A True/False Translation Test for English Vocabulary Size Assessment in a Japanese Context

Eoin Jordan

Abstract

This article describes the development and initial trial of a translation-based True/False Vocabulary Size Test (TFVST) for Japanese learners of English, created with the aim of improving on established vocabulary size tests in terms of: (i) greater time efficiency, and (ii) utilization of L1 translations, rather than L2 definitions, for more efficient confirmation of form-meaning link knowledge. A group of Japanese university students (n = 58) and Japanese teachers of English (n = 2) took the new test together with a corresponding written translation task, and the results provided some preliminary evidence for TFVST's reliability, validity, and practicality. Most notably, TFVST results correlated strongly with those of the translation task criterion measure (r_s = .791, p < .001), and the new test also appeared to be more time efficient than existing vocabulary size tests. The findings of this research suggest that TFVST may be a useful alternative for measuring English vocabulary size in a Japanese context, and accordingly that translation-based True/False vocabulary size tests of this type are worthy of further investigation with both Japanese and other L1 groups.

Key terms: English vocabulary size, vocabulary assessment, true/false, translation, Japanese

Vocabulary knowledge is widely regarded as a central component of second language proficiency (Albrechtsen, Haastrup, and Henriksen, 2008; Alderson, 2005; Laufer and Goldstein, 2004; Meara, 1996a; Nation, 1990), and accordingly its accurate measurement is of great importance within the field of applied linguistics, both for research and educational purposes. Indeed, Beglar (2010) notes that tests measuring vocabulary 'size' or 'breadth' (Anderson and Freebody, 1981) may be used to chart vocabulary growth, compare the vocabulary sizes of different groups, assess whether courses are their meeting lexical targets, and determine how successful learners might be at different everyday tasks such as talking with friends, watching movies or reading newspapers.

A number of currently existing tests provide information about learners' vocabulary sizes. In particular, a great deal of attention has been paid over the past two decades to the Vocabulary Levels Test (VLT) (Beglar and Hunt, 1999; Nation, 1990; Schmitt, Schmitt, and Clapham, 2001) and the Yes/No format (Eyckmans, 2004; Meara, 1992, 2005; Meara and Buxton, 1987; Meara and Jones, 1988, 1990). It should be noted that the VLT was not strictly speaking designed to measure breadth of vocabulary, but its results do provide a considerable amount of information about participants' vocabulary profile, and therefore size; thus it is loosely defined here as a vocabulary size test. In recent years other new tests of vocabulary size such as the Computer Adaptive Test of Size and Strength (CATSS) (Laufer, Elder, Hill, and Congdon, 2004; Laufer and Goldstein, 2004) and the Vocabulary Size Test (VST) (Beglar, 2010; Nation, 2008; Nation and Gu, 2007) have also been developed. While all of these tests are supported by some degree of validity evidence, one area in which this appears lacking is in the provision of convincing concurrent validity evidence. This type of evidence is produced through the comparison of participant test scores with some other criterion measure taken at approximately the same time by the same participants (Alderson, Clapham, and Wall, 1995). A strong correlation between the test scores and the criterion measure would indicate an overlap in the constructs being measured. The lack of such evidence is quite surprising, given how common this type of validation is in language education (Bachman, 1990).

The VLT has more validity evidence behind it than the other three tests discussed here (Beglar 1999; Read 1988; Schmitt et al. 2001). Much of this is highly compelling; however, the results presented for its concurrent validity appear less convincing. In what is arguably the most in-depth validation study, Schmitt et al. (2001) used post-test interviews to assess whether a selection of their participants knew certain words on the test. The interviewers' assessments of whether

these words were known or not were then used as a criterion measure against which the participants' responses to the relevant items on the VLT were compared, and a Phi coefficient of .749, $p < .001$ was recorded. This coefficient indicates a reasonably high degree of similarity between words participants were judged as knowing in the interviews and those they gave correct answers for on the VLT. Although this could be viewed as evidence for the concurrent validity of the test, it should be noted that Schmitt et al. conducted their interviews after administering the VLT. This meant it was possible that some participants could have correctly matched definitions with words during the test through guesswork, then recalled and repeated these definitions in the interview. Accordingly, the interview results may not give an accurate indication of participants' vocabulary knowledge before taking the VLT, which would surely be the most appropriate criterion measure.

In another study, Mochida and Harrington (2006) compared results of a Yes/No test against those from the VLT and found a high degree of correlation ($r > .8$, $p < .001$). However, their work presupposed that the VLT is suitable to be used as a criterion measure for other vocabulary size tests, a presumption that may be called into question given the VLT's own lack of convincing concurrent validity evidence. Other than this, the highest correlation coefficient that has been reported in a positive light with regard to any of these four formats is $r = .58$, $p < .001$ between the passive recognition and passive recall sections of CATSS (Laufer and Goldstein, 2004).

It is the author's opinion that the lack of strong concurrent validity evidence highlighted here is the result of some test design problems. Even though the existing tests do appear to function reasonably well in practice, there is scope to investigate new test formats that may provide superior combinations of accuracy and practicality in specific teaching and research contexts. Accordingly, this article will first outline some of the key problems with existing vocabulary size tests, and then describe a new format of test for Japanese learners of English that was developed in response to these issues. Finally, a small-scale pilot study of the new test will be reported.

Issues with Existing Vocabulary Size Tests

The vocabulary size tests mentioned in the previous section have several significant design problems. Firstly, Meara (1996b) suggests that the sampling rate of the original VLT may be too low to estimate vocabulary size reliably - an issue that may be partly a result of the VLT's relatively low time efficiency per item in comparison to Yes/No type tests. Indeed, the completion times reported by Schmitt et al. (2001) for the VLT are considerably longer than those reported by Eyckmans (2004) for the Yes/No format when calculated on a per item basis. Test times were not reported by Beglar (2010) for the VST or by Laufer and Goldstein (2004) for CATSS, but it seems reasonable to assume that neither of these designs are as time efficient as the Yes/No format given the relatively large amount of reading that they require for each item on the part of the test-taker.

Despite having a time efficiency advantage, it is not clear whether the Yes/No format really measures a consistent level of knowledge of the form-meaning link. Knowledge of the form-meaning link of a vocabulary item indicates the degree to which a person can associate the visual form of a word with aspects of its meanings. For instance, it is quite feasible that participants recognize the form of a word, but have misinformation about the words meaning. In this case, if they select the word as known then they will receive marks for it. The most confident answering strategy for a Yes/No test involves the test-taker interpreting the 'do you know this word?' instruction as 'is this a real word?', in which case the test simply becomes a True/False examination that only assesses knowledge of word form. On the other hand, some test-takers may be less confident in their own knowledge and therefore be less inclined to select words unless they are very familiar with them. This makes it very difficult to draw accurate comparisons between different students' scores as they may be more reflective of test-takers' personalities than their knowledge of vocabulary.

While the other test formats discussed here employ multiple-choice items that do require participants to demonstrate some knowledge of the form-meaning link, this type of item has the disadvantage of being susceptible to strategic guessing. Participants are more likely to choose distractors than the key option if they guess blindly; however,

they may employ distractor elimination strategies in order to improve their chances of guessing the right option (Schmitt et al., 2001). Accordingly these tests can be seen as assessing participants' test taking strategic competence in addition to their vocabulary knowledge (although the VST does attempt to deal with this by including distractors that are close in meaning to the key (Beglar, 2010)).

A further issue with the VLT and VST is that, despite employing a restricted vocabulary (Beglar, 2010; Nation, 1990), examinees may have trouble understanding some words or grammar in the definitions that are used as response options. If this is the case, then their ability to understand the definition is being tested as well as their knowledge of the target item. It should be noted, however, that versions of the VLT using translations into other languages for different L1 groups in place of definitions have also been developed and are available from Paul Nation's webpage (Nation, 2009a).

One other common failing of all of the above tests mentioned here is that none of them attempt to take account of words that are cognate for test-takers (that is, L2 words that share a common root with L1 words). This has been noted as an issue of importance with relation to the VLT by Nation (1990, p.262) and Read (2000, p.123), and with respect to the Yes/No format by Eyckmans (2004, p.84). Of course, considering this may only be practical when all participants are from the same L1 background; however, failure to do so may result in disproportionately large or small numbers of cognate words at different frequency levels being randomly sampled for tests, which may in turn affect results. In particular, the chance of this happening is higher when there is a low sampling rate, as is particularly the case with the VST, which uses only ten words to represent a frequency band of 1000.

New Test Development

In response to the problems highlighted in the previous section, the author developed a new test for Japanese learners of English, named the True False Vocabulary Size Test (TFVST), to improve on existing designs by:

- Minimizing reading to the greatest extent possible (thus allowing for the maximum possible number of items to be tested per unit of time) while still requiring participants to demonstrate some knowledge of the form-meaning link for vocabulary items.
- Utilizing translation into L1 to avoid problems with non-comprehension of L2 definitions.

The design of the test also attempted to ensure that the proportion of cognates sampled at each word frequency level was representative of the proportion in the frequency band itself; thus, all potential items were categorized as either cognate or noncognate. There is not sufficient space in this article to detail this procedure fully (see Jordan (2012) for a full description of the procedure); however, it should be noted that the final set of items used in TFVST could also have been selected using random sampling of each frequency band with no control for cognates. Accordingly, the stratified sampling of cognates did not enhance any of the results achieved in the trial of the new test – it simply acted to prevent an unrepresentative sample of cognates distorting them. The basic test item design for TFVST is shown in Figure 1.

	True	False
Dog = 犬		

Figure 1: Example TFVST item.

Rationale

Why True/False? The True/False format is used regularly in other areas of educational assessment (Burton, 2005; European Society of Anaesthesiology, 2008) and has the favourable property of being able to cover a large number of items in a short period. When compared with existing vocabulary size tests, the True/False format's time efficiency advantage is likely only to be superseded by the Yes/No format. Greater time efficiency allows for a larger, and therefore

more representative, sample of the words in a frequency band than a multiple-choice format would. With regard to strategic guessing too, True/False items provide fewer hints to facilitate distractor elimination than their multiple-choice counterparts do (for true items in particular); although blind guessing is likely to present a greater problem.

It was decided not to include a '*Don't know*' option in the test, even though there is disagreement in the literature on this point (Burton, 2001, 2002, 2005; Ebel, 1979; Muijtjens, Mameren, Hoogenboom et al., 1999). The main issue is whether the variance in the amount of guessing between individual participants when such an option is included has a greater effect on scores than the inevitable, yet more quantifiable, blind guessing that occurs when participants have to provide answers for every item. Calculations using the binomial distribution (Mendenhall, Beaver, and Beaver, 2008) indicate that pure chance guessing results in increasingly predictable overall scores as the number of events increases. This is similar to the situation where a coin flipped an infinite number of times will have proportions of heads and tails that both tend towards 50%. However, differences in answering strategy when a *Don't know* option is available are impossible to predict and their effect is likely to remain constant regardless of the length of the test. With this in mind, it seems clear that longer tests should require participants to provide answers for every item (that is, there should not be a *Don't know* option), although it is less obvious exactly how many items there should be in order for elimination of the option to be advisable. With these uncertainties in mind it was decided that adding a *Don't know* option was not appropriate for this test, particularly given the expectation that TFVST's time efficiency would allow for a high sampling rate.

Why translation? Although translation in language teaching has been highly unfashionable ever since the advent of the communicative approach (Swan, 1985a; Swan, 1985b), some recent research suggests that in fact the form-meaning link is learned no better through the context of a sentence than it is through L2 words paired with their L1 translations (Webb, 2007). Buck (1992) and Ito (2004) also found that translation may be an effective way of testing reading comprehension, and Eyckmans (2004, p.77), Laufer and Goldstein (2004) and Nation (2001, p.351) all favour the use of translation as the most thorough method by which to test receptive vocabulary knowledge.

Design

Word lists. Filtered versions of Paul Nation's (2009b) BNC word family lists were used as the basis for TFVST. These lists consist of word families (Bauer and Nation, 1993) determined according to Bauer and Nation's Level 6 criteria (Nation, personal communication). While lists of lemmas and individual words from the BNC are also available (Kilgarriff, 1996), word families are generally considered by researchers to be the most appropriate choice when measuring receptive vocabulary size (Schmitt, 2010). There is also empirical psycholinguistic evidence that supports the validity of the word family concept (Bertram, Baayen, and Schreuder, 2000; Bertram, Laine, and Virkkala, 2000; Nagy, Anderson, Schommer, Scott, and Stallman, 1989). Nation's lists were filtered using the JACET (Japan Association of College English Teachers) 8000 lists (Aizawa, Ishikawa, and Murata, 2005), which are based partly on the BNC, but are also reflective of English teaching materials in Japan, and claim to represent the most important 8000 English word families for Japanese students. This process was intended to eliminate any 'outlier' words that might be considerably more or less familiar to Japanese students than Nation's frequency banding suggested.

The filtering procedure involved entering each of Nation's original frequency banded lists of headwords into the JACET 8000 Level Marker (Shimizu, 2013). This showed in which thousand word JACET 8000 band each of Nation's headwords would be placed. The results are displayed in Table 1. There was a small group of words that placed very differently on the two lists; for instance, the word 'confer' was ranked in the 1k band of Nation's lists, but in the 5000 band of JACET 8000.

Table 1

Breakdown of Nation's BNC lists into JACET 8000 levels

JACET 8000 lists	Nation's BNC lists				
	1k	2k	3k	4k	5k
1000	745	187	15	12	7
2000	162	418	144	52	8
3000	31	172	270	140	60
4000	24	127	148	154	77
5000	8	42	160	142	143
6000	3	23	118	136	137
7000	0	2	41	98	124
8000	0	1	15	74	102
Other	9	26	89	191	342
Unrecognized	18	2	0	1	0
Total	1000	1000	1000	1000	1000

It was planned that TFVST would feature 20 headwords from each 1000 word frequency band, which meant that 50 words would be represented by a single test item. Accordingly, it was felt that JACET 8000 word bands that had an overlap of less than 50 with any of Nation's individual frequency levels should be excluded from that particular level. If included this would be an over-representation of that JACET 8000 word band. These exclusions were made, resulting in pools of words for each frequency level that contained less outliers in terms of word difficulty for Japanese learners of English. It was considered desirable to continue using Nation's lists as a basic framework as they provided a more relevant basis for international comparison than the JACET 8000 lists.

Filtered versions of the 1000 to 5000 level (referred to as 1k to 5k) lists from Nation (2009b) were used in the test; thus testing knowledge of the first 5000 word families of English. Words were selected for inclusion from randomly ordered cognate and noncognate sampling lists at each frequency level. In total, 20 headwords were sampled from each 1000 word frequency band, resulting in a test that had a reasonable sampling rate, yet would not take too long to administer. Preventing test-taker fatigue was an important consideration here, as a more time-consuming translation task would be administered ahead of TFVST in its initial validation study.

The test was limited to the first 5000 words of English, the same range as covered by the X_Lex test (Meara, 2005), as it was felt that this would be sufficient to record the vocabulary sizes of the Japanese university students who would

make up the majority of the participants in the study. Existing estimates of Japanese university students' average vocabulary sizes were 2000 (Shillaw, 1995) and 2300 (Barrow et al., 1999). 5000 has also been cited as the number of word families required for learners to engage in conversations of a general nature (Hu and Nation, 2000). Of course, the range of TFVST could easily be extended in order to meet the needs of different research or teaching situations.

Translations. A textbook on the JACET 8000 list (Aizawa, Ishikawa, and Murata, 2005) which employed Japanese translations was used as the main source of translations for the test. The translations used in this book describe 'core' meanings for the word families that the headwords represented, and thus were felt to be more appropriate than the most frequent meanings according to corpus data as listed in the Wisdom English-Japanese Dictionary (Inoue and Akano, 2008). Only if a word could not be found in the JACET 8000 textbook, then the most frequent translation from the Wisdom dictionary was inserted. Some words were also discarded from the test for the following reasons:

- They were listed as being archaic in the Wisdom dictionary (for example, 'ere').
- They were listed as being specific to British English in the Wisdom dictionary (the majority of students were presumed to have been taught a predominantly American English model in school (Matsuda, 2003)).
- The words were not 'content words' – that is, nouns, verbs, adjectives or adverbs – as in Nation (1993). The other 'function words' were not included as translating them would have been problematic. This study therefore assumes that the number of function words known is directly proportional to the number of content words known.

Discarded items were replaced by the next word on their sampling lists. Half of the cognate and noncognate words at each frequency level were then given false translations.

False translations were chosen by searching down the randomly ordered sampling lists for the next word that was of the same word class, and then attaching the translation for that word. Thus, a 2k noncognate adjective would be assigned the translation of another 2k noncognate adjective. This method prevented test-takers from being able to guess whether a translation was true or false on the grounds of word class, and meant that they were not presented with translations for very high frequency words together with lower frequency English words (or vice versa), which may have acted as a common sense clue that a pair was false. The list of Japanese translations for the test items was shown to a native speaker of Japanese to check that they were fully comprehensible, but no changes were made.

Presentation. The fact that TFVST did not feature a *Don't know* option meant that participants would have to provide a response to every item, even if they had to guess. In order to decrease participants' opportunity to strategically guess correct responses based on the (valid) assumption that there would be an equal number of true and false items in each frequency band, it was decided to divide the test into four sections of 25 items each, with the makeup of each section as described in Table 2. The items in each section were set to appear in a random order for each test to balance out any advantage or disadvantage that they might have had from being near to the start or end of a section. Lists of the items selected for the four sections of TFVST can be found in the Appendix.

Table 2

Breakdown of true and false items in each section of TFVST

Section	Number of true items	Number of false items
A (20*1k words, 5*2k words)	15	10

B (15*2k words, 10*3k words)	12	13
C (10*3k words, 15*4k words)	13	12
D (5*4k words, 20*5k words)	10	15

This format ensured that the number of true and false items varied with each section, and a sentence was added to the initial instructions to inform test-takers that the balance of true and false items may not be the constant. Care was taken to word the instructions in a manner that did not imply that the balance could never be the same, as this might have resulted in participants questioning their responses if they ended up with the same number of true and false options selected on consecutive sections. The instructions also informed participants that they had to provide responses for every item, and that the Japanese translations represented either 'core' or highest frequency meanings.

Scoring. As TFVST did not feature a *Don't know* response option, participants would almost certainly gain some marks for guessed items, unless they knew every word on the test. With this in mind, it was clear that number correct marking would not provide plausible estimates of vocabulary size; thus, negative marking was adopted as the scoring scheme for the test. The rationale for this was that if students stand a 50% chance of guessing the correct option for items that they have no knowledge of, then it is to be expected that the number of these items that they respond to incorrectly should also tend towards 50%. Accordingly it can be argued that deducting an additional mark for each incorrect response should account approximately for the number of items guessed at, and that the accuracy of this 'estimate of guessing' should increase with test length.

Investigation of TFVST

Research Questions

The following two research questions were investigated in a small-scale pilot study on TFVST, which aimed to provide a preliminary evaluation of its usefulness:

- Is the correlation between TFVST test scores and a same-vocabulary translation task significantly higher than the previous 'best result' for a vocabulary size test (Laufer and Goldstein (2004) between the passive recall and passive recognition sections of CATSS ($r = .58$, $p < .001$))?
- Is the test quicker to administer than other vocabulary size tests?

Method

Participants. Permission was obtained from the Centre for Language Education at Ritsumeikan Asia Pacific University for the study to be implemented in English course classes, subject to individual instructors' agreement. Teachers in every course level were contacted and asked if they would be able to administer the tests during lesson time. However, the schedules for the different levels meant that it was only possible to do them in two lower ability 'Fundamental English' classes, and to offer them as one of a selection of optional activities in two 'Intermediate English' classes. Fundamental English courses have a paper-based TOEFL target score of 450, while Intermediate English courses have a target of 500. In total 66 responses were received from the lower ability classes. Only three students from the intermediate level classes opted to take the test.

Two Japanese English teachers at the university, both of whom have completed Masters degrees at UK or US universities, also took the test in their own time. This gave a total of 71 participants. Ten of the students from the lower English classes were, however, of Korean or Chinese nationality; thus, their results were excluded from analysis on the

grounds that, despite having a high level of Japanese ability, there existed too great a chance that their scores would be influenced by problems understanding Japanese. Accordingly, a total of 61 sets of results were used for analysis. No rewards were given for participation, none of the participants had taken TFVST, and all of them were required to provide personal information and complete an online consent form in Japanese that formed the first page of the instrument.

Instruments. It was decided that, given the time and resources available for this project, the most appropriate criterion against which the concurrent validity of TFVST could be tested was a translation task, taking an approach similar to that of Eyckmans (2004, p.75-77) when investigating the Yes/No format. Although interviews, as used by Schmitt et al. (2001) with the VLT, may arguably have provided a more reliable criterion measure, they would have considerably restricted the number of students from whom data could have been collected in the study.

The research instrument was created to be administered online. The consent form and the two tests (the translation task and TFVST) were combined using the Surveygizmo (2005) tool, and all necessary instructions for students were included in Japanese. Instructions were initially written by the author and then checked over by a native speaker of Japanese. Adjustments were then made until both the author and the native speaker were both satisfied that they were as easy to understand as possible. Japanese was used in order to avoid any confusion that might arise from misunderstanding of English.

The online translation task, where students were presented with English words for which they had to type in Japanese translations, was created using identical vocabulary and structure to TFVST. Unlike in Eyckmans' study, the translation task in this experiment was administered first, as TFVST would have provided students with too many hints as to the correct translations of words should it have been the initial test. Although convenient from the perspective of marking, the first letter of the expected Japanese translation of words in the translation task was not given (as it was in Laufer and Goldstein's (2004) CATSS passive recall section) as this may have assisted students with their responses to TFVST.

For TFVST, the 25 items selected for each of the four sections of the test were placed in randomly ordered lists and then inserted into single-page tables with *True* and *False* tick box options (displayed in Japanese), only one of which could be selected for each word pair. The order of the word pairs on each page was also set to vary randomly for each new test-taker. This test therefore consisted of four single-page sections.

The instrument was designed to be fully self-explanatory in order that participants could just follow the instructions on the screen to take it. For the translation task, test-takers were informed that they should fill in 'X' if they did not know a translation for an English word. The first page of both of the tests stated that test-takers would not be able to return to previous pages once they had completed them, and that they could not use their browser's back button. It was also not possible to advance on to the next page of the instrument until responses had been provided for all items on the current page. Therefore, there was no possibility of participants missing questions.

Test-takers were also informed that there was no time limit; however, the instrument did require students to enter the time shown on their computer screen following completion of the consent form, completion of the open translation task, and completion of TFVST. The appearance of the instrument was checked in Microsoft Internet Explorer version 8, the default browser on all university computers, and no problems were noted.

Procedures. Students in Fundamental and Intermediate English classes were directed to the test by their teachers during class time via a temporary link placed in their own specific class page of the English course's online learning system. The two teachers who helped administer the study were briefed about the nature and content of the instrument in advance. They then gave a short explanation to their respective classes. Students were free to ask questions if they wished to, and were supervised by their teacher for the duration of their participation.

The two Japanese English teachers who participated in the study were contacted individually, and the content and nature of the instrument was explained to them. They both agreed to do the tests on their own, and they were sent a link

to the instrument by email. All participants were instructed to follow the instructions on the screen during the study, and all of the data was collected over a one week period.

Data analysis. The data from the tests was downloaded as a spreadsheet file. The translations were then marked in accordance with the lenient marking scheme employed by Eyckmans (2004, p. 81). This allowed for levels 1, 2 and 3 of her taxonomy to be accepted as correct (see Figure 2 for details of her taxonomy).

1. Correct translation
2. Correct translation but wrongly spelled or typed
3. Mistakes due to grammatical category
4. Undoubtedly incorrect translation or no response (X)

Figure 2: Marking taxonomy for the translation task (from Eyckmans, 2004 p.81)

A lenient marking scheme was chosen, as the main aim was to assess whether students had some level of knowledge of the form-meaning link for vocabulary items, rather than to evaluate their grammatical knowledge or ability to render words correctly in Japanese. Translations were permitted if they could be found in either Aizawa, Ishikawa, and Murata (2005) or the Wisdom English-Japanese Dictionary (Inoue and Akano, 2008), or if they appeared to have the same meaning as translations listed in one of these two sources and could also be found listed in the online ALC database (SPACEALC, 2000). Correct responses were awarded 1 mark, incorrect answers 0.

For TFVST, a negative marking scoring scheme where correct answers were awarded 1 mark and incorrect answers -1 was computed. A preliminary calculation of scores on both the translation task and TFVST revealed that one student had obtained a negative score on TFVST, yet had scored reasonably well on the translation task. A comparison scatterplot of the translation task against the TFVST results revealed that this student's results deviated considerably from the general correlation shown by the other participants. Further investigation revealed that the student had scored very lowly even on the 1000 frequency level words in the TFVST, despite having inputted translations in the translation that very closely resembled those given in TFVST task for these items. This suggested that the participant had not attempted to answer the TFVST section of the test earnestly, and accordingly their results were excluded from further analysis.

Normality of the data was initially checked by inspecting histograms. The TFVST data appeared to fit closely with the normal curve; however, the translation task data deviated slightly from normality. Kolmogorov-Smirnov tests then corroborated the evidence for non-normality of the translation task results. Based on this it was felt that the non-parametric Spearman Coefficient would be a more appropriate selection as a measure of correlation that the Pearson Product Moment Coefficient, which is regarded as being highly sensitive to data distribution (Bachman, 2004, pp.87-88).

Results and Discussion

General results. The descriptive statistics and Cronbach Alpha reliability coefficients for the tests are displayed in Table 3. The difference in means suggested that TFVST measured a lower level of knowledge of the form-meaning link than the translation task (passive recognition as opposed to passive recall (Laufer and Goldstein, 2004)). The reliability coefficient for TFVST was a little low; however, it was felt that this could be partly attributable to the large concentration of participants at a similar general English ability level. General English ability has previously been found to correlate well with vocabulary size test results (Read, 2007) and a wide range of abilities among participants has the effect of increasing the Cronbach Alpha coefficient (Burton, 2005). Additionally, given the expected favourable time efficiency of TFVST, there should be scope in many situations to include considerably more than 100 items when using this format. As a greater number of items is likely to produce higher values for Cronbach Alpha (Dörnyei, 2007, pp.206-207), the figure observed in this trial was considered to be acceptable.

Scores on TFVST decreased on average with each frequency band for the first four 1000 word bands; however, the mean score for the 5000 band was actually higher than both the 3000 and 4000 band means (see Table 4). This pattern was not observed in the translation task results, where the mean scores decreased for every frequency band. This finding suggests that TFVST may not produce reliable frequency profiles at lower frequency levels where participants have very little knowledge of the items being tested. It should also be noted, however, that the minimal difference in means between the 4000 and 5000 bands on the translation task suggests that the words that were randomly sampled for these two frequency levels may have exerted some influence on scores here.

Table 3

Descriptive statistics and reliability coefficients for TFVST and the translation task

N = 60	*M* (/100)	*SD*	Reliability (Cronbach α)
TFVST Negative marking	57.17	13.59	.73
Translation task	45.28	12.95	.94

Table 4

Word frequency level comparisons of scores on TFVST and the translation task

N=60	***M* (*SD*)**	
	TFTVT2 Negative marking	Translation task
1000 frequency level (/20)	16.63 (2.59)	17.13 (2.40)
2000 frequency level (/20)	13.03 (4.04)	9.90 (3.43)
3000 frequency level (/20)	8.73 (4.18)	6.73 (3.24)
4000 frequency level (/20)	8.10 (4.02)	5.93 (3.41)
5000 frequency level (/20)	10.67 (4.16)	5.83 (2.73)

Correlation between TFVST and the translation task. Spearman's coefficient showed a high degree of correlation between participants' total scores on the translation task and on the TFVST (r_s = .79, p < .001), which is an improvement on the .58 figure reported by Laufer and Goldstein (2004). The Fisher r-to-z transformation revealed that the difference with Laufer and Goldstein's result was statistically significant (z = 2.65, r < .01). Based on the assumption that the criterion of the translation task was an accurate measure of a higher level of participants' vocabulary knowledge, the strong correlation observed here suggests that the TFVST is worthy of further investigation.

Table 5

Time efficiency comparison of TFVST with the VLT and Yes/No format

Test	Average time taken (minutes:seconds)	Number of items tested	Average time per item (seconds)
TFVST2	5:44	100	3.44
Yes/No (Eyckmans, 2004 p.153)	4:45	60	4.45
VLT (Schmitt et al., 2001)	31:00	150	12.40

Test times. The average reported time taken to complete TFVST was 5 minutes 44 seconds (*SD*=1 minute 52 seconds). This result is contrasted with previously reported times for the VLT and the Yes/No format in Table 5. Although the figures reported for TFVST are only rough estimates and were not measured with stopwatch precision, they do suggest that TFVST may be more time efficient on average than both the VLT and the Yes/No format for Japanese learners of English. In particular the advantage over the VLT is striking, and this is likely to allow for a considerably higher sampling rate of words within a given time period than for that format.

Conclusion

The Usefulness of the Test

The results obtained suggest that the True/False translation test format investigated here holds considerable potential as a measure of vocabulary size in a Japanese context. TFVST performed well both in terms of correlation with a translation task criterion measure and time efficiency per item. The correlation observed was significantly stronger than that reported by Laufer and Goldstein (2004) in their comparison of CATSS passive recall and passive recognition components. The average time per item for TFVST was lower than previously reported figures for the VLT and Yes/No tests, suggesting that the new format may be able to achieve a higher sampling rate in a given time period than either of these tests.

Limitations

It is essential that the limitations of this investigation be taken into account when interpreting its results. The study described here was conducted on one small group of Japanese university students and teachers, a large number of whom were in the same level English class. It is not clear what effect a wider and more evenly spread range of English proficiencies among participants would have had on results; accordingly the generalizability of the findings here is somewhat limited. Additionally, although students were supervised throughout their participation, it was possible that some of them may have been able to cheat by copying their neighbours' answers or by looking up translations on the internet. A lack of incentive to perform well on the tests may also have resulted in participants not taking the tests in an earnest manner.

Suggestions for Future Research and Improvement of the Test Design

Following the results achieved here, it is hoped that TFVST type tests will be trialled with other more diverse groups of Japanese learners of English in the future. In particular, there is a need to investigate longer tests and the effect of changing the balance of true and false items, both of which may facilitate improved performance. The test format

presented here can also be readily adapted to other L1 groups. Accordingly, it is the hope of the author that investigations into its usefulness and performance will be conducted in different L1 contexts too.

References

Aizawa, H., Ishikawa, & Murata, T. (2005). *JACET 8000 EITANGO*. Kirihara.

Albrechtsen, D., Haastrup, K., & Henriksen, B. (2008). *Vocabulary and writing in a first and second language: Process and development*. Basingstoke: Palgrave Macmillan.

Alderson, J.C. (2005). *Diagnosing foreign language proficiency*. London: Continuum.

Alderson, J.C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. New York: Cambridge University Press.

Anderson, R.C. & Freebody, P. (1981). Vocabulary knowledge. In: Guthrie, J.T. (Ed.) *Comprehension and teaching: Research reviews*. Newark, DE: International Reading Association.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L.F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Barrow, J., Nakanishi, Y., & Ishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System 27*(2): pp.223-247.

Bauer, L. & Nation, I.S.P. (1993). Word families. *International Journal of Lexicography 6*: pp.253-279.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing 27*: pp.101-118.

Beglar, D. & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing 16*: pp.131-162.

Bertram, R., Baayen, R., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language 42*: pp.390-405.

Bertram, R., Laine, M., and Virkkala, M. (2000). The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology 41*(4): pp.287-296.

Buck, G. (1992). Translation as a language testing procedure: does it work? *Language Testing 9*(2): pp.123–148.

Burton, R.F. (2001). Quantifying the effects of chance in multiple-choice and true/false tests: item selection and guessing of answers. *Assessment and Evaluation in Higher Education 26*: pp.41-50.

Burton, R.F. (2002). Misinformation, partial knowledge and guessing in true/false tests. *Medical Education 36*(9): pp.805-811.

Burton, R.F. (2005). Multiple-choice and true/false tests: myths and misapprehensions. *Assessment and Evaluation in Higher Education 30*(1): pp.65-72.

Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford: Oxford University Press.

Ebel, R.L. (1979). *Essentials of educational measurement*. 3rd ed. Englewood Cliffs: Prentice Hall.

European Society of Anaesthesiology (2008). Abandoning negative marking. *European Journal of Anaesthesiology 25*: pp.349-351.

Eyckmans, J. (2004). *Measuring receptive vocabulary size*. LOT (Landelijke Onderzoekschool Taalwetenschap).

Inoue, N. and Akano, I. (2008). *The Wisdom English-Japanese dictionary*. Tokyo: Sanseido.

Ito, A. (2004). Two types of translation tests: their reliability and validity. *System 32*(3): pp.395-405.

Jordan, E. (2012). Cognates in Vocabulary Size Testing-a Distorting Influence? *Language Testing in Asia* 2(3): pp.5-17.

Kilgarriff, A. (1996). BNC frequency lists [online]. Available at: http://www.kilgarriff.co.uk/bnc-readme.html [accessed 2 September 2013].

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing 21*(2): pp.202-226.

Laufer, B. & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning 54*(3): pp.399-436.

Matsuda, A. (2003). Incorporating world Englishes in teaching English as an international language. *TESOL Quarterly 37*(4): pp.719–729.

Meara, P. (1992). EFL vocabulary tests. University College, Swansea: Centre for Applied Language Studies.

Meara, P. (1996a). The classical research in vocabulary acquisition. In Anderman, G. & Rogers, M. (Eds.) *Words, words, words*. Clevedon: Multilingual Matters, pp.27-40. Available at: http://www.lognostics.co.uk/vlibrary/meara1996b.pdf [accessed 2 September 2013].

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

Meara, P.M. (2005). X_Lex: the Swansea vocabulary levels test. v2.05. Swansea: Lognostics [online]. Available at: http://www.lognostics.co.uk/tools/index.htm [accessed 2 September 2013].

Meara, P. & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing 4*: pp.142-154.

Meara, P. & Jones, G. (1988). Vocabulary size as a placement indicator. In: Grunwell, P. (Ed.) *Applied Linguistics in Society*. CILT, London, pp.80–87.

Meara, P. & Jones, G. (1990). The Eurocentres' 10K vocabulary size test. Eurocentres Learning Service, Zurich.

Mendenhall, W., Beaver, R.J., & Beaver, B.M. (2008). *Introduction to probability and statistics*. Duxbury Press.

Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing 23*: pp.73-98.

Muijtjens, A.M.M., van Mameren, H., Hoogenboom, R.J.I., Evers, J.L.H., & van der Vleuten, C.P.M. (1999). The effect of `don't know' option on test scores: number-right and formula scoring. *Medical Education 33*(4): pp.267-275.

Nagy, W.E., Anderson, R., Schommer, M., Scott, J.A., & Stallman, A. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly 24*(3): pp.263-282.

Nation, I.S.P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I.S.P. (1993). Measuring readiness for simplified material: a test of the first 1,000 words of English. In Tickoo, M.L. (Ed.) *Simplification: Theory and Application*. RELC Anthology Series 31, pp.193-203.

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

Nation, I.S.P. (2008). *Teaching vocabulary: strategies and techniques*. Boston: Heinle.

Nation, I.S.P. (2009a). Personal Homepage [online]. Available at: http://www.victoria.ac.nz/lals/staff/paul-nation.aspx [accessed 2 September 2013].

Nation, I.S.P. (2009b). BNC 14k lists [online]. Available at: http://www.lextutor.ca/vp/bnc/nation_14/ [accessed 2 September 2013].

Nation, P. & Gu, P.Y. (2007). *Focus on vocabulary*. Sydney: National Centre for English Language Teaching and Research.

Rasch, D. & Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science 46*: pp.175-208.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal 19*: pp.12-25.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Read, J. (2007). Second language vocabulary assessment: current practices and new directions. *International Journal of English Studies 7*(2): pp.105-125.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke: MacMillan.

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): pp.55-88.

Shillaw, J. (1995). Using a word list as a focus for vocabulary learning. *The Language Teacher 19*(2): pp.58-59.

Shimizu, S. (2013). JACET 8000 level marker [online]. Available at: http://www.tcp-ip.or.jp/~shim/J8LevelMarker/j8lm.cgi [accessed 2 September 2013].

SPACEALC (2000). EITAROO on the web. Available at: http://www.alc.co.jp/ [accessed 2 September 2013].

SurveyGizmo (2005). SurveyGizmo online survey tool. [online]. Available at: http://www.surveygizmo.com/ [accessed 2 September 2013].

Swan, M. (1985a). A critical look at the communicative approach. Part 1. *English Language Teaching Journal 39*: pp.2-12.

Swan, M. (1985b). A critical look at the communicative approach. Part 2. *English Language Teaching Journal 39*: pp.77-87.

Webb, S. (2007). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research 11*(1): pp.63-81.

Appendix

List of Items Used in TFVST

SECTION A

ADVANCE	進歩
CHARACTER	特性
DRY	確かな
EASY	重い
FARM	給料
GERMANY	ドイツ
GROW	育つ
HAND	写し
HERE	最も
INDUSTRY	教会
INFORM	得点する
MAJOR	専攻する
NEW	幅の広い
PARTY	パーティー
PROCESS	過程
PROMOTE	昇進させる
PURE	純粋な
SIT	座る
SITE	場所
SOCIAL	楽しい
SUGGEST	示唆する
THEN	...さえ
TIME	時間
WINE	ワイン
WISE	賢い

SECTION B

BIBLE	キリスト教の聖書
CASUALTY	死傷者
CHAPEL	毛布
CONTRIBUTE	移動する
CRITERION	特徴
CUSHION	クッション
DIMENSION	側面
FETCH	締め出す
FORMAL	まるごとの
HILL	形
INTELLIGENCE	技術
JEANS	ジーパン
LIBERAL	自由主義の
MANUAL	体を使う
NEVERTHELESS	それにもかかわらず
ORCHESTRA	ギャング
PUBLISH	出版する
REPLY	返事をする
RIVER	皿
SKY	キャンプ場
SPIN	ブレーキ
TREMENDOUS	巨大な
TRIVIAL	取るに足りない
ULTIMATE	厚い
VEGETABLE	苦痛

SECTION C

ANTIQUE	古くて価値のある
ARTIFICIAL	人工の
BAIL	保釈
BEE	革
BUBBLE	泡
CEASE	終わる
DELIBERATE	大胆な
EMPIRE	悲嘆
FIN	ひれ
GALLON	ガロン
IRONY	のど
ISRAEL	イスラエル
IVY	パンの1塊
MUTUAL	相互の
OUTRAGE	平鍋
PALM	泥棒
PUNISH	罰する
RAVE	激賞する
RECITE	同行する
REVEAL	癒す
SHATTER	保持する
SOAP	ラテン語
TRACTOR	農業用トラクター
VERIFY	確認する
WHEELCHAIR	干草

SECTION D

ADMINISTER	管理する
BLEND	分解させる
BLOUSE	ブラウス
BROOM	人質
DEADLINE	締め切り
DEFICIENCY	薄暗がり
DIVINE	現代の
EMIGRATE	統合する
FLUID	悪魔
FOG	歳入
GASP	息をのむ
INTERCEPT	割り込む
JAZZ	バレエ
JURY	混合物
MINIATURE	小型の
MULTITUDE	多数
PHYSICS	飾り
PIERCE	妨げる
PUBLICIZE	損失を負う
RENDER	…の士気をくじく
RETREAT	退却する
SNAKE	ヘビ
TESTAMENT	芝土
TRADESMAN	マホガニー
UNDERWEAR	下着