Articles

A Japanese-English Bilingual Version of the New General Service List Test

Tim Stoeckel
University of Niigata Prefecture
Tomoko Ishii
Meiji Gakuin University
Phil Bennett
University of Niigata Prefecture

This paper describes the development and initial validation of a Japanese-English bilingual version of the New General Service List Test (NGSLT; Stoeckel & Bennett, 2015). The New General Service List (NGSL; Browne, 2013) consists of 2,800 high frequency words and is intended to provide maximal coverage of texts for learners of English. The NGSLT is a diagnostic instrument designed to identify gaps in knowledge of words on the NGSL. The NGSLT is a multiple-choice test that consists of 5 levels, each assessing knowledge of 20 randomly sampled words from a 560-word frequency-based level of the NGSL. A bilingual version of the NGSLT was developed to minimize the risk of conflating vocabulary knowledge with understanding of the answer choices. A validation study with 382 Japanese high school and university learners found the instrument to be reliable (α = .97) and unidimensional and to demonstrate good fit to the Rasch model.

本論文では New General Service List (NGSL) に基づく語彙サイズテスト(NGSLT)の日本語版の開発及び検証を論じる。NGSL (Browne, 2013) は高いテキストカバー率を目指して編集された2800語の高頻度語彙のリストであり、NGSLT (Stoeckel & Bennett, 2015) はそ

のリストについての学習者の知識を診断するテストである。NGSLを560語ごとの５レベルに分割し、各レベルから20語を無作為に抽出し計100問の多肢選択式のテストを作成した。選択肢の理解不足によって不正解になる懸念があるため、日本語版を作成した。大学生・高校生合わせて382人の学習者による検証により、この日本語版の信頼性が高いこと（α = .97）、測定が一次元的に行われていること、またラッシュモデルに適合することが確認された。

Keywords: bilingual tests; New General Service List; New General Service List Test; Rasch model; second language vocabulary testing

Vocabulary is now widely regarded as a critical component of L2 learning (Hunt & Beglar, 2005) with research revealing a close relationship between lexical knowledge and the skills of reading, writing, listening, and speaking (Milton, Wade, & Hopkins, 2010). In the case of reading, learners need to have knowledge of the basic form–meaning relationship of approximately 98% of the running words in a written text to facilitate unassisted comprehension (Schmitt, Jiang, & Grabe, 2011). Because most of this coverage is provided by high-frequency vocabulary, which is typically defined as the most frequent 2,000-3,000 word families (Nation, 2013; Schmitt & Schmitt, 2012), it is important for both pedagogy and research that reliable instruments to measure knowledge of such words be developed. This paper introduces one such instrument. We begin with overviews of two high-frequency vocabulary lists, the General Service List (GSL; West, 1953) and its modern replacement, the New General Service List (NGSL; Browne, 2013). We then outline the development and initial validation evidence of a Japanese–English bilingual version of a diagnostic test of high-frequency terms sampled from the NGSL.

West's Original General Service List

For many years, West's (1953) General Service List was used pedagogically to provide coverage of high-frequency words. The GSL consists of approximately 2,000 headwords plus their related constituents (e.g., *nation* plus *nations*, *national*, *nationally*, and *nationwide*). The criteria for inclusion of words in the GSL included both frequency and a subjective evaluation of how useful each word would be for L2 learners of English. Since its development, the GSL has been used in the creation of both learning materials, such as graded readers, and vocabulary assessment instruments, such as the Vocabulary Levels Test (Schmitt, Schmitt, & Clapham, 2001).

Despite its usefulness, the GSL was probably due for substantial revision. The approximately 2.5-million-word corpus used to create the list is small

by modern standards, meaning the GSL may not be representative of a wide range of language use. Additionally, the list itself has aged and is no longer completely reflective of modern lexis. For example, it contains items such as *telegraph* and *mankind*, which have gradually faded from use, but other terms that have become more widely used such as *computer* and *climate* are absent. Finally, the organization of the GSL may not be optimal for the types of learners (i.e., those of relatively low proficiency) most likely to benefit from a list of high-frequency vocabulary. Originally, entries were grouped as headwords plus inflected forms as well as many frequent and regular derivatives. This was later standardized by Bauman and Culligan (1995) so that each entry included all word forms through level 4 of Bauer and Nation's (1993) word family levels. Though word families appear to be actual psychological constructs (see, e.g., Nagy, Anderson, Schommer, Scott, & Stallman, 1989), and adult native speakers of English familiar with one member of a word family are also likely able to recognize other family members (Tyler & Nagy, 1989), this may not hold true for nonnative speakers unless they are highly proficient in the L2 (Gardner, 2007; McLean, Nation, Pinchbeck, Brown, & Kramer, 2016).

From the perspective of teaching and learning, this means that the burden for mastery of the GSL consists of learning not only each headword but also other family members that may not be readily recognizable from knowledge of the headword. However, for many word families, constituents differ greatly in the text coverage they provide. For instance, in the word family for *hard*, the constituents *hard*, *harder*, and *hardest* represent approximately 95.5% of occurrences in the Corpus of Contemporary American English (COCA; http://corpus.byu.edu/coca/), while the remaining seven word forms (*harden*, *hardness*, *hardship*, plus inflected forms) provide little additional coverage. When the improvement in coverage is so limited, learning those word family members should be deprioritized in favor of learning other more frequently occurring word forms. From the perspective of assessment, if learning one member of a word family does not automatically result in the ability to recognize other family members for many learners, then when tests are configured to measure knowledge of a sampling of headwords, correct responses do not necessarily indicate knowledge of all related family members.

The New General Service List

To address these limitations, a New General Service List was developed by Browne, Culligan, and Phillips (Browne, 2013). The NGSL is derived from an

analysis of a much larger 273-million-word subset of the Cambridge English Corpus (CEC). The complete CEC is comprised of materials with both reported (71.0% of the corpus) and unreported (29.0%) publication dates. Of the former, 85.1% of the words are derived from sources dated 2000 or later (S. Grieves, Cambridge University Press, personal communication, September 23, 2016). This, combined with the fact that the subset of the CEC used to create the NGSL was carefully balanced to include both written and spoken discourse from nine separate subcorpora (Browne, 2013), suggests that the NGSL is representative of a more modern and broader range of English language use than the GSL. Additionally, entries in the NGSL are organized into "modified lemmas" rather than word families. A regular lemma consists of a headword plus its inflected (but not derived) forms. Under the headword *hard*, for example, are only the inflected forms *harder* and *hardest*. In a "modified lemma," orthographically identical headwords and their constituent derivations are grouped together. For instance, the modified lemma for *approach* consists of the nominal inflection *approaches* plus the verbal inflections *approaches*, *approaching*, and *approached*. This modified lemma grouping facilitates accurate text analysis with tools such as the Lextutor VocabProfilers (http://www.lextutor.ca/vp/), which are currently incapable of distinguishing between orthographically identical word forms that belong to separate lemma groupings. It also aligns relatively well with Gardner's (2007) recommendations that words be grouped as base forms plus regular inflections for low-proficiency learners and extended to include irregular inflections and derivational prefixes for those at an intermediate level.

In total, there are 2,800 modified lemmas in the NGSL, ordered according to frequency and dispersion across the various subcorpora used to create the list (Browne, 2014). Though this figure exceeds the 2,000 word families in the original GSL, it may represent a smaller learning burden than the approximately 3,600 lemmas present in the GSL (Browne, 2013). In terms of coverage, at 90.34%, the NGSL offers about 6% more coverage of the subset of the CEC used to create the list than does the original GSL (84.24%; Browne, 2013).

The English Version of the New General Service List Test

The New General Service List Test (NGSLT) is a diagnostic instrument designed to identify gaps in learners' written receptive knowledge of words on the NGSL, to assist in setting vocabulary learning goals, and to aid in designing lexically appropriate educational experiences (Stoeckel & Bennett,

2015). This section describes the test in its original monolingual English format.

The test consists of 100 items, 20 for each of five 560-word frequency-based levels of the NGSL. Items are written with specifications similar to those of the Vocabulary Size Test (VST; Nation & Beglar, 2007). Thus, a multiple-choice format is used in which item stems consist of a target word followed by a short sentence using the word. These decontextualized sentences are intended to indicate the tested word's part of speech, help examinees view it as an authentic element of language, and provide "a little extra associational help in accessing the meaning" (Nation & Beglar, 2007, p. 11). Four answer choices follow in the form of short definitions or synonyms of the tested word and of three other words of similar frequency. To keep answer choices as simple and intelligible as possible, they are written with high-frequency vocabulary. Specifically, whenever possible, items testing words in the first three levels of the NGSL were written only with words from the first two levels. Moreover, items testing words in the fourth and fifth levels were written exclusively with words of higher frequency than the target word. Because of these restrictions, the correct answer is worded only specifically enough to distinguish it from the three distractors. For instance, for the item testing the word *slide*, the correct answer *move* defines *slide* in terms that are only precise enough to distinguish it from the distractors *break*, *make power*, and *become bigger* (for details see Stoeckel & Bennett, 2015). This approach is unavoidable for items testing words that cannot succinctly be defined with high-frequency vocabulary, but it means that examinees are sometimes not required to demonstrate precise meaning recognition. (As described below, the bilingual test format overcomes this limitation by using a direct translation of most tested words.)

The NGSLT is designed to determine whether examinees have made an initial link between the form and meaning of each tested word. To increase the accuracy of the test in this regard, three steps were taken during item writing, each of which was informed by frequency counts or a tally of concordance lines in the COCA. First, when the modified lemma included more than one part of speech, the most frequently occurring part of speech was the form that was tested. Thus, for the headword *approach*, the noun form was tested because it occurs more frequently than the verb form (shown in Figure 1). Second, for the example sentence in most item stems, the most frequently occurring word form for the tested part of speech was used. For *approach*, this meant the singular *approach* rather than the plural *approaches* was utilized. Third, when the tested word had multiple meanings, the most frequently oc-

curring sense was used to define the word. For *approach*, the definition was "way of doing something" rather than "movement toward something" or other less common meanings. This use of frequency as a guide in item writing was intended to increase the likelihood that examinees would be familiar with the word form and meaning sense used in the test.

approach: We like your **approach**.
a. way of doing something
b. part of a book
c. house and land
d. facts and information

Figure 1. Example item from the monolingual version of the NGSLT.

Japanese–English Bilingual Version of the NGSLT

For tests of written receptive vocabulary knowledge such as the NGSLT, bilingual formats have become popular because they are thought to reduce the risk of scores being influenced by poor knowledge of the syntax or vocabulary used in the answer choices (Elgort, 2013; Karami, 2012; Nguyen & Nation, 2011). Supporting this view, research with the VST has shown that scores are generally higher with bilingual versions than with the monolingual variant (Elgort, 2013; McDonald, 2015). Because the NGSLT is intended for low- and intermediate-level learners who may have limited grammatical knowledge or gaps in knowledge of even high-frequency vocabulary, a bilingual format may be particularly suitable for enabling examinees to fully demonstrate their actual lexical knowledge. In terms of washback, the bilingual format may be beneficial in encouraging learners to utilize the L1 to establish initial form–meaning linkage, an approach that is often more effective than using L2 definitions (see Schmitt, 2008).

In developing the Japanese–English bilingual version of the NGSLT, Nguyen and Nation's (2011) guidelines for writing such tests were adopted. Thus, answer choices are usually not direct translations of the definitions that are used in the English version but are instead translations of the words that are defined by the answer choices in the English version. For instance, in the monolingual item shown in Figure 2, "hit this hard" defines *knock*, "follow this" defines *pursue*, and "exchange this for something" defines *switch;* these three words and the target word *justify* are translated directly into Japanese in the bilingual version.

justify: We cannot **justify** this. a. hit this hard b. follow this c. exchange this for something else d. show that this is right	**justify**: We cannot **justify** this. a. 打ち砕く b. 追求する c. 変更する d. 正当化する

Figure 2. A comparison of monolingual and Japanese-English bilingual formatting in the NGSLT.

In addition, four conventions employed by McLean, Ishii, Stoeckel, Bennett, and Matsumoto (2016) in their revisions to the Japanese–English version of the VST were used here. First, when the answer choices in the monolingual version defined words that exist as loanwords in Japanese, the *katakana* forms of these words were avoided in the bilingual format to prevent the use of phonological matching to guess the correct answer or to eliminate distractors. Instead, paraphrases or alternative words with Japanese etymological origins were used. For instance, for the tested word *hall*, rather than using the *katakana* ホール (*hooru*), which is phonologically similar to the tested word, the alternative 集会場 (*shukaijo*) was used. Second, consistency in part of speech and inflection was pursued across the four answer choices. For instance, though Japanese adjectival forms have several possible endings (e.g., ～い [*-i*], ～な [*-na*], ～である[*-dearu*]), an effort was made to use the same ending across the four options wherever it was possible and sounded natural. Third, the answer choices were written so that none would stand out from the others due to a difference in length. Finally, the wording of each item stem was checked to make sure that it sounded natural together with each of the four answer choices from the point of view of a native speaker of Japanese. For this purpose, passive verbs in some item stems were changed to the active voice, which is less awkward in Japanese. For instance, the item stem "**hide:** It was **hidden**" in the monolingual version was changed to "**hide:** Please **hide** these" in the bilingual variant, so that the base form 隠す (*kakusu,* meaning *hide*) could be used instead of its inflected passive form 隠された (*kakusareta,* meaning *hidden*).

Initial Validation Evidence

An initial validation study of the Japanese–English bilingual version of the NGSLT was conducted with a convenience sample of 386 native speakers of

Japanese. These included 285 1st-year students at a prefectural university in northern Japan and 101 students in their 1st, 2nd, or 3rd year of study from three high schools in southern Japan.

The instrumentation consisted of Form A of the Japanese–English bilingual variant of the New General Service List Test (available from the authors).[1] For the university participants, the test was administered by computer via the ClassMarker website (https://www.classmarker.com), which prevented examinees from skipping test questions.[2] For the high school participants, the test was administered in a paper and pencil format because computers were unavailable. These examinees were instructed to answer all test questions; for items testing unknown words, the instructions were to make a best guess after carefully reading all answer choices. Four of these participants responded to fewer than 70% of test items and were therefore removed from the study (final n = 382). For the remaining high school participants, there were a total of 10 unanswered questions (by nine separate persons). The data were analyzed with these values missing and again with imputed randomly generated answers (see Garson, 2012), and there were no discernible differences in any of the analyses. Reported here are the results with imputed answers.

Estimates of reliability were satisfactory for the entire sample (α = .97) and for the separate groups of university (α = .80) and high school (α = .89) students (Table 1). The reliability coefficients for the five individual test levels ranged from .82 to .88 (Table 2), suggesting that this version of the NGSLT also provides reliable estimates of vocabulary knowledge at each level of the test. Table 2 also shows that mean scores across the five levels of the test were consistent with the frequency-based model of vocabulary acquisition in which there is a general trend for higher frequency words to be learned before less commonly occurring lexis (Milton, 2009). Scores of the university students (n = 285) were also found to correlate moderately with performance on the Computerized Assessment System for English Communication (CASEC; http://global.casec.com), a test of general English proficiency, r = .586, p < .001.

Table 1. Descriptive Statistics

Group	*n*	*M*	*SD*	α
All	382	80.1	18.1	.97
University	285	89.4	6.0	.80
High School	97	52.6	12.8	.89

Note. There were 100 items on the test.

Table 2. Descriptive Statistics by Test Level

Test Level	*M*	*SD*	α
1	17.0	3.5	.87
2	17.0	3.8	.88
3	15.7	3.9	.86
4	15.2	3.7	.82
5	15.2	4.3	.86

Note. There were 20 items in each level of the test.

Rasch analysis was performed with the Winsteps software (version 3.92) to assess person fit, item fit, construct dimensionality, and the responsiveness of the instrument to changes in the measured construct. Using standardized outfit values > 2.0 as the criterion for person and item misfit (Wolfe & Smith, 2007), 27 persons were flagged as misfitting the model and were temporarily removed. With this smaller dataset (n = 355), 20 items were identified as having poor fit to the Rasch model. To investigate reasons for this high number of misfitting items, the 27 persons were reinstated and a principle component analysis (PCA) of Rasch residuals was conducted to inspect person and item dimensionality. Rasch analysis identifies the primary dimension in a dataset, presumably lexical knowledge in the case of the NGSLT, and a nonrandom pattern in the residuals would be indicative of a secondary dimension (Linacre, 2007). Using Stevens' (2002) criteria, meaningful dimensions in the PCA were defined as those with components having 10 or more loadings above .40, four or more above .60, or at least three above .80. No secondary dimension was identified in the item residuals. However, in the PCA of person residuals 11 persons (all university students) had loadings above .40 and 16 (all high school students) below -.40 in the first contrast.

This suggested the possibility of differential item functioning (DIF), a situation in which different groups of examinees respond differently to one or more test items even after differences in ability are accounted for (Zumbo, 1999). Thus, a separate calibration *t*-test approach was employed to explore DIF between the high school and university groups. Based on Linacre (1994), the criteria for DIF was set at $p < .01$ and effect size > 1.0 (defined as the group difference in Rasch item measures). Thirty items were found to demonstrate DIF. Fourteen were more difficult than expected for university students and 16 for high school students. Items displaying DIF were examined to determine whether they were sources of unfair bias or were instead indicative of real differences in knowledge between the groups (Zieky, 2006). Consistent with the general pattern of vocabulary growth from high school to university, all test items, including those exhibiting DIF, were more difficult for the high school group in *absolute* terms. For example, although Q74 *currency* exhibited DIF in favor of the high school group ($p < .001$, DIF contrast = -2.79), the university students still performed slightly better in absolute terms (38.2% versus 36.1% correct). Moreover, an examination of each DIF item revealed no obvious cause of unfair bias between the two groups. From this we tentatively conclude that DIF with the present sample was caused by actual dissimilarities in lexical knowledge, perhaps due to differences in curriculum or language exposure.

To remove the effect of DIF in assessing item fit, a separate analysis was conducted for each group. For the high school group, six items misfit the Rasch model (standardized outfit values in parentheses): Q21 *observe* (2.7), Q25 *extra* (2.7), Q27 *solution* (2.1), Q49 *guarantee* (3.6), Q67 *impose* (2.6), Q 99 *accurate* (2.4); for the university group, there were four misfitting items: Q1 *charge* (4.8), Q26 *instance* (2.5), Q52 *label* (2.9), and Q79 *shadow* (2.2). These items will be monitored in future test administrations; however, considering that for each group approximately five items should exceed a standardized outfit value of 2.0 by chance alone, these findings suggest acceptable fit to the Rasch model when the effect of DIF is removed.

Instrument responsiveness, the capacity of an instrument to detect changes in a measured construct (Wolfe & Smith, 2007), was assessed visually with the person–item map shown in Figure 3. Persons are arranged on the left according to ability, and items are arranged on the right according to difficulty. The numbers along the left margin represent the logit-based scale for both person and item measures. In the present dataset, item measures from approximately -3 to +3 logits indicate that the instrument provided coverage over a range of person abilities and is sensitive to changes in the

measured construct. The presence of test items well below the lowest ability person indicates that the test did not have a floor; however, the presence of a number of learners above the most difficult item suggests a ceiling effect as learners gain mastery of the NGSL.

Taken together, the findings of this initial investigation indicate that the instrument appears to have sound measurement properties and behaves in ways that are expected by theoretical understanding of the tested construct. However, the presence of DIF between high school and university learners warrants further investigation.

Score Interpretation and Test Use

The NGSLT assesses written receptive vocabulary knowledge, the kind of lexical knowledge needed for reading (Stoeckel & Bennett, 2015). As stated above, it assesses knowledge of form–meaning linkage. It does not evaluate aspects of vocabulary depth such as collocation or register. Additionally, it does not measure lexical comprehension in listening, nor should it be used to assess productive vocabulary in speaking or writing. Moreover, score interpretations need to account for the overestimation of lexical knowledge due to the use of test-taking strategies and blind guessing that has been observed to occur in multiple-choice vocabulary tests (Gyllstad, Vilkaitė, & Schmitt, 2015). For bilingual multiple-choice tests of written receptive vocabulary knowledge, recent research found that this overestimation is equivalent to approximately 40 to 45% of unknown words (Stoeckel, 2016; Stoeckel & Stewart, 2016).

With this in mind, a good way to use the test is to examine the pattern of scores across the five test levels and to use the point at which scores drop and stay below a threshold of 85 to 90% (i.e., 17 or 18 correct out of 20 at a given level) as a target for intentional vocabulary study. This threshold, and not 100%, is based in part on work by Milton (2009) indicating that even high proficiency learners commonly have small gaps in knowledge of high-frequency vocabulary. The threshold also allows for a small amount of miskeying on the part of examinees. Factoring in the use of test strategies and random guessing, this benchmark represents a level of mastery of perhaps 75% of the entries in a 560-word band, meaning a gap in knowledge of at least 140 words. Because of the importance of high-frequency vocabulary, scores below this threshold indicate a need to study and learn the unknown words at the level in question.

Rasch Measure (logits)	Persons (more ability)	Items (more difficult)
6		
	UUU	
5		
	UUUUUU	
	UUUUUUUUUUUUUU	
4	UUUUUUUUUUUUUUU	
	UUUUUUUUUUUUUUUUUUUU	
	UUUUUUUUUUUUUUUUUUUUUUUUUU	
	UUUUUUUUUUUUUUUUUUUUUUUU	
	UUUUUUUUUUUUUUUUUUUUUUUUUUUU	Q1 Q74
3	UUUUUUUUUUUUUUUUUUUUUUUUUU	Q67
	UUUUUUUUUUUUUUUUUUUUUUUUUUU	
	UUUUUUUUUUUUUUUUUUUUU	Q60 Q68 Q93
	UUUUUUUUUUUUUUUUUUUUUUUU	Q43 Q58
	UUUUUUUUUUUUUUUUUUUUUUUUUUU	Q70 Q76
2	UUUUUUUU	Q10 Q52 Q89 Q94
	UUUUUUUU	
	HHHHUUUU	Q91
	HHUUUU	Q35
	HHHHHHUUUU	Q24 Q47 Q71 Q80 Q84
1	HHHHHH	Q18 Q26 Q49 Q50 Q61 Q85 Q87 Q99
	HHHHHHHHH	Q42 Q90
	HHHHHHHHH	Q13 Q33 Q92 Q97
	HHHHH	Q21 Q31 Q54 Q62 Q63 Q98
	HHHHHHHHH	Q10 Q11 Q19 Q3 Q40 Q5 Q7
0	HHHHHHHHHHH	Q22 Q25 Q27 Q2 Q37 Q38 Q46 Q57 Q66 Q96
	HHHHHHHH	Q17 Q30 Q56 Q77 Q78 Q81
	HHHHHHH	Q45 Q59 Q69
	HHHHH	Q48 Q72 Q83
	HHH	Q23 Q29
-1	HHH	Q6 Q88 Q95
	HHHH	Q20 Q53
	HH	Q55 Q65 Q8
	H	Q12 Q15 Q16 Q73
		Q39 Q51 Q79 Q82 Q9
-2		Q32 Q34
		Q28 Q36
		Q41 Q44 Q64
		Q75
-3		Q86
		Q4
-4		Q14
	(less ability)	(less difficult)

Figure 3. A Rasch person–item map showing the person ability and item difficulty measures for the high school and university students in the validation study. The persons are designated "U" for university and "H" for high school participants. The numbers along the left-hand side are the Rasch-based logit values for persons and items.

Additional Resources

Once a level of the NGSL has been identified for intentional study, learners may benefit from several additional steps. First, it would be useful for examinees to review a complete list of headwords from the targeted NGSL level and to highlight unfamiliar words. This can serve as an initial step in awareness raising and can help to verify whether the target level matches learner needs. Second, it would be valuable for learners to begin a principled program of study of the targeted words (Hunt & Beglar, 2005; Schmitt, 2008). For this, there are a number of free, well-designed resources available on the NGSL website (http://www.newgeneralservicelist.org/), including several different English-only and Japanese–English bilingual spaced repetition flashcard applications that can be used on both PCs and smartphones. Third, periodic reassessment would be helpful to monitor lexical growth, to see whether learning goals have been met, and to guide further goal-setting.

To this end, Form B of the Japanese–English bilingual NGSLT has been completed, and Forms C and D are under development. These instruments use the same test blueprint but each assesses knowledge of 100 different randomly sampled NGSL words. A comparison of the relative difficulty of Forms A and B has been conducted with a sample of Japanese learners using a Rasch-based anchor item approach (see Wolfe, 2000). This examination suggests that Forms A and B yield similar but not identical scores for learners of the same ability. The tests could therefore be considered equivalent for low stakes diagnostic and classroom use but need more careful scrutiny and perhaps a reassignment of some items across the two forms before they could be considered equivalent for high stakes purposes.

Conclusion

This paper introduced and described the ongoing development of a Japanese–English bilingual version of the New General Service List Test. The initial validation evidence presented here suggests that the instrument provides a psychometrically sound measure of written receptive knowledge of words on the NGSL. Test results can be used to establish learning goals, to monitor progress in achieving those goals, and to ascertain whether educational materials are lexically suitable for a given group of learners. Both the monolingual and bilingual versions of Forms A and B of the NGSLT are freely available from any of the authors' Academia.edu profile pages.

Notes

1. The test has been periodically revised to reflect both updates to the NGSL and item-performance data collected from ongoing testing. The version used in the present study is dated March 2016.
2. Research on instructions to skip unknown words or the addition of "I don't know" as an answer choice in multiple-choice vocabulary tests shows that examinees use such conventions differentially (Bennett & Stoeckel, 2012; Stoeckel & Stewart, 2016), which introduces "willingness to skip" as a nonrelevant construct impacting test scores and weakening test validity (Stoeckel, Bennett, & McLean, 2016).

Tim Stoeckel is an associate professor at the University of Niigata Prefecture. His primary interests are in L2 reading and vocabulary development and testing.

Tomoko Ishii is a lecturer at Meiji Gakuin University. Her primary research interests are in memory studies and vocabulary development.

Phil Bennett is an associate professor at the University of Niigata Prefecture. His research interests are mainly in vocabulary development, testing, and metaphorical vocabulary acquisition.

References

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

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

Bennett, P., & Stoeckel, T. (2012). Variations in format and willingness to skip items in a multiple-choice vocabulary test. *Vocabulary Education & Research Bulletin, 1*(2), 2-3. Retrieved from https://jaltvocab.weebly.com/uploads/3/3/4/0/3340830/verb-vol1.2.pdf

Browne, C. (2013). The New General Service List: Celebrating 60 years of vocabulary learning. *The Language Teacher, 37*(4), 13-16. Retrieved from https://jalt-publications.org/files/pdf-article/37.4tlt_featureds.pdf

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction, 3*(2), 1-10. Retrieved from http://www.charlie-browne.com/wp-content/downloadable-files/vli130026.pdf

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

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

Garson, D. (2012). *Missing values analysis and data imputation.* Asheboro, NC: Statistical Publishing.

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL - International Journal for Applied Linguistics, 166,* 278-306. https://doi.org/10.1075/itl.166.2.04gyl

Hunt, A., & Beglar, D. (2005). A framework for developing EFL reading vocabulary. *Reading in a Foreign Language, 17,* 23-59. Retrieved from http://nflrc.hawaii.edu/rfl/April2005/hunt/hunt.pdf

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

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7,* 328.

Linacre, J. M. (2007). *Dimensionality: Contrasts and variances*. Retrieved from http://www.winsteps.com/winman/principalcomponents.htm

McDonald, K. (2015). The potential impact of guessing on monolingual and bilingual versions of the Vocabulary Size Test. *Osaka JALT Journal, 2,* 44-61. Retrieved from http://www.osakajalt.org/storage/Osaka_JALT_Journal_2015.pdf

McLean, S., Ishii, T., Stoeckel, T., Bennett, P., & Matsumoto, Y. (2016). An edited version of the first eight 1,000-word frequency bands of the Japanese-English version of the Vocabulary Size Test. *The Language Teacher, 40*(4), 3-7. Retrieved from https://jalt-publications.org/files/pdf-article/40.4-tlt-art1.pdf

McLean, S., Nation, P., Pinchbeck, G. G., Brown, D., & Kramer, B. (2016). Revisiting the word family: What is an appropriate lexical unit for Japanese EFL learners? Paper presented at the Vocab@Tokyo Vocabulary Conference, Tokyo, Japan. Abstract retrieved from https://drive.google.com/file/d/0B6pqkaroKu33OWZFR3RJX21ZSVU/view

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

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

Nagy, W., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24,* 262-282. https://doi.org/10.2307/747770

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge, England: Cambridge University Press.

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

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

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

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95,* 26-43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47,* 484-503. https://doi.org/10.1017/S0261444812000018

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. https://doi.org/10.1177/026553220101800103

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Erlbaum.

Stoeckel, T. (2016, September). A serial multiple-choice format to reduce overestimation of meaning recall knowledge on the Vocabulary Size Test. Paper presented at the 20th Anniversary Conference of the Japan Language Testing Association, Kanagawa, Japan.

Stoeckel, T., & Bennett, P. (2015). A test of the New General Service List. *Vocabulary Learning and Instruction, 4*(1), 1-8. Retrieved from http://vli-journal.org/wp/wp-content/uploads/2015/10/vli.v04.1.2187-2759.pdf

Stoeckel, T., Bennett, P., & McLean, S. (2016). Is "I don't know" a viable answer choice on the Vocabulary Size Test? *TESOL Quarterly, 50,* 965-975. https://doi.org/10.1002/tesq.325

Stoeckel, T., & Stewart, J. (2016, September). The "I don't know" option and L1 answer choices: A comparison of four variants of the Vocabulary Size Test. Paper presented at the Vocab@Tokyo Vocabulary Conference, Tokyo, Japan. Abstract retrieved from https://drive.google.com/file/d/0B6pqkaroKu33OWZFR3RJX21ZSVU/view

Tyler, A., & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language, 28,* 649-667. https://doi.org/10.1016/0749-596X(89)90002-8

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

Wolfe, E. W. (2000). Understanding Rasch measurement: Equating and item banking with the Rasch model. *Journal of Applied Measurement, 1,* 409-434.

Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II—validation activities. *Journal of Applied Measurement, 8,* 204-234.

Zieky, M. J. (2006). Fairness review in assessment. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359-376). Mahwah, NJ: Erlbaum.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF).* Retrieved from http://faculty.educ.ubc.ca/zumbo/DIF/