (1) and (2), the two participants took two versions of the 100-item VST (one without the DK option and one with the DK option), and their responses were coded through the test takers' retrospective think-aloud protocol. Lucovich reported that 4% of the items were categorized as "uninformed guesses," which implies that even with the DK option included, it is difficult to eliminate uninformed guessing completely. However, the effect of the DK option was still clear because 12% of the items were coded as "DK usage." In addition, the interview data reveal that the test takers chose the DK option ONLY when they did not have any knowledge of the correct answer. In this way, Lucovich successfully showed that a large proportion of purely uninformed guessing could be avoided with the DK option. By categorizing the answers of the VST and using an interview technique, Lucovich shed light on psychological aspects of test-taking behavior that could not be revealed by quantitative methods alone.

However, to make her study more interpretable, a couple of points are recommended. First, the focal point of this study is the difference in test-taking strategy between tests with or without the DK option. Thus, it would be better to show the coding results of both versions of the test separately for comparison. Second, the four figures of the results of the two tests by the two participants are informative. Thus, for the sake of comparison, each participant's two figures should be placed next to each other. Moreover, all the figures should mark off the same divisions on the scale of the *x*-axis. In her draft, those divisions are slightly different across the four figures (for example, "0, 3, 6, 9, 15" and "0, 3, 6, 10, 13, 16" for Participant A and "0, 3, 6, 9, 15" and "0, 3, 5, 8, 11, 13, 16" for Participant B). In addition, the figures need to be explained. Likewise, careful layout of the tables and figures would help readers to interpret them correctly.

Concerning RQ (3), Lucovich reported that the two Greek participants used more DK options than the intermediate-proficiency L1 Japanese users of English she had analyzed in her previous study (Lucovich, 2013). It would have been helpful to have identified whether this difference was the result of differences in proficiency, personality, or test-taking conditions. She also should have provided more information about the participants in the previous study in the Introduction of this current study if such a comparison was to be made.

With regard to the limitations of the study, as Lucovich acknowledges, it is difficult to generalize from the results of this study because there were only two Greek participants, and they had similar proficiency levels. Thus, to determine the factors affecting the use of the DK option, research that is more comprehensive is required, perhaps by combining the current study with the previous one in which the L1 Japanese and L1 English users participated, and by adding some less proficient learners of English.

Zhang (2013) argues that whether the DK option works well or not may depend on the level of the stakes. If test takers' scores reflect their grade, they may not dare to choose the DK options unless some penalties are added. If, however, the scores are used for placement or diagnostic purposes, without affecting test takers' grades, the DK option may work well, and a more accurate estimation of their vocabulary size would be obtained. Therefore, it might be an interesting topic to explore what kinds of instructions and/or penalties are necessary to make the DK option function properly, even in a high-stakes situation.

4 Estimating Learners' Vocabulary Size under Item Response Theory (IRT)

The last study, by Gibson, is about an alternative method for estimating vocabulary size using IRT. The equating of several test forms using IRT and the estimation of all the difficulties of the items on the same scale merit attention because learners' vocabulary knowledge can be estimated using a smaller number of words than could be done by a test without using IRT. In addition, words fit to the test taker's ability can be selected in a non-random way for a test as long as these words have been calibrated.

Gibson deserves particular credit for the following three points. First, he adopted a counterbalanced design in a careful manner. The five forms to be pre-tested each semester were shuffled and distributed randomly to students, even within a single class. Consequently, the sample size and test takers' ability levels were equal across the test forms, which made more precise estimation of items possible. Second, he tried to make IRT scores (i.e., logit values) more interpretable for teachers and students. He calculated the estimated number of words known out of 150 words to be taught in a semester, based on "the probability of correct guess by ability level." For example, a student with a logit value of zero, which is 50% probability of answering the total number of words correctly, is estimated to know 75 words out of 150 words. The third point is that he used a productive vocabulary test called "Active MC," in which students identified the first letter of a word by looking at L1 definitions, the part of speech, the second and third letters in the word, and the number of letters contained in the entire word (Stewart, 2012a). Assessing students' productive vocabulary knowledge is more valid than assessing receptive knowledge because the test takers are students in English conversation classes who learn these words productively. In addition, the use of a receptive vocabulary test such as the VST and VLT would overestimate learners' mastery of words as mentioned by three studies above.

Although estimating learners' vocabulary size by using IRT has great potential, there are some drawbacks. First of all, under IRT, all the words need to be calibrated beforehand, which would take time. Second, related to the first point, students' vocabulary size can only be estimated with the words calibrated. Therefore, a learner's total vocabulary size cannot be measured as it can be by using the frequency-based VLT or VST. In this study, the total vocabulary size is 150 words, not 1,000 or 2,000 words. However, it is not impossible to estimate the total size out of a frequency level of 2,000 if all 2,000 words have been calibrated. One way to do this is to expand the item bank that stores calibrated items bit by bit.

Whenever a pretest is given to students at the beginning of each semester, some new items can be mixed with the calibrated items, similar to the way that *Educational Testing Service* (ETS) includes some unscored listening or reading items in TOEFL tests. Later, these items can be calibrated by anchoring them to the rest of the calibrated items.

Since the aim of Gibson's study is to outline the equating procedures of an in-house testing project, it is difficult to understand the whole picture of the equating project without reading other relevant papers, such as Stewart (2012a, 2012b) and Stewart and Gibson (2010). However, the project seems to be promising and certainly helps teachers who want to minimize testing time but still get an accurate evaluation.

5 Conclusion

One important issue commonly raised in these studies was the overestimation of vocabulary size under the current VLT and VST. Some measures may need to be taken to deal with this problem, such as improving the format or individual distractors, adding a "DK" option, penalizing random guesswork, deducting a certain percentage from the score obtained (i.e., estimated vocabulary size), or using a different method such as IRT. Given that it is impossible to create one single perfect test, as these studies have suggested, we need to improve or evaluate currently available vocabulary tests such as the VLT and VST by considering the purpose of their use and aspects of their assessment, including validity, reliability, washback, and practicality.

References

Hambleton, R.K., Swaminathan, H., & Roger, H.J. (1991). *Fundamentals of item response theory (Vol. 2)*. London, UK: Sage.

Lucovich, D. (2013). The inclusion of "I don't know" on the Vocabulary Size Test. *Tokyo JALT Journal*, *1*, 28–31.

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

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

Nation, I.S.P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31* (7), 9–13.

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

Stæhr, L.S. (2008). Vocabulary size and the skills of listening, reading and writing. *Langauge Learning Journal*, *36* (2), 139–152, doi:10.1080/09571730802389975

Stewart, J. (2012a). A multiple-choice test of active vocabulary knowledge. *Vocabulary Learning and Instruction*, *1* (1), 53–59. doi:10.7820/vli.v01.1.stewart

Stewart, J. (2012b). The LERC Vocabulary Program: Score gains for first and second year students. *Kyushu Sangyo University Language Education and Research Center Journal*, *7*, 87–93.

Stewart, J., & Gibson, A. (2010). Equating classroom pre and post-tests under item response theory. *SHIKEN*, *14* (2), 11–18.

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