Vocabulary Learning and Instruction
Volume 8, Issue 1, March 2019
http://vli-journal.org

Evaluating the Efficacy of Yes–No Checklist Tests to Assess Knowledge of Multi-Word Lexical Units

Raymond Stubbe[a] and Yumiko Cochrane[b]
[a]*Nagasaki University;* [b]*Kyushu Sangyo University*
https://doi.org/10.7820/vli.v08.1.stubbe.cochrane

Abstract

One of the many challenges facing Japanese university students studying English is the multi-word phrase. The English language contains a large number of such multiple-word items, which act as single words with a single meaning. This study is concerned with evaluating the efficacy of yes/no checklist tests to assess knowledge of multi-word units. Participants (n = 206) took a yes–no test of 30 real words and 15 pseudowords. The 30 real words were selected from the students' textbook, based on the teacher's intuition of the words and multi-words posing the greatest learning burden for the students. Twenty-one of the selected words were single-word items. The remaining nine were multi-words, such as "get up" and "take turns". Forty-five minutes following completion of the yes–no test, an English to Japanese translation test of the same 30 real words was taken by the same participants to evaluate the efficacy of yes/no test. Results suggest that the yes–no vocabulary test format may be able to measure student knowledge of multi-word lexical units as (or more) effectively than single-word units.

Key words: Multi word lexical units; yes no vocabulary tests; translation tests; overestimation; Japanese EFL learners.

1 Introduction

One of the many challenges facing Japanese university students studying English is the multi-word phrase. This study is concerned with evaluating the efficacy of Yes/No checklists to assess student knowledge of multi-word units. As this enquiry involves comparing multi-word lexical units with single-word items using a yes–no checklist test and an English to Japanese (L2 to L1) translation test, these three entities will be briefly reviewed.

1.1 Yes–No Checklist Tests

Recognized as the easiest vocabulary test to administer to a group of learners, yes–no checklist tests may be an efficient means of enquiring about student knowledge of multi-word units. According to Read (2007, pp. 112–113), "Despite its simplicity, the Yes/No format has proved to be an informative and cost-effective means of assessing the state of learners' vocabulary knowledge, particularly for

Please cite this article as: Stubbe, R. and Cochrane, Y. (2019). Evaluating the Efficacy of Yes–No Checklist Tests to Assess Knowledge of Multi-Word Lexical Units. *Vocabulary Learning and Instruction, 8* (1), 62–69. https://doi.org/10.7820/vli.v08.1.stubbe.cochrane

placement and diagnostic purposes." Yes–no tests present learners with a list of decontextualized words and have them signify their knowledge of each item by either checking (or circling) that word, or by selecting either "yes" or "no". As yes–no tests depend on self-reporting, the actual lexical knowledge of test-takers cannot be verified. One concern with this format is whether yes–no results accurately reflect the test-takers' knowledge of the tested items, or do they overestimate the number of words truly known (Read, 1993, 2000). To compensate for the potential of students claiming knowledge of words they do not actually know the meaning of (labeled *overestimation*), pseudowords were introduced to the yes–no format by Anderson and Freebody (1983). In such yes–no tests, claiming knowledge of a real word is known as a "hit", while claiming knowledge of a nonword is called a "false alarm" (FA). Not claiming knowledge of a real word is labeled a "miss" and not claiming knowledge of a nonword is a "correct rejection". A number of scoring formulae have been devised to adjust yes–no test scores using FA and real-word hit data. The simplest formula *h-f,* subtracts the FA rate from the hit rate. Along with overestimation, yes–no tests are also liable to underestimation, where students do not signal knowledge of items they actually do know if tested with a translation test or interview. Unlike pseudowords for overestimation, nothing has been developed to indicate possible underestimation in yes–no tests.

A number of prior studies have investigated yes–no test results by comparing them with a criterion measure. Mochida and Harrington (2006) used a multiple-choice test, the *Vocabulary Levels Test* (VLT) (Nation, 1990; Schmitt, Schmitt & Clapham, 2001). Pellicer-Sánchez and Schmidt (2012) utilized personal interviews as the criterion measure to ascertain participants' actual vocabulary knowledge in order to determine the amount of overestimation of vocabulary knowledge on their yes–no test. Eyckmans (2004, p. 77) selected L2 to L1 translation tests as her criterion measure because "asking participants to provide mother-tongue equivalents of the target language words was the most univocal way of verifying recognition" on the yes–no tests. As multiple-choice test results are liable to guessing effects (Stewart & White, 2011) and personal interviews are very time-consuming, the L2 to L1 translation test was selected as the criterion measure.

1.2 L2 to L1 Translation Tests

Similar to yes–no tests, the L2 to L1 translation test to be used in this study presents learners with a list of decontextualized words and has them translate the English words into Japanese. L2 to L1 translation tests measure *passive recall* ability (Laufer & Goldstein, 2004), which is the ability to recall the meaning of an L2 word in the first language, or the ability to recall the meaning of an L3 word in the L2. Echoing Eyckmans (2004), other researchers agree that translation ability is a strong indicator of which words students can actually understand while reading (Waring & Takaki, 2003) and listening (Pellicer-Sánchez & Schmitt, 2012).

1.3 Multi-word Units

As the name suggests, multi-word units are lexical items made up of two or more words, which combine to have a single meaning that differs (however slightly)

from the meanings of each individual word (Bennett, 2017; Schmitt & McCarthy, 1997, p. 329). According to Schmitt (2007):

> English has a large number of these multiple-word-item lexemes that behave as a single word with a single meaning … There are a number of different kinds of multiword units, including compound words (playpen), phrasal verbs (give up), fixed phrases (ladies and gentlemen), idioms (put your nose to the grindstone), and proverbs (A stitch in time saves nine). (p. 747)

2 Methodology

Participants in this study were first-year university students enrolled in one of five mandatory English classes (n = 206) at a public university in southern Japan. Although English levels were not available for each student, the five participating departments tend to have students at the high beginner level.

All tested items (k = 30) were selected from the students' textbook, based on the teacher's intuition of the words and multi-words posing the greatest learning burden for the students. Twenty-one of the selected items were single-word items. The remaining nine were the following multi-words: brush (your) teeth; easy to; fish bowl; get dressed; get ready; get up; put on makeup; take turns; and, wake up.

For the yes–no test, the 30 items were combined with 15 pseudowords, and all 45 items were randomly ordered and then numbered 1 to 45. Each numbered item was followed by a bubbled "y" (for yes I know this item) and a bubbled "n" (for no I do not know this item). For the translation test, the 30 items were also randomly ordered, then numbered 1 to 30. Following each numbered item, two spaces appeared (_____ _____) and students were encouraged to provide two answers if possible. This was done to avoid the possibility of students writing only one answer (as naturally occurs when only one space is provided) that happens to be incorrect while still knowing a different, correct answer.

The yes–no test was given at the beginning of the second class of the term, prior to commencing the textbook, and took under 10 min. The textbook commenced in the third class. The translation test was given toward the end of that second class, to maximize test-pairings and avoid any possible acquisition of some of the tested items between testing. This second test also was completed in about 10 min. The yes–no test was scored using an optical scanner. The translation test was hand-marked by a native Japanese English university teacher with high-level English ability. Ten percent of the translation tests (21) were copied prior to marking and also marked by a second native Japanese university teacher with high-level English ability. Co-rater agreement was good at 92%.

3 Results

One participant signaled knowledge of 11 of the total 15 pseudowords. As an extreme outlier, his results have been deleted from the data. The overall mean on the yes–no test was 21.97 of the 30 items (73%; Table 1). Compared to other studies

Table 1. Descriptive Stats for Yes–No and Translation Tests

Test	Mean	SD	Min	Max	Mean%	YN%-Tr%	*r*
YN 30	21.95	3.37	7	28	73.2%		0.539
FAs 15	1.61	1.54	0	6	10.7%		0.086
FAs 10	0.44	0.79	0	4	4.4%		0.043
Tr 30	18.52	3.53	10	30	61.7%	11.5%	1

Note: YN denotes yes–no test real-word items; FAs denote false alarms; Tr denotes translation test; *r* denotes correlation (Pearson Product Moment) on translation test scores; $n = 205$, $k = 30$.

Table 2. Single versus Multi-Word Items, Both Tests

Test	*k*	Mean	SD	Min	Max	Mean%	YN-Tr
YN Single	21	16.19	2.38	5	21	77.1%	
Tr Single	21	13.38	2.64	6	21	63.7%	13.4%
YN Multi	9	5.77	1.51	2	9	64.1%	
Tr Multi	9	5.37	1.68	1	9	57.2%	6.9%

involving Japanese learners (Mochida & Harington, 2006; Stubbe, 2012; 2015), the FA rate was high, likely resulting from the composition of the pseudowords used in this study. Five of the 15 created by this author were based on Laufer's (1998) concept of *synforms*, and they proved unexpectedly attractive. Removing these 5 synform pseudowords would result in a FA rate of 4.4%, which is in line with the rates mentioned in the other aforementioned studies. For this reason, only the 10 pseudowords will be used hereafter. The overall mean for the translation test was 18.52 of the 30 words (62%). The correlation (Pearson Product Moment) between the yes–no test and the translation test was moderate at 0.539.

Table 2 presents a comparison of the single and multi-word results for both tests. As there were different numbers of single words (21) and multi-words (9), the mean percentage figures are the most useful. On the yes–no test, students signaled knowledge of 77.2% of the 21 single words and 64.1% of the nine multi-words. The students signaled that they knew 13.1% more single words. However, on the translation test, the students correctly translated 63.7% of the single words (a decrease of 13.5%) and 57.2% of the multi-words (a decrease of 6.9%).

Comparing the responses of the yes–no test directly with the answers on the translation test on a student-by-student item-by-item basis yields four possible outcomes:

1. *yes* on the yes–no test matched with a correct answer on the translation test, labeled "known".
2. *no* on the yes–no test matched with an incorrect answer on the translation test, labeled "unknown".
3. *yes* on the yes–no test matched with an incorrect answer on the translation test, labeled "overestimation" (on the yes–no test).
4. *no* on the yes–no test matched with a correct answer on the translation test, labeled "underestimation" (on the yes–no test).

Table 3 presents these possible matches between the two tests. The second and third columns present outcomes *a* and *b*, in which the yes–no test response is verified by the translation test (*known* and *unknown*). The single-word items were more accurate than the multi-word items by 3.9% (79.4% – 75.5%). The last two columns show the amount and type of inaccurate yes–no test responses. With the single-word items, the participants overestimated their word knowledge by 17% but underestimated it by only 3.6%. However, with the multi-word items, less overestimation occurred (15.7%), while much more underestimation occurred (8.8%). This greater amount of underestimation for the multi-word items offsets their overestimation amount, which accounts for the smaller decrease in mean scores between the yes–no and translation tests (6.9%), compared to the single-word decrease (13.5%; see Table 2).

In a further analysis, the scoring formula *h-f* was used to determine the usefulness of the 10 pseudowords for adjusting yes–no results to better reflect demonstrable knowledge of the tested items (Table 4). For each group (the 21 single-word items, and the 9 multi-word items) applying the *h-f* scoring formula resulted in adjusted yes–no scores that were closer to the translation test scores than the original yes–no test scores, but considerably closer for the multi-word items. However, the correlations between the translation test scores and the yes–no results were weaker for the multi-words, especially when adjusted by *h-f*.

An item analysis was also undertaken. As can be seen in Table 5, the mean difference between the yes–no scores and the translation test scores for the single words was 27.52, representing 13.4% of the 205 students. Noteworthy items included *helmet*, the only underestimated single word (-11), as well as *monitor* and *schedule*, which had equal scores. The word *mechanic* was the most overestimated single word on the yes–no test.

Table 3. Direct Comparison of the Yes–No and Translation Tests – Person-by-Person, Item-by-Item

Item	Known	Unknown	Overestimation	Underestimation
Single words	60.1%	19.3%	17.0%	3.6%
Multi-words	48.3%	27.1%	15.7%	8.8%
Differences Single–Multi	11.7%	–7.80%	1.30%	–5.20%

Table 4. Applying h-f Correction for Guessing Formula to Yes–No Test Results

Test/items	Mean%	Diff (__ - Tr)	*r*
YN single-word items (21)	77.1%	13.5%	0.534
Single-word items *h-f*	68.8%	5.2%	0.387
Tr single-word items	63.6%		1
YN multi-word items (9)	64.1%	6.9%	0.383
Multi-word items *h-f*	59.7%	2.5%	0.313
Tr multi-word items	57.2%		1

Note: As the number in each group differs (i.e., 21, 9), means are given as percentages; diff denotes difference; *r* denotes correlation with Tr scores.

Table 5. Item Analysis: Single words (*n* = 205)

Word item (21)	YN single	Tr single	Diff YN-Tr
Boring	193	152	41
Deliver	190	161	29
Dorm (dormitory)	55	37	18
During	203	192	11
Helmet	184	194	–10
License	178	164	14
Mechanic	196	43	153
Monitor	187	187	0
Qualification	59	33	26
Saw (noun)	108	29	79
Schedule	197	197	0
Shave	144	122	22
Shelf	142	114	28
Sink	157	141	16
Skill	205	197	8
Strategy	191	166	25
Textbook	202	201	1
Tool	199	177	22
Tough	196	156	40
Tutor	32	13	19
Vase	100	64	36
Mean (*n* = 205)	158	130.48	27.52

Table 6. Item Analysis: Multi-Words (*n* = 205)

Word item	YN Multi	Tr Multi	Diff YN-Tr
Brush (your) teeth	196	196	0
Easy to ~	158	142	16
Fish bowl	57	19	38
Get dressed	104	99	5
Get ready	162	150	12
Get up	197	173	24
Put on makeup	26	53	–27
Take turns	84	62	22
Wake up	198	161	37
Mean (*n* = 205)	131.33	117.22	14.11

Table 6 presents an item analysis for the multi-words. The mean scores of 131.33 and 117.22 for multi-words on the yes–no and translation tests were substantially lower than for the single-word items (158.00 and 130.48). However, the difference between the yes–no scores and the translation test scores for the multi-words was 14.11, only 6.8% of the 205 participants. In addition, the correlation between the number of correct responses for each target word in the two

test formats was stronger for the multi-word items than for the single-word items (r = 0.951 and 0.852, respectively). Finally, the item *put on makeup* accounted for 21.5% of the underestimation noted in Table 3.

4 Conclusion

This has been an investigation into the efficacy of yes–no checklist tests for assessing student knowledge of multi-word units. A yes–no test comprises 21 single-word items, and 9 multi-word items were taken by 205 university students in Japan. Students checked more single-word items than multi-word items on this yes–no test (77.1% and 64.1%, respectively; see Table 2). The yes–no test was then followed by a translation test of the same items. Again participants correctly translated more single-word than multi-word items (63.6% and 57.2%, respectively). However, a greater decrease in scores between the two tests was observed with the single-word items compared to the multi-word items (13.4% vs. 6.9%, respectively; see Table 2). While the single-word items primarily suffered from overestimation on the yes–no test (17.0%) with only a small amount of underestimation (3.6%), the multi-word items suffered from both overestimation (15.7%) and underestimation (8.8%). This underestimation helped to reduce the effect of the overestimation for the multi-words on the yes–no test, resulting in closer yes–no and translation test scores. The application of the *h-f* correction formula to the yes–no test scores resulted in adjusted scores that were much closer to the translation test result for the multi-words than for the single words. While correlations presented in Table 4 favored the single words (0.534 vs. 0.383), the item analysis correlations between the two test formats favored the 9 multi-words (0.951) over the 21 single words (0.852).

The above results suggest that the yes–no vocabulary test format may be able to measure student knowledge of multi-word lexical units as (or more) effectively than single-word units. One weakness of the present study is the small number of items tested, especially the nine multi-word items. Future studies should increase the number of multi-word items to match the number of single-word items.

References

Anderson, R. C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. A. Hutson (Ed.), *Advances in Reading/Language Research* (Vol. 2, pp. 231–256). Greenwich, CT: JAI Press.

Bennett, P. (2017). Using cognitive linguistic principles to encourage production of metaphorical vocabulary in writing. *Vocabulary Learning and Instruction, 6*(2), 31–39. doi: 10.7820/vli.v06.2.Bennett.

De Veaux, R., Velleman, P., & Bock, D. (2008). *Stats: Data and models*. Essex, UK: Pearson Education Ltd.

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

Laufer, B. (1998). The concept of 'synforms' (similar lexical forms) in vocabulary acquisition. *Language and Education, 2*(2), 114–132.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399–436. doi:10.1111/j.0023-8333.2004.00260.x.

Laufer, B., & McLean, S. (2016). Loanwords and vocabulary size test scores: A case of different estimates for different l1 learners. *Language Assessment Quarterly, 13*(3), 202–217. doi: 10.1080/15434303.2016.1210611.

Mochida, A., & Harrington, M. (2006). YN test as a measure of receptive vocabulary. *Language Testing, 23*(1), 73–98. doi: 10.1191/0265532206lt321oa.

Nation, I. S. P. (1990). *Teaching and learning vocabulary.* Boston, MA: Heinle and Heinle.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing, 29*(4), 489–509. doi:10.1177/0265532212438053.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. doi:10.1111/lang.12079.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing 10*(3), 55–71. doi: 10.1177/026553229301000308.

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

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

Schmitt, N. (2007). Current perspectives on vocabulary teaching and learning. In: J. Cummins & C. Davison (Eds.), *International Handbook of English Language Teaching* (p. 15). Boston, MA: Springer International Handbooks of Education.

Schmitt, N., & McCarthy, M. (1997). *Vocabulary: Description, acquisition and pedagogy.* Cambridge, UK: Cambridge University Press.

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), 55–89. doi: 10.1177/026553220101800103.

Stewart, J., & White, D. (2011). Estimating guessing effects on the vocabulary levels test for differing degrees of word knowledge. *TESOL Quarterly, 45*(2), 370–380. doi: 10.5054/tq.2011.254523

Stubbe, R. (2012). Do pseudoword false alarm rates and overestimation rates in Yes/No vocabulary tests change with Japanese university students' English ability levels. *Language Testing, 29*, 471–488. doi:10.1177/0265532211433033.

Stubbe, R. (2015). Replacing translation tests with yes/no tests. *Vocabulary Learning and Instruction, 4*(2), 38–48. doi: 10.7820/vli.v04.2.stubbe.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2), 130–163.