Vocabulary Learning and Instruction
Volume 5, Issue 1, December 2016
http://vli-journal.org

Commentary

"I Don't Know" Use and Guessing on the Bilingual Japanese Vocabulary Size Test: Clarifications and Limitations

Kurtis McDonald[a] and Mayumi Asaba[b]
[a]Kobe College; [b]Kwansei Gakuin University

doi: http://dx.doi.org/10.7820/vli.v05.1.mcdonald.asaba

Abstract

This paper offers a response to Hutchinson's comments on our preliminary report of "I don't know" use and guessing on the bilingual Japanese Vocabulary Size Test (VST), which was published in *Vocabulary Learning and Instruction*. In particular, it provides greater clarification of the English proficiency levels used throughout that paper and a reiteration of what we see as its key findings regarding the range of vocabulary size estimates that were able to be calculated for the participants. Finally, it addresses the methodological limitations of the original study, which, we believe, reduce any determinations about the participants' personality types or general test-taking behaviors to mere speculation.

1 Introduction

In this paper, we offer a response to Huchinson's (2015) comments on our preliminary report of "I don't know" use and guessing on the bilingual Japanese Vocabulary Size Test (VST; Nation & Beglar, 2007), which was published in *Vocabulary Learning and Instruction* (see McDonald & Asaba, 2015). Not only do we welcome the chance to provide greater clarification of what we see as the key findings of that paper, but we also value the opportunity to again promote greater awareness to several test factors known to impact the validity and reliability of the vocabulary size estimates garnered through the use of the VST with second language (L2) learners. Such factors include determinations on what word family frequency levels should be included, whether a monolingual or bilingual version is most appropriate, and how effective test directions, announced penalties, and/or the inclusion of an "I don't know" option can be in deterring random, uninformed guessing. We hope that revisiting some of these issues in this paper can again highlight the need for potential users of the VST to consider how they can best ensure that the test will align the purposes for which it is intended.

In our initial report, we presented the findings of a small scale, qualitative study of the test-taking behavior of four first-year, Japanese university students who completed the 140-item bilingual Japanese VST (translated by Sasao and Nakata and available on Paul Nation's website) modified to include a fifth "I don't know" answer option in two passes. On the first pass, the students were explicitly told to select the "I don't know" option on all items that they were unsure about. Then, on

Please cite this article as: McDonald, K., & Asaba, M. (2016). "I don't know" use and guessing on the bilingual Japanese vocabulary size test: Clarifications and limitations. *Vocabulary Learning and Instruction*, *Advance online publication*, 34–39. doi: 10.7820/vli.v05.1.mcdonald.asaba

the second pass, the students were told to return to all of these self-identified unknown items and to select the best answer from the original four multiple-choice options. Individual retrospective interviews were subsequently carried out with the four participants to ascertain how these test takers eventually arrived at their guesses to these self-identified unknown items. The recalled thought processes elicited in the interviews were then used to code the answer selections as *informed guesses* based on *true partial knowledge*, *false partial knowledge*, and/or *test strategies* or completely *uninformed guesses* chosen at random. This information allowed five distinct scores (vocabulary size estimates when multiplied by 100) to be calculated for each participant as the various types of guesses identified were selectively excluded or included. The range of scores provided for each individual suggested to us that there may be little difference in the vocabulary size estimates that either include or exclude all guesses for higher proficiency learners but a great deal of difference between these estimates for lower proficiency learners.

In his commentary, Huchinson (2015) raised concerns about the conclusions we reached in that article, offered an alternative interpretation of the data from the participant we considered the most proficient, and discussed a call for further research at the interface between test-taking behavior and test-taker confidence levels to be more theoretically driven. Hutchinson concluded by pointing out that a possible starting point would be developing a greater understanding of reluctant responses and second attempts, as our original study sought to explore. In what follows, we hope to address what appears to be two critical misunderstandings of our initial report and outline the methodological limitations that prevent us from drawing any firm conclusions about the participants' personality types or the reasoning they may have applied when deciding whether or not to respond to items on their first passes through the test.

2 Clarification of Proficiency Levels

The first area of concern raised by Huchinson (2015) seems to be based on a misunderstanding of the proficiency levels used in our study. Throughout the initial report, we exclusively used an average of the overall scores from two administrations of the Test of English for International Communication Institutional Program (TOEIC (IP)) conducted in April 2014 and February 2015 as an external measure of proficiency for the participants, a point which we stated explicitly on page 20. The TOEIC (IP) test is widely used as a proficiency measure in Japan and university students' scores from this test are often used in level placement decisions (TOEIC Steering Committee of the Institute for International Business Communication, 2008), as is the case at the institution attended by the participants in our study. The TOEIC (IP) test comprises two sections, Listening and Reading, which are equally scored out of 495 points and are typically reported as a combined score out of 990. To give further clarity, an enhanced version of Table 1 from our original study is reproduced below with all of the scores listed. General descriptors for various score levels on each subsection of the test can be found athttps://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_Score_Desc.pdf

Given the noted importance of L2 vocabulary knowledge to overall L2 proficiency (Meara, 1996), a strong positive relationship between a learner's

English vocabulary size and her total score on the TOEIC (IP) would be expected. Furthermore, as the VST was designed to test written receptive vocabulary knowledge, valid and reliable scores from this test would be expected to most highly correlate with those from the Reading subsection, particularly given the close ties that have been identified in the literature between vocabulary knowledge and reading comprehension (Grabe, 2009; Hu & Nation, 2000; Nation, 2006). Viewed in terms of either their total scores or their Reading subsection scores, we believe that the participants' average TOEIC (IP) scores provide a strong and widely interpretable proxy for their English proficiency levels as described in the original study and listed again in Table 1 by the order of ability: from the most proficient to the least proficient.

3 Clarification of the Findings

Although it remains unclear how much of Huchinson's (2015) subsequent claims stem from this lack of understanding about the external proficiency measures used in the original study, another key misunderstanding seems to lie in his interpretation of the findings we discuss in relation to our third research question: "How much do vocabulary size estimates differ when different approaches to scoring guesses are applied?" (McDonald & Asaba, 2015, p. 17). For reasons that remain unclear to us, Hutchinson claims that we are "giving precedence to the score without guesses" and "going beyond the data" (p. 50). Let us begin our clarification of these findings by revisiting the data we used to answer this research question, reproduced in Table 2.

Looking at how the scores for each individual increase across the table from left to right, as more and more guesses based on reasoning less and less demonstrative of actual knowledge of the target words, led us to draw two sides of the same conclusion from this very small sample of participants. While the scores from the strictest end of the spectrum (the scores without guesses) and those from the most sensitive or lenient end of the spectrum (the scores with all guesses) varied little for the most highly proficient student, Rena (+7 points), they varied a great deal for the three other less proficient students: Risako (+35 points), Rika (+32 points), and Mari (+31 points).

Besides pointing out the striking difference between the variation exhibited by Rena and the other participants in the scores at the two ends of the continuum, we hoped to illustrate the potential risks of relying on either of these scores alone as an

Table 1. Summary of Participant Details and TOEIC (IP) Scores

Participant	Major	TOEIC (IP) total score average	TOEIC (IP) listening average	TOEIC (IP) reading average
Rena	English	775	410	365
Risako	Intercultural Studies	543	288	255
Rika	Psychology	380	230	150
Mari	Bioscience	305	200	105

Note. The TOEIC (IP) total score average listed represents the average from two administrations of the test conducted in April 2014 and February 2015.

Table 2. Scores for Each Participant on the Bilingual Japanese Vocabulary Size Test

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. Each score can be multiplied by 100 to arrive at a vocabulary size estimate out of a maximum total of 14,000 word families possible (Nation, 2012).

indicator of a learner's English vocabulary size, especially given that the sampling which underlies the selection of words tested on the VST means that each item answered correctly is multiplied by 100 to arrive at a learner's vocabulary size estimate. As the original test instructions for the VST make no accommodation for limiting or penalizing guessing of any kind (Nation, 2012), scores similar to those in the rightmost column would be expected for these learners if this were the only version administered. Classified by these estimates alone, an instructor might conclude that the English vocabulary sizes of these learners would all be in the 8,100–9,100 word family range and that they should all be comparably able to handle texts written using only the 8,000 most frequent word families of English. However, if a version of the test modified to include an "I don't know" option were given to students who conscientiously followed the directions stressing its use for items which they were unsure about, scores similar to those in the leftmost column would be expected from these learners. Classified by these estimates alone, an instructor might conclude that two different sets of reading texts would be required to appropriately match the two distinct tiers of vocabulary size estimates garnered (8,400 word families for Rena and 5,000–5,400 word families for the others).

Rather than giving precedence to either set of these rather blunt estimates, we expressly suggested that the scores listed in this paper in bold in the third column of Table 2 would seem to be the "most sensible figures" to consider since they were based only on answers derived from the kind of vocabulary knowledge that the learners would be most likely to benefit from if they were to encounter these words while reading (McDonald & Asaba, 2015, p. 23). This determination was made following arguments like that of Nagy, Herman, and Anderson (1985), which posit that since word knowledge accrues incrementally and no vocabulary test is able to capture all aspects of this knowledge, a demonstration of partial knowledge should be acknowledged as acceptable evidence of word recognition. To put the rationale behind our preference another way, we believe that the scores in this third column are the ones that would seem to be most relevant to the construct of written receptive vocabulary knowledge that were able to be obtained given the test instructions and research methodology that we employed. Not only does this preference seem best aligned with the underlying construct intended to be measured, but it also seems to negate much, if not all, of Huchinson's (2015) concern that certain personality types would be unfairly disadvantaged if their guesses were entirely excluded since all correct answers arising from either initial *self-perceived knowledge* or subsequent guesses based on *true partial knowledge* contributed equally to these resulting scores.

4 Limitations

One final area that we would like to address in this response involves acknowledging more directly some of the limitations inherent to the original study. As Huchinson (2015) and the reviewers of our original article have pointed out, there are several key questions that remain unresolved in relation to each participant's interpretation and subsequent application of our first pass test instructions, which explicitly directed them not to guess on items that they were unsure about. We completely agree that this is a key component missing from the original study which is certainly worthy of greater investigation. However, although Hutchinson's reorganization of our results in his Table 1 allows for a worthwhile reconsideration of the results, which does indeed hint at likely differences in the ways the participants responded to items during their first passes through the test, we believe that the methodology we employed, with retrospective interviews only about the items that the learners self-identified as unknown, severely limits any further discussion in this area to mere conjecture and speculation. Given these limitations, any differences potentially attributable to each participant's individual interpretation of the test instructions, threshold level of what constitutes "sure"-ness, susceptibility to faulty intuitions and/or deceptive transparency (Laufer, 1997), and/or personality type remain unclear. A much more rigorous and involved series of interviews aimed at uncovering the conscious reasoning applied to all 140 answer selections would be required before we would feel at all comfortable drawing any tentative conclusions here. Indeed, even if such a methodology were carried out, it seems highly unlikely that it would provide enough information to support the sort of conclusions that Hutchinson already seems more than willing to suggest regarding the participants' English proficiency levels and personality types, much less the occupations which they may be best suited for in the future.

5 Conclusion

We hope that the clarification of the proficiency levels used throughout our original study and the reiteration of the level of sensitivity we feel is most closely linked to the underlying construct targeted by the bilingual Japanese VST that are offered in this response allows for the main points of that paper to be better understood. We appreciate Huchinson's (2015) concern that test takers with certain personality types may achieve better scores than they would have otherwise received if they were advised to select the "I don't know" option while others may, in turn, receive worse scores than they would have otherwise. However, this concern seems misplaced in relation to our study since we advocated that the scores that include both initial self-perceived knowledge and true partial knowledge would seem to be the most valid of the scores that we were able to tabulate. While there are clearly many important unresolved questions regarding the validity of estimates garnered from the bilingual Japanese VST, we acknowledge that the most important takeaway from our preliminary report is a simple warning that the validity of test scores that do not take factors like test language and non-construct-related guessing into account may be seriously threatened, particularly for lower proficiency learners. These issues are certainly worthy of further consideration and constructive debate.

References

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice.* New York, NY: Cambridge University Press.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430.

Huchinson, T. P. (2015). Low-confidence responses on the vocabulary size test. *Vocabulary Learning and Instruction*, *4*(2), 49–51. doi:10.7820/vli.v04.2.hutchinson

Laufer, B. (1997). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 20–34). Cambridge: Cambridge University Press.

McDonald, K., & Asaba, M. (2016). "I don't know" use and guessing on the bilingual Japanese vocabulary size test: Clarifications and limitations. *Vocabulary Learning and Instruction*, *5*(1), 1–6. doi:10.7820/vli.v05.1.mcdonald.asaba

Meara, P. M. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Performance and competence in second language acquisition* (pp. 35–53). Cambridge: Cambridge University Press.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, *20*(2), 233–253. doi:10.2307/747758.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82. doi:10.1353/cml.2006.0049

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

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

TOEIC Steering Committee of the Institute for International Business Communication. (2008). Trends survey of TOEIC® test utilization 2007. Retrieved from http://www.toeic.or.jp/library/toeic_data/toeic_en/pdf/data/TOEIC_Utilization_2007.pdf