Vocabulary Learning and Instruction
Volume 4, Issue 2, December 2015
http://vli-journal.org

Article

Replacing Translation Tests With Yes/No Tests

Raymond Stubbe
Kyushu Sangyo University
doi: http://dx.doi.org/10.7820/vli.v04.2.stubbe

Abstract

Along with personal interviews, individual word translation tests from the target language to the mother tongue are recognized as a reliable method of determining students' actual lexical knowledge. However, as most English as a foreign language teachers are aware, the marking of these tests can be a laborious task. A far easier vocabulary testing format is the Yes/No (YN) checklist test, which can examine a large number of words while not over-burdening the marker. Pseudowords, which look like real words but do not bear meaning, have been added to the YN format to check for evidence of overestimation of lexical knowledge by test-takers. Four scoring formulae, which adjust YN results according to the number of pseudoword reports, have become established in the literature. Of these, the *h-f* formula has become recognized as the simplest to use for adjusting YN scores. This study presents a regression-based prediction formula derived from the *h-f* results in a pilot study, which was then applied to the YN *h-f* adjustments in a second study (the main study) to predict actual vocabulary knowledge as demonstrated by a meaning recall translation test of the same items. This prediction formula, labeled *h-fRF*, was compared with another regression-based formula as well as the original *h-f* formula. Results showed that 54% of the 455 individual *h-fRF* predictions were within 5% (4.8 of 96 words) of matching translation test scores, and 82% were within 10%, which were better than the other formula predictions. These results may be of interest to classroom teachers as they suggest that by using the *h-fRF*, the burden of marking translation tests can be reduced by the far easier YN test format.

1 Background

The ability to recall the meaning of individual words while reading has long been recognized as a pre-requisite for successful reading comprehension (Beglar & Hunt, 1999; Qian, 1999, Stæhr, 2008 and others). For second/foreign language (L2) learners, the easiest way to demonstrate this recall ability is by translating the encountered words into their native language (L1). Consequently, when judging the lexical suitability of texts for use in the English as a foreign language (EFL) classroom, teachers often utilize L1–L2 passive recall translation tests. This study examined whether or not the easier to administer Yes/No (YN) test can be used to predict scores on a meaning recall translation test and possibly replace this latter, more cumbersome format.

Please cite this article as: 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

1.1 YN Vocabulary Tests

YN vocabulary tests present learners with a list of words, usually selected from word frequency lists, and ask them to signify their knowledge of each item by either checking that word or by selecting either "yes" or "no". Read (2007, pp. 112–113) notes: "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 placement and diagnostic purposes." As YN tests rely solely on self-reporting, the actual lexical knowledge of the students cannot be verified. One concern with the YN format is whether test results accurately reflect the test takers' knowledge of the selected items, or if the results overestimate the number of words actually known (Read, 1993, 2000). To compensate for the potential of students claiming knowledge of words they actually do not know the meaning of (overestimation), pseudowords, or non-real words, were introduced to the vocabulary checklist test by Anderson and Freebody (1983). Pseudowords were introduced to the field of L2 acquisition by Meara and Buxton (1987).

The use of pseudowords in YN tests has remained widespread through present-day versions. In these tests, knowledge of a real word is known as a hit, while claiming knowledge of a pseudoword is a false alarm (FA). Not claiming knowledge of a real word is labeled a miss and not claiming knowledge of a pseudoword is a correct rejection. Claiming knowledge of words that do not exist is seen as an indication of falsely claiming knowledge of real words (overestimation).

There presently exists three approaches to utilizing YN test pseudoword data. One use is to set a maximum acceptable number of pseudowords beyond which "the data are discarded as unreliable" (Schmitt, 2010, p. 201). Schmitt, Jiang, and Grabe (2011) set their acceptance limit at three (10% of their 30 pseudowords). Barrow, Nakanishi, and Ishino (1999) set the same cut-off point for the 30 pseudowords used in that study. Stubbe (2012b) demonstrated that a cut-off point of four (12.5% of the 32 pseudowords) better suited those YN test results.

Another use of YN pseudowords is to adjust the YN scores using a correction for guessing formula. The test results from learners checking pseudowords are adjusted using a variety of formulae, to better reflect their actual vocabulary knowledge. Four such established formulae were compared in Huibregtse, Admiraal, and Meara (2002): *h-f, cfg, △m, and Isdt.* With the first formula, *h-f,* (Anderson & Freebody, 1983), the proportion of FAs relative to the total number of pseudowords, the FA rate (*f*), is subtracted from the proportion of hits relative to the total number of real-word items, the hit rate (*h*), to create the formula: true hit rate = *h-f.* The remaining correction for guessing formulas are slightly more complicated and are presented below:

cfg (correction for guessing: Meara & Buxton, 1987):

$$cfg = \frac{h\text{-}f}{1 - f}$$

$\triangle m$ (Meara, 1997):

$$\Delta m = \left(\frac{h\text{-}f}{1-f}\right) - \left(\frac{f}{h}\right)$$

Isdt (Huibregtse et al., 2002):

$$Isdt = 1 - \frac{(4 * h * (1-f)) - (2 * (h\text{-}f)) * (1 + h\text{-}f)}{(4 * h * (1-f)) - (h\text{-}f) * (1 + h\text{-}f)}$$

Huibregtse et al. (2002) found that their *Isdt* formula had the best prediction ability of the four correction formulae, but that the simpler *h-f* formula (Anderson & Freebody, 1983) worked just as well under most conditions. Mochida and Harrington (2006) and Stubbe (2012a) similarly report that *Isdt* had the highest correlation of the four correction formulae with a second multiple-choice test of the same items, while YN raw hits had the lowest correlation. Eyckmans (2004) however, comparing YN test results with a meaning recall (L2 to L1 translation) test reported that the *cfg* formula had higher correlations than *Isdt*. Eight years following Huibregtse et al.'s (2002) study, Schmitt (2010, p. 201) noted that "it is still unclear how well the various adjustment formulae work."

Pellicer-Sánchez and Schmitt (2012) compared the same four scoring formulae, using subsequent meaning recall student interviews as the criterion measure. They found that each formula provide a mean score that was higher than the matching interview score, with $\triangle m$ providing the closest mean score. In other words, all four of the established YN scoring formulae overestimated the testees' demonstrable lexical knowledge. A correlational analysis found that all formulae provided high correlations with the interview results for the non-native speakers ($r > .796$), with *h-f* proving superior. Despite the high correlations, each formulae provided adjusted YN scores that overestimated the testees' actual vocabulary knowledge.

In a more recent study, Stubbe and Hoke (2014; using pre-existing data from Stubbe, 2013) compared the same four scoring formulae evaluated in Huibregtse et al.'s (2002) study. It was also found that *h-f* had the highest correlations with translation scores (see Table 1). Results also suggested that residuals, which are the differences between the correction formula predicted score and the actual translation score for each participant, were lowest for *h-f.* Residuals were calculated using the root mean square error (RMSE) method described in De Veaux, Velleman, and Bock (2008). As Stubbe and Hoke (2014) demonstrated that *h-f* was superior to *cfg,* $\triangle m$ and *Isdt* in terms of predicting translation scores, only *h-f* was chosen for inclusion in this present study.

1.2 Improving YN Scores using Regression Analysis

A third usage for pseudowords (or false alarm data) was introduced by Stubbe and Stewart (2012): the creation of a standard least squares (multiple regression) model and formula which can be used to predict translation test scores using self-reports of lexical knowledge (real-word and pseudoword) on a YN test.

Table 1. Means, SDs, Range, Correlations, and Residuals of Applying the Four Correction Formulae ($n = 455$; from Stubbe & Hoke, 2014, p. 74)

Test/Formula	Mean	SD	*r*	residual
Tr Scores	27.05	12.16	1	–
YN hits	48.82	17.23	0.721	24.82
FA Counts	2.17	3.16	-0.142	–
h-f	42.29	16.53	0.833	17.84
cfg	45.46	17.50	0.807	21.19
△ m	33.11	26.57	0.739	20.31
Isdt	50.02	13.55	0.775	24.56

Note. Tr = translation test; FA = false alarms (pseudoword reports); SD = standard deviation; *r* = correlation (Pearson Product-Moment) with translation test scores; Residual was calculated by squaring the differences between each translation test score and each of the five predictions, summing those squares, calculating the mean (df = 453) and finally acquiring the square root.
Source. Stubbe and Hoke (2014) "Comparing Yes/No Test Correction Formula Predictions of Passive Recall Test Results" which first appeared in The 2013 Pan-SIG Conference Proceedings published by the Japan Association of Language Teaching (pp. 72–78).

The use of regression analysis with YN test results, though not that common, is not unprecedented (Mochida & Harrington, 2006).

Stubbe and Stewart (2012) presented two scoring formulae derived using multiple regression analysis, with YN test real-word scores and pseudoword scores as two independent (predictor) variables and translation test scores as the dependent variable. The first formula was based on the full 120 real-word and 32 pseudowords YN item list, and had an r^2 of 45.2%. This formula was reported as "True number of words known $= 8.14 + (0.41 \times \text{YN Score}) - (1.94 \times \text{FAs})$" (Stubbe & Stewart, 2012, p. 5), where 8.14 words represents the intercept on the *y*-axis. For every word reported known on the YN test, add 0.41 words truly known. For every FA, subtract 1.94 words. To illustrate, one student reported 78 words as known on the YN test, and checked two pseudowords, so her true score would be calculated as 35.34, $\{8.14 + (.41 \times 78) - (1.94 \times 2)\}$.

This original prediction formula was improved by utilizing item analysis to select 40 of the 120 real words on the YN test which had "the highest phi correlations to translation test results, and the 9 pseudowords with the highest negative point biserial correlations to overall translation test scores" (Stubbe & Stewart, 2012, p. 6). The resulting prediction formula was reported as "True number of words known $= 3.26 + (.51 \times \text{YN Score}) - (2.39 \times \text{FAs})$" and had an r^2 of 59.1% (Stubbe & Stewart, 2012, p. 6; hereinafter referred to as *S&SRF*, for Stubbe & Stewart Regression-based Formula). Using this prediction formula, the same student as above would receive a true score of 38.77, which is considerably closer to her actual translation score of 39 than the original prediction of 35.34.

2 Aim

Three of the established YN scoring formula, *cfg, Δm,* and *Isdt*, may be too cumbersome to be of much use to regular classroom teachers. Though easier to use, the *h-f* adjusted YN scores were not very close to the actual translation test scores (see Table 1). The aim of this study is to introduce and assess a simple regression-based approach to scoring YN tests and to compare that approach to *S&SRF* and *h-f.* As *h-f* had a stronger correlation and a smaller residual than *cfg, Δm,* and *Isdt*, these latter three formulae will not be included.

3 Method

For clarity, the methodologies employed in the pilot study (Stubbe & Yokomitsu, 2012) and the main study (Stubbe, 2013) are reviewed in this section.

3.1 Pilot Study

For the pilot study, four English loanwords (LWs) and four non-loanwords (NLWs) were randomly selected from the top half and the bottom half of each of the eight word frequency levels in the *JACET List of 8000 Basic Words* (JACET Basic Word Revision Committee, 2003; hereinafter the *JACET8000*); for a total of 64 items for each group. Regrettably, three words were found to be in the wrong frequency level and one NLW turned out to be a LW. These four items along with their corresponding member from the opposite group (LW or NLW) had to be deleted from the item pool, leaving 120 words to be tested (Stubbe & Yokomitsu, 2012). A YN test was created containing these 120 words plus 32 pseudowords, all randomly ordered. All pseudowords were randomly selected from *Tests 101–106* of the *EFL Vocabulary Tests* (Meara, 2010). A translation test (English to Japanese, L2–L1) was also created which contained the same 120 words, also randomly ordered. The L2–L1 format test was chosen because translation ability is a strong indicator of which words students can actually understand while reading (Waring & Takaki, 2003) and "asking participants to provide mother-tongue equivalents of the target language words was the most univocal way of verifying recognition" (Eyckmans, 2004, p. 77).

Both the YN and translation tests used in the pilot study were given to Japanese university students enrolled in mandatory English classes ($n = 71$). TOEIC Bridge scores for the participants ranged from 90 though 140, roughly equivalent to 200 through 240 on the TOEIC. The YN test was given at the beginning of class and the translation test was given towards the end of that same class. This was done to ensure each YN test was paired with a translation test.

3.2 Main Study

To improve the separation between adjacent *JACET8000* levels in the main study (Stubbe, 2013), words were sampled only from the bottom half of each level. It was also decided to reduce the total number of tested items to 96; six LWs and six NLWs from each of the eight levels of the *JACET8000*. Forty-four of the 120 words

in the pilot study were included in the main study's item pool. The other 52 words were randomly selected from the various levels of the *JACET8000* as required to complete the desired six LWs and six NLWs per frequency level. Also, only 16 of the best predicting 40 words used to create the prediction formulae (*S&SRF* and the *h-fRF*) were included in the main study's item pool. Again two tests were created: a YN test with 96 words, plus 32 pseudowords including the nine best predicting pseudowords identified in Stubbe and Stewart (2012), with 23 more randomly selected from Tests 101–106 of the *EFL Vocabulary Tests* (Meara, 2010). An L2–L1 translation test, which contained the same 96 words, also randomly ordered, was also created.

Participants in the main study (Stubbe, 2013) took the YN test at the beginning of a class. As in the pilot, this was a paper test in which the students signaled whether they knew a word by filling in either a "Yes" bubble or a "No" bubble beside each item. The same students ($n = 455$) took the paper translation test towards the end of that same class in order. The YN test was scored by means of an optical scanner; the translation test was hand-marked by three native Japanese raters. Interrater reliability was 92%, and Facets analysis (Linacre, 2012) indicated that the raters were basically equal with overall measures of 0.02, 0.02, and –0.04 logits. Participants were all EFL students enrolled in one of four Japanese universities. About 40% of these participants had TOEIC scores in the 350–450 range, considerably higher than the pilot study range of 200–240.

3.3 Regression-Based Prediction Formulae

As discussed above, the improved prediction formula *S&SRF* was generated using the reduced 40-word and 9-pseudoword item set in the pilot study (Stubbe and Stewart, 2012). Two outliers with six FAs each out of a possible nine were deleted ($n = 69$ of 71) from the data. Again, the *S&SRF* formula (p. 6) is:

"True knowledge of tested words $= 3.26 + (0.51 \times \text{YN Score}) - (2.39 \times \text{False Alarms})$".

The formula for *h-fRF,* generated by running a simple regression analysis with the pilot study *h-f* adjusted YN scores as the independent variable and matching translation scores as the dependent variable ($n = 71$; using the reduced item set), was calculated to be:

$$\text{True knowledge of tested words} = 3.28 + (0.51 \times \textit{h-f}).$$

Both of these formulae were created using pilot study data only, to be applied to the main study YN test results. Also it is notable that only 16 of the 40 words as well as the 9 pseudowords used to create the two formulae above were also included in the main study's item set of 96 words and 32 pseudowords. Further, the ability levels of 40% of the participants in the main study were considerably higher than the pilot study participants.

4 Results and Discussion

Means and standard deviations (SDs) for the main study YN test as well as the translation test are presented in Table 2 (Stubbe, 2013). Similar to the pilot

Table 2. Summary of YN and Translation Test Results

Test	Mean	SD	Reliability
YN hits	48.82	17.23	0.96
YN FAs	2.17	3.16	n/a
Tr score	27.06	12.16	0.92

Note. Tr = translation test; SD = standard deviation; Reliability = Cronbach's alpha; $n = 455$; $k = 96$ real-words and 32 pseudowords on the YN test and 96 real-words on the translation test. The mean and SD figures reported in Stubbe (2013) were percentages, and are thus slightly higher.

study (Stubbe & Yokomitsu, 2012), YN test means were considerably higher than the translation mean (48.82 versus 27.06, respectively). This 44.6% decrease between YN and translation test scores is less than the nearly 50% decrease found in the pilot study, likely because the larger population in this main study (Stubbe, 2013) included English learners of higher proficiency. The reliability (Cronbach alpha) for these two tests was high at .96 and .92, respectively.

Means, SDs, correlations with the translation test scores (r), and residuals for the two *regression formulae* (RF) as well as *h-f* are presented in Table 3. Of the three formulae, *h-fRF* had the closest mean to the translation mean (24.85 and 27.06, respectively). Both *h-fRF* and *h-f* shared the highest correlation with translation scores (.833). Using Chen and Popovich's (2002) *t (difference)* formula for paired t-tests as adapted by Field (2009), the difference between the *h-fRF* correlation and *S&SRF* (.833 and .789, respectively) was statistically significant ($t = 6.14$, $df = 452$, $p < .0001$). The *h-fRF* formula also had the lowest of all residuals (RMSE) at 7.30. With the closest mean to translation scores, the highest correlation and smallest residual, *h-fRF* is clearly the best prediction formula.

A one-way analysis of variance revealed that the differences between the means of the translation scores and the two regression-based scoring formulae were statistically significant ($F(2, 1362) = 18.68$, $p < .0001$). Post hoc analysis revealed that the difference between the means of all three pairings was also statistically significant: (a) translation score with *h-fRF*; (b) translation scorewith *S&SRF*; and (c) *S&SRF* with *h-fRF* ($t = 6.69$, 11.70, and 14.83, respectively; df = 454; $p < .0001$, Bonferroni adjustment: .05/3 = .017). Effect sizes (Cohen's d) between translation scores and the two formulae were: *S&SRF* = .377; *h-fRF* = .209. As Cohen (1988) considered an effect size of .2 to be small and .5 to be medium, the difference between the translation scores and the *h-fRF* predicted scores was small. Although the difference between the means of the translation test and the *h-fRF* predictions was significant, the small effect size (.209) suggests that this formula may be able to predict recall knowledge reasonably well.

4.1 Proximity of Individual Predicted Scores

In a subsequent analysis, the score predicted by the two regression-based formulae, *S&SRF* and *h-fRF,* as well as *h-f*, were subtracted from the translation

Table 3. Means, SD, Correlations, and Residuals of Applying the Formulae: *S&SRF*, *h-fRF*, and *h-f*

Test/Formula	Mean	SD	*r*	Residual
Tr Score	27.06	12.16	1	n/a
YN hits	48.82	17.23	0.721	24.90
YN FAs	2.17	3.16	-0.142	n/a
S&SRF	22.96	9.39	0.789	8.54
h-fRF	24.85	8.43	0.833	7.29
h-f	42.29	16.53	0.833	17.90

Note. Tr = translation test; SD = standard deviation; *r* = correlation (Pearson Product-Moment) with Tr (translation test) scores. Residuals were calculated as per Table 1 Note, above ($N = 455$).

score for each of the 455 individual participants to evaluate the usefulness of the individual predictions. Table 4 displays the number of individual participants with predicted scores within 1 percentage point (.96 of 96 words), 5 percentage points (4.8 of 96 words), and 10 percentage points (9.6 of 96 words) of his/her actual translation score. With 54% of predicted scores within 5 percentage points of translation results and 81.5% within 10 percentage points, the *h-fRF* again appears to predict translation scores reasonably accurately. The efficacy of applying regression analysis to YN test results is clearly demonstrated by the *h-fRF* residuals, which are substantially lower than *h-f* (7.3 versus 17.9, see Table 3). Remembering that *h-fRF* and *h-f* share the same *r* value (.833), these results also demonstrate the necessity of calculating residuals, and not just correlations, when comparing YN scoring formulae.

A couple of important differences exist between the pilot study (Stubbe & Yokomitsu, 2012; from which *h-fRF* and *S&SRF* were developed) and the main study (Stubbe, 2013; upon which the prediction formulae were tested). Only Stubbe and Stewart's (2012) *reduced item set* of the best 40 words and 9 pseudowords was used to create the regression-based formulae. This *reduced item set* shared only 16 words and the nine pseudowords with the main study item set of 96 words and 32 pseudowords. As 103 of the total 128 items, in the main study YN item set (82.4%) were used only in the main study, it appears as if *h-fRF* is not overly item dependent. In other words, this formula seems to work well even when tests contain different vocabulary items. Another important difference is between the participants in the two studies in terms of sample size and English proficiency levels.

Table 4. Proximity of Predicted Scores to Actual Translation Scores

Formula	within 1%	within 5%	within 10%	outside 10%
S&SRF	44 (9.7%)	224 (49.2%)	353 (77.6%)	102 (22.4%)
h-fRF	58 (12.7%)	247 (54.3%)	373 (82.0%)	82 (18.0%)
h-f	5 (1.1%)	40 (8.8%)	122 (26.8%)	333 (73.2%)

Note. $n = 455$; 'within x%' denotes percentage points, i.e., 'within 5%' means within five percentage points of the total 96 items, i.e. 4.8 words.

Whereas *h-fRF* was based on 71 low level learners with TOEIC scores of about 200–240, the participants in the main study numbered 455 with TOEIC scores ranging from 200 through 450. Hence, *h-fRF* appears to work well with a wider range of proficiency levels.

5 Conclusion

This study was an investigation into predicting meaning recall (L2–L1) translation test scores from YN test real-word and pseudoword results. Two regression-based prediction formulae, developed using test results from the pilot study following the method described in Stubbe and Stewart's (2012) study, were compared with the established *correction for guessing* scoring formula *h-f*. Results suggest that the two regression-based formulae delivered better predictions. The *h-fRF* formula proved to be the overall best predictor. It was also found that *h-fRF* was not overly item dependent, nor ability level dependent.

This study has demonstrated the usefulness of a new YN scoring formula derived from a simple regression analysis of the *h-f* adjusted YN scores in one study and applied to the *h-f* adjusted YN scores in a different study. By first calculating *h-f* adjustments to YN test results and modifying those adjusted scores using a regression-based prediction formula, such as *h-fRF*, the prediction ability of the YN test can be substantially improved. These results have certain implications for EFL teachers. One trusted means of checking vocabulary knowledge is to test the students on a selection of words they will encounter in a language activity, or text, using a meaning recall translation test. However, the marking of this testing format can be quite cumbersome, especially with large numbers of students and/or items. The present study has found that by adjusting the results of a YN test of the same words, using a simple regression-based scoring formula, *h-fRF*, produced predicted scores that are within 10 percentage points of the actual translation test scores for 82% of participants. For example, an individual *h-fRF* predicted score of 60% of the total number of tested words has an 82% chance of falling within an actual knowledge range between 50% and 70% as demonstrated by a translation test, and a 54% chance of falling between 55% and 65%. Thus, teachers can have reasonable confidence in YN scores adjusted by *h-fRF*, while avoiding the drudgery of marking translation tests.

The *h-fRF* formula, "True knowledge of tested words $= 3.28 + (0.51 \times h\text{-}f)$", was derived from and consequently appears to work well with low-level Japanese EFL students. This formula may require re-calibration for learners of different ability levels, and/or cultures. By first giving students a YN test followed by a translation test of the same items, calculating the *h-f* adjusted YN scores, and then performing a simple regression analysis (with *h-f* adjusted YN scores as the independent variable and translation scores as the dependent variable), a revised *h-fRF* can be calculated. From the regression table, coefficients similar to "3.28" for the intercept and "0.51" for *h-f* will be provided. Using the revised formula, teachers should be able to replace translation tests with the easier YN test format.

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.

Barrow, J., Nakanishi, Y., & Ishino, H. (1999). Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System*, *27*(2), 223–247. doi:10.1016/S0346-251X(99)00018-4

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing*, *16*, 131–162. doi: 10.1177/026553229901600202

Chen, P., & Popovich, P. (2002). *Correlation: Parametric and non-parametric measures*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-139. Thousand Oaks, CA: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

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 (LandelijkeOnderzoekschoolTaalwetenschap).

Field, A. (2009). *Discovering statistics using SPSS (3rd ed.)*. London, UK: Sage Publications.

Gyllstad, H., Vilkaite, L., Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL International Journal of Applied Linguistics 166*(2), 276–303.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes–no vocabulary test: correction for guessing and response style. *Language Testing*, *19*(3), 227–245. doi:10.1191/0265532202lt229oa

JACET Basic Word Revision Committee. (2003). *JACET list of 8000 basic words*. Tokyo: Japan Association of College English Teachers.

Linacre, J. M. (2012). *Facets computer program for many-facet Rasch measurement, version 3.70.0*. Beaverton, Oregon: Winsteps.com. Retrieved from: http:// www.winsteps.com/index.htm

Meara, P. (1997). Towards a new approach to modelling vocabulary learning. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. *109–121*). Cambridge, UK: Cambridge University Press.

Meara, P. (2010). *EFL vocabulary tests*. Swansea: Lognostics second edition. Retrieved from: http://www.lognostics.co.uk/vlibrary/meara1992z.pdf

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142–154.

Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing*, *23*(1), 73–98. doi:10.1191/ 0265532206lt321oa

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

Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*(2), 282–308. doi:10.3138/cmlr.56.2.282

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355–371. doi:10.1177/026553229301000308

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511732942

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

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York: Palgrave Macmillan. doi: 10.1057/9780230293977

Schmitt, N., Jiang, X., & Grabe, W. (2011).The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. doi:10.1111/j.1540-4781.2011.01146.x

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*, 139–152. doi: 10.1080/09571730802389975

Stubbe, R. (2012a). Searching for an acceptable false alarm maximum. *Vocabulary Education & Research Bulletin*, *1*(2), 7–9.

Stubbe, R. (2012b). Do pseudoword false alarm rates and overestimation rates in YN vocabulary tests change with Japanese university students' English ability levels? *Language Testing*, *29*(4), 471–488. doi: 10.1177/0265532211433033

Stubbe, R. (2013). Comparing regression versus correction formula predictions of passive recall test scores from yes-no test results. *Vocabulary Learning and Instruction*, *2*(1), 39–46.

Stubbe, R., & Hoke, S. (2014). Comparing YN test correction formula predictions of passive recall test results. In R. Chartrand, G. Brooks, M. Porter, & M. Grogan (Eds.), *The 2013 PanSIG conference proceedings* (pp. 72–78). Nagoya, Japan: JALT.

Stubbe, R., & Stewart, J. (2012). Optimizing scoring formulae for YN vocabulary checklists using linear models. *Shiken Research Bulletin*, *16*(2), 2–7.

Stubbe, R., & Yokomitsu, H. (2012). English loanwords in Japanese and the JACET 8000. *Vocabulary Education & Research Bulletin*, *1*(1), 10–11.

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. Retrieved from: http://nflrc.hawaii.edu/rfl/October2003/waring/waring.pdf.