

Adding *I Don't* Know to the Vocabulary Size Test

Dawn Lucovich
Teachers College Columbia
University, The Writing
Center

Reference Data:

Lucovich, D. (2014). Adding *I don't know* to the Vocabulary Size Test. In N. Sonda & A. Krause (Eds.), *JALT2013 Conference Proceedings*. Tokyo: JALT.

In this pilot study, an *I don't know* (DK) option was added to the 20,000-word family, 100-item Vocabulary Size Test (VST; Nation & Beglar, 2007). The participant (current MA student, L1 American English) took the original and modified versions of the VST. In a semi-structured qualitative interview, based on Nagy, Herman, and Anderson's (1985) word interview protocol, the participant was asked about the answers given, items on the VST, and test-taking strategies. The study was aimed at uncovering how a test taker determined answers, and how she qualitatively perceived and used the DK option. The participant only used the DK option if all other answering strategies (such as partial knowledge or process of elimination) were not available. Future research should further investigate the VST as a learning and testing tool, the addition of a DK option to other multiple choice tests, and how DK usage differs between test takers.

この予備研究では、頻度上位20000語の中の100項目から構成される語彙サイズテスト (VST; Nation & Beglar, 2007) に「I don't know (わからない) (DK) という選択肢を追加した。米語を第一言語とする大学院生1名がこの研究に参加し、VSTとその修正版の両方を受けた。半構造的インタビューでは、ナギー・ハーマン・アンダーソンによる語彙インタビュープロトコール(1985)に基づき、被験者に自身の回答や、VSTの項目、そしてテスト方略について尋ねた。この研究は、テスト受験者がどのように回答を決定し、またDKオプションをどう質的に捉え活用したかを明らかにすることを目的とした。その結果、被験者は他の全てのテスト方略(部分的な知識や消去法)が使用不可能な時のみDKオプションを使うことが分かった。今後の研究では、学習や試験のツールとしてのVST、他の多肢選択試験でのDKのオプションの追加、そして受験者間におけるDKオプションの活用の違いを明らかにする必要がある。

DUE TO the importance of vocabulary in language learning, vast classroom, material, and teacher resources are invested in vocabulary. Although the methods of teaching and learning vocabulary are actively and thoroughly researched, methods of testing vocabulary learning are generally polarized into short, low-stakes classroom quizzes and high-stakes standardized testing such as TOEIC (Test of English for International Communication) and TOEFL. Therefore, a research agenda or curriculum design should prioritize identifying, selecting, and refining a vocabulary test that occupies a middle ground between a low-stakes quiz and a high-stakes exam. Furthermore, the use of such a test to assess learner growth and effectiveness of instruction, to pinpoint current gaps in knowledge, as well as to identify possible directions for future learning and instruction should be investigated.

There are several tests of written receptive vocabulary, including the Eurocentres Vocabulary Size Test (Meara & Jones, 1988, 1990), the Word Associates Test (Read, 1993, 1998; Schmitt,



Ng, & Garras, 2011), and the Vocabulary Levels Test (Laufer & Nation, 1999; Nation, 1993; Schmitt, Schmitt, & Clapham, 2001). The Vocabulary Size Test (VST; Nation & Beglar, 2007) has emerged as a valid, reliable, simple, easy-to-use, and easy-to-interpret method of estimating receptive vocabulary size. It is free to access, administer, and score, and can be completed in a relatively short amount of time (estimated at 20-40 minutes). It is also increasingly available in validated bilingual versions. However, few published studies have investigated the effect of the multiple-choice format, test strategies, guessing, or a DK option on the VST (Stewart, in press; Zhang, 2013).

In this paper I describe the methodology and preliminary findings of a case study that was aimed at exploring the effects of the addition of a DK option on the Vocabulary Size Test, specifically how one test taker qualitatively perceived and used the *I don't know* option.

The Vocabulary Size Test (VST)

The Vocabulary Size Test is a 100-item multiple choice test with four options. In its current version, it tests up to 20,000 word families (Nation, 2012b). The VST has several distinct advantages. First, it can be used to estimate vocabulary sizes of native speakers of English, as well as those of ESL and EFL learners. Second, the VST displays a high level of psychometric unidimensionality (Beglar, 2010); that is, it has been shown to test one, and only one, attribute (i.e., vocabulary size). In addition, several versions of the monolingual (English) test are considered equivalent (Nation, 2012b), which is an advantage for researchers. Work continues on new or equivalent bilingual VSTs (Karami, 2012; Nguyen & Nation, 2011). Finally, the VST can be freely accessed and downloaded for research and pedagogical purposes from the Victoria University at Wellington website (<http://www.victoria.ac.nz/lals/about/staff/paul->

nation#paul_nation_vocabulary_resources_downloads). The following sample item is from the VST:

1. see: They <saw it>.
 - a closed it tightly
 - b waited for it
 - c looked at it
 - d started it up

As an added advantage to ESL and EFL learners, the VST is available in bilingual formats. That is, the tested word and stem are written in the target language (English), but the options are written in the test takers' L1. As of 2013, it is available in Korean, Japanese, Mandarin, Russian, and Vietnamese (Nguyen & Nation, 2011; Victoria University of Wellington, n.d.). Other bilingual versions, such as Persian, continue to be developed and refined (Karami, 2012). The following sample item is from the 140-item bilingual Japanese version of the VST (Victoria University of Wellington, n.d.):

1. see: They **saw** it.
 - a 切った
 - b 始めた
 - c 待った
 - d 見た

As the VST continues to be researched and improved upon, it could become a very powerful tool for researchers, instructors, and learners alike to gauge their vocabulary learning progress and current vocabulary size, to identify future needs, and to function as a learning exercise in and of itself.

I Don't Know in Language Testing

Nation (2012b) disclosed that the VST “does not use an *I don't know option*, because such an option discourages informed guessing” (Writing the choices section, para. 9). On tests that do not employ penalties for incorrect guessing, informed guessing should be encouraged, and a DK option may not be appropriate for such contexts. However, in contexts that employ penalties for incorrect guessing, offering a DK option and using a DK option as a test-taking strategy is sound advice. From an administrative standpoint, high-stakes testing that offers a DK option—particularly when coupled with a penalty—could provide a better, more accurate, and more confident measure of test taker ability due to its discouragement of guessing (Zhang, 2013).

There are several points of consideration when regarding the VST, which also serve to provide the rationale for this study. First, there is currently no opt-out response on the VST. An unanswered item can only be coded as NR (no response), and the right-wrong or score calculations can be adjusted according to the test administrator's intended test strictness. For the test taker, however, there are unknown consequences for not responding to an item, as this condition is not covered in the test directions. For the test administrator, NRs give zero information about an item or a test taker, as the item could have been not responded to for a variety of reasons: accidentally missed, intentionally skipped, or intentionally skipped and then forgotten. Furthermore, these reasons could have occurred due to a number of variables, only some of which can be controlled by the test administrator: testing environment (e.g., temperature, noise level, or suitability), test taker condition (e.g., alertness, health, attention, or interest), or the test itself (e.g., difficulty, legibility, or length). For these reasons, providing an opt-out response and conducting qualitative interviews with test takers can provide critical information.

Second, guessing can cause a misleading increase in scores on the VST. For test takers, there is currently no penalty for guessing, which incentivizes guessing—or, at the very least, treats it as neutral. For test administrators, there is currently no score correction for guessing. Finally, there is no way to differentiate between random guessing, guessing that is informed by partial knowledge, and guessing that is prompted by test-taking strategy. Again, providing test directions that clearly state scoring procedures and offering an option that could potentially help eliminate random guessing is desirable. Furthermore, qualitative interviews can assist with differentiating between random guessing, knowledge-informed guessing, and test-taking strategy, and under what conditions each of these occur on the VST.

Case Studies in Language Testing

The intensive qualitative interview format of this preliminary case study was chosen for several reasons. First, prior research on the VST has investigated and considered mostly quantitative data on validity, scoring, or equivalency (Beglar, 2010). Second, the collection of both quantitative and qualitative data on the process of test taking, in addition to the VST results, has not yet been extensively undertaken and has been recently encouraged (Nation, 2012a; Nation, 2013, personal communication; Ngyuen & Nation, 2011). When describing a study that paired test takers with interviewers during and after the VST, Nation (2012a) wrote, “The reason for this very labour-intensive method of testing was to make sure that the results meant something” (p. 10).

Due to the individual and voluntary nature of the test administration, there is high motivation for the test taker to perform to the best of his or her ability. Due to the low stakes of the test, no negative effects or disadvantages are experienced beyond those usually associated with taking a multiple-choice test; that is, no decisions are made about or for the test taker based on the VST results. Finally, the intensive interviews allowed for otherwise

obscured qualitative data to be obtained. Employing these kinds of interviews in a classroom, peer-peer, or instructor–learner format is especially valuable to discover what learners know (or think they know) and don't know, why they know or don't know something, how they perceive a multiple-choice test and its options, as well as their test-taking, learning, and reasoning processes.

Methodology

Instruments

For the purposes of this study, two equivalent monolingual paper versions of the Vocabulary Size Test were used. One VST was utilized in its original form (stem with four options), but the other version was modified to include a fifth option in addition to the original four options (stem, four options, and *I don't know*). The following sample items are from the unmodified and modified VSTs, respectively:

Unmodified VST with With Four Options

1. drive: He <drives> fast.

- a swims
- b learns
- c throws balls
- d uses a car

Modified VST with With Five Options

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

For the purposes of this study, directions were written for both versions of the VST.

Unmodified VST Directions

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.

Modified VST Directions

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.

Procedure

For the pilot study, potential participants were identified in order to explore the possibility of a ceiling effect on the VST. As the VST can be used with L1 or L2 English users, an American L1 English user with a graduate-level education from a prestigious American university and a high English proficiency level (as determined by field of study, work experience, chosen industry, and extracurricular interests) was identified as having the highest chance of illustrating a potential ceiling effect. In this case study, a well-educated (MA Education in progress), L1 English user from the United States with TESOL experience who worked as an English literature and writing instructor in the Kanto area was chosen (participant A). Outside of her studies and career, the participant was also regularly involved in reading- and writing-centered activities.

The participant was met and briefed about the study. Both the original and the modified VSTs were taken by the participant. Each test-taking period was approximately 10 minutes long. Each interview lasted approximately 15 to 25 minutes. A short

debriefing was conducted. The entire procedure, from start to finish, lasted no longer than 2 hours.

The interview combined elements of think-aloud protocol and a semi-structured qualitative interview that followed the vocabulary interview protocol established by Nagy, Herman, and Anderson (1985). The interview was retrospective, conducted immediately after the test was administered and scored, and was later transcribed and coded by the researcher. Questions mainly focused on the following points:

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

Results

Original VST

Participant A's score was 95 points out of a total possible 100 points: out of 100 attempts, 95 items were correct and 5 items were answered incorrectly (#41, #75, #82, #83, and #96). During the retrospective interview, she stated that she did not guess on any item, but did use two partial knowledge techniques on items #70 and #80 (see Figure 1). Correct answers were scored as one point and incorrect answers were given zero points. As per the VST scoring instructions (<http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf>), by multiplying the number of correct answers by 200 (95 x 200), her receptive vocabulary size estimate would be 19,000.

For #41, participant A argued that the answer wasn't "necessarily wrong, actually. Gimmick is a cheap trick to get attention. To me, gimmick has more of a trick feeling than just an 'attention-getting action or image'—but it's not really a 'clever

plan,' I guess." She showed both knowledge of the word, its use in different contexts, and the subtlety of the definition. Similarly, she displayed partial knowledge for #75: "Coven. I always think of a group of witches, and I didn't like any of the answers, so I went with a group of church women, even though I know that they're trying to get you with 'convent.' I thought it was specifically only for women, so that's why I circled that one." For this item, she points out that knowledge of the distractors may actually interfere with answering correctly. Items such as this one (i.e., where the correct answer and a distractor are very similar sounding) may need to be more closely examined for item bias.

However, for #82 and #83, she stated that she actually knew the definition of both words, but just made a mistake. Finally, for #96, she said that she thought she knew the correct definition, but was ultimately incorrect.

She responded that for #70 and #80, she used a word root (for "substantiate") and process of elimination: "I knew it wasn't the first one. I didn't think it was the second one. I had to have one of those tests before that measure your amount of breath. So it was either c or d. D sounds silly, so I went with c." In both cases, partial knowledge allowed for a successful guess.

41. gimmick: That's a good <gimmick>.
- a thing for standing on to work high above the ground
 - b small thing with pockets for holding money
 - c attention-getting action or image
 - d clever plan or trick
75. coven: She is the leader of a <coven>.
- a small singing group
 - b business that is owned by the workers
 - c secret society
 - d group of church women who follow a strict religious life

82. vitreous: These rocks are <vitreous>.
- a very heavy
 - b easy to break
 - c full of small holes
 - d like glass
83. cerise: Her dress was <cerise>.
- a a bright red colour
 - b made of a thin, soft material
 - c a pale blue-green colour
 - d made of expensive fabric with pretty patterns and small holes
96. maladroit: He is <maladroit>.
- a feeling sick to his stomach
 - b physically awkward
 - c rather silly but likeable
 - d quickly angry and easily depressed
- *****
70. instantiate: you need to <instantiate that>.
- a make that happen quickly
 - b put that into the correct place
 - c give a real example of that
 - d explain that
80. pantograph: The <pantograph> is broken.
- a instrument which plays music from a metal tube
 - b instrument which measures the amount of breath a person has
 - c framework of moving bars for copying plans
 - d pen with a metal point for writing on hard surfaces

Figure 1. Original VST. Participant's incorrect items: 41, 75, 82, 83, and 96; correct partial knowledge items: 70 and 80.

Modified VST

Participant A's results were scored at 95 points out of a total possible 100 points: out of 100 attempts, 95 items were correct, 4 items were answered with *I don't know* (#74, #82, #86, #98), and 1 item was answered incorrectly (#92). During the retrospective interview, she admitted to guessing (correctly) on items #52 and #75 (see Figure 2). Correct answers were scored as one point and incorrect answers and *I don't know* were attributed zero points. Her receptive vocabulary size estimate was 19,000.

For the two items that participant A did not know, but guessed correctly, she used a combination of previous knowledge and elimination: (#52) "I had a word association in my head, and I don't think they have a glass wall for growing plants . . . I probably should have circled *I don't know* for that"; (#75) "This was a word I had heard before, and I thought I could narrow it down."

Notably, participant A said this about #74, the first item on which she chose the DK option: "At that point, I didn't really stop to think what the answer might have been due to the directions." This answer may point to an inhibiting factor of a DK option, which may suppress partial knowledge and artificially depress the overall score and estimate of vocabulary size. However, she also went on to say, "It's a word I've never heard before. [The answers are] all really similar." In this case, partial knowledge did not exist for participant A, either within the word itself or the definition. In such cases, the DK option is clearly the best choice and preserves a more accurate measure of vocabulary size.

For the remaining items (#82, #86, and #98) on which she chose DK, she professed to not know the meaning at all (#82), to having heard the word before but not knowing the meaning (#86), and to never having heard the word before (#98).

92. panzer: They saw the <panzers> getting nearer.
- a players in a marching band
 - b fighter planes
 - c large, slow windowless army cars
 - d policewomen
 - e I don't know
74. swingeing: They got <swingeing fines>.
- a very large fines
 - b very small fines
 - c fines paid in small amounts at a time
 - d fines that vary depending on income
 - e I don't know
82. gobbet: The cat left a <gobbet> behind.
- a strip of torn material
 - b footprint
 - c piece of solid waste from the body
 - d lump of food returned from the stomach
 - e I don't know
86. copra: They supply <copra>.
- a a highly poisonous substance used to kill unwanted plants
 - b the dried meat from a large nut used to make oil
 - c an illegal substance which makes people feel good for a short time
 - d strong rope used on sailing ships
 - e I don't know
98. casuist: Don't <play the casuist> with me!
- a focus only on self-pleasure
 - b act like a tough guy
 - c make judgments about my conduct of duty
 - d be stupid
 - e I don't know

52. refectory: We met in the <refectory>.
- a room for eating
 - b office where legal papers can be signed
 - c room for several people to sleep in
 - d room with glass walls for growing plants
 - e I don't know
75. cenotaph: We met at the <cenotaph>.
- a large and important church
 - b public square in the centre of a town
 - c memorial for people buried somewhere else
 - d underground train station
 - e I don't know

Figure 2. Modified VST. Participant's incorrect item: 92; items answered with DK: 74, 82, 86, and 98; correctly guessed items: 52 and 75.

Discussion

First, participant A chose her correct answers in consistent ways: She had either explicitly studied or heard the word previously. She provided example sentences, expanded on content or context, or cited specific situations when describing these answers.

Second, participant A utilized several consistent strategies for words that she had not explicitly studied or heard. Participant A used the following strategies: (a) partial (correct or incorrect) knowledge, (b) (correct or incorrect) process of elimination, and (c) (correct or incorrect) guessing.

Finally, participant A only used the DK option on the modified VST under extremely strict conditions. These conditions may be said to be self-imposed, as neither the test directions, nor the interviewer, stated or implied a reward or penalty for guessing or for using the DK option. On the Modified VST, participant A only used the DK option when "[she] did not

know the meaning at all," "[she had] heard the word before but [didn't] know the meaning," and when "[it was] a word [she'd] never heard before."

As this was a preliminary case study, the results cannot be generalized to other populations or contexts. However, the results are important as they provide insight into the test-taking strategies and effects observed when adding a DK option to the Vocabulary Size Test.

Conclusion

This study, as well as future studies in this area, will add more evidence to the growing body of literature around the VST and the effect of adding a DK option. Due to the VST's robustness when used in a bilingual format, other future directions might include the effects of guessing, and a DK option on these bilingual versions. Differences not apparent when using the monolingual VST as an instrument may emerge on bilingual VSTs or between categories of test takers (e.g., L1 English, L1 Japanese, and L1 other).

As a follow-up to this preliminary study, several studies are in progress. In order to test if the findings of this study are consistent across test takers with different levels of education and English language proficiency, additional sets of participants were identified and tested using the same methodology: highly educated (graduate degree or higher), educated (university degree), or degree in progress; proficient L2 English users with L1 Japanese, proficient L2 English users with L1 other, and intermediate L2 English users with L1 Japanese. Another study in 2014 will be conducted using a larger sample size of beginner and intermediate L2 English users with L1 Japanese. The participants will take the original VST and modified VST. Selected interviewees will undergo identical interview methodology to this study by a trained L1 Japanese interviewer.

In conclusion, attention should be paid to not only the VST itself—its format, stems, distractors, and development—but also to its administration and contextual use. Finally, further research should examine the effects of adding a DK option in testing and in L2 learning, and how a DK option in language learning compares to the use of a DK option in other research areas.

Bio Data

Dawn Lucovich is currently enrolled in Temple University's PhD Education program. She earned her MA from Teachers College Columbia University, and has served as Director and tutor for The Writing Center at Teachers College Columbia University's Tokyo campus. <lu covich@temple.edu>

References

- Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, 27(1), 101-118.
- Karami, H. (2012). The development and validation of a bilingual version of the Vocabulary Size Test. *Regional Language Centre Journal*, 43(1), 153-167.
- Laufer, B., & Nation, P. (1999). A vocabulary size test of controlled productive ability. *Language Testing*, 16(1), 33-51.
- Meara, P., & Jones, G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied Linguistics in Society* (pp. 80-87). London: CILT.
- Meara, P., & Jones, G. (1990). *Eurocentres vocabulary size tests 10KA*. Zurich: Eurocentres Learning Service.
- Nagy, W., Herman, P., & Anderson, R. (1985). Learning words from context. *Reading Research Quarterly*, 20(2), 233-253.
- Nation, I. S. P. (1993). Measuring readiness for simplified reading: A test of the first 1000 words of English. *Regional Language Centre Journal*, 31, 193-203.

- 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). 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.
- Nguyen, L. T. C., & Nation, P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *Regional Language Centre Journal*, 42(1), 86-99.
- Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, 10(3), 355-371.
- Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 41-60). Mahwah, NJ: Erlbaum.
- Schmitt, N., Ng, J. W. C., & Garras, J. (2011). The Word Associates Format: Validation evidence. *Language Testing*, 28(1), 105-126.
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18(1), 55-88.
- Stewart, J. (in press). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*.
- Victoria University of Wellington. (n.d.). Vocabulary Size Test (bilingual Japanese version). Retrieved from: http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocab_Size_Test_Japanese.pdf
- Zhang, X. (2013). The *I don't know* option in the Vocabulary Size Test. *TESOL Quarterly*, 47(4), 790-811. <http://dx.doi.org/10.1002/tesq.98>