Article

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

Kurtis McDonald[a] and Mayumi Asaba[b]

[a]Kobe College; [b]Kwansei Gakuin University

Abstract

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

1 Background

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

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

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

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

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

1.1 *Bilingual Versions*

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

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

Figure 1. Example Bilingual Question.

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

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

1.2   Guessing

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

1.3   The "I Don't Know" Option

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

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

2 Method

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

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

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

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

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

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

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

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

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

3 Findings and Discussion

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

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

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

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

3.2   How Do Learners Answer Self-identified Unknown Items?

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

(1) *True partial knowledge-informed guess*: A guess based on accurate self-perceived knowledge of the target word, one or more of its word parts, or a concept clearly associated with the word.
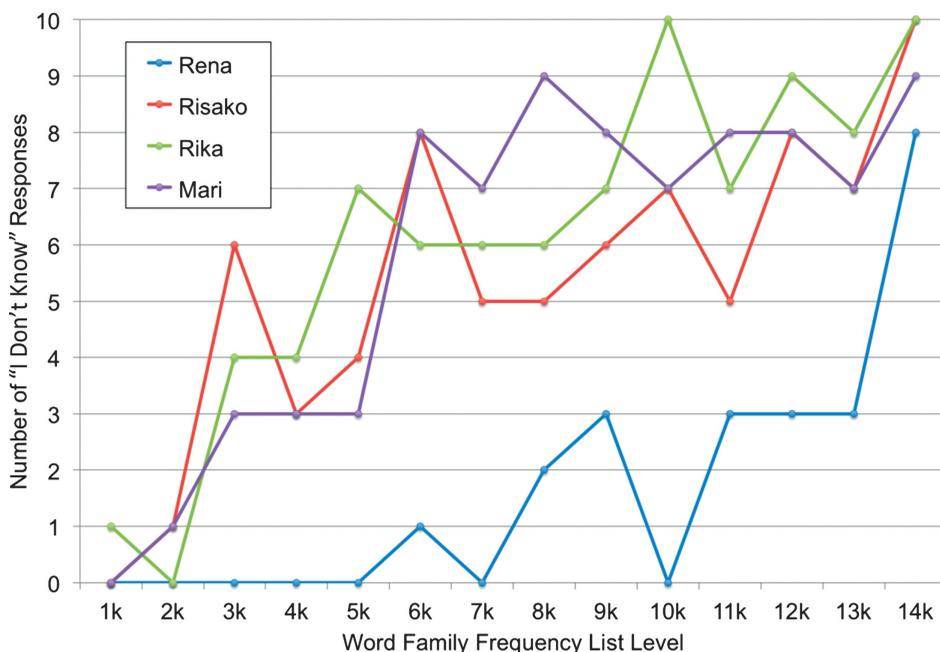
Figure 2. Number of "I don't know" responses for each word family frequency list level on the bilingual Japanese VST administered in January ($k = 140$).

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

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

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

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

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

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

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

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

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

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

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

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

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

Note: $k = 140$.

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

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

4 Conclusion

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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