Optimizing scoring formulas for yes/no vocabulary tests with linear models

Raymond Stubbe
raymondstubbe@gmail.com
Kyushu Sangyo University

Jeffrey Stewart
jeffjrstewart@gmail.com
Kyushu Sangyo University, Cardiff University

Abstract

Yes/No tests offer an expedient method of testing learners' vocabulary knowledge, although a drawback of this method is that since the method is self-report, actual knowledge cannot be confirmed. "Pseudowords" have been used within such lists to test if learners are reporting knowledge of words they cannot possibly know, but it is unclear how to use this information to adjust scores. Although a variety of scoring formulas have been proposed in the literature, empirical research (e.g., Mochida & Harrington, 2006) has found little evidence of their efficacy.

The authors propose that a standard least squares model (multiple regression), in which the counts of words reported known and counts of pseudowords reported known are added as separate predictor variables, can be used to generate scoring formulas that have substantially higher predictive power. This is demonstrated on pilot data, and limitations of the method and goals of future research are discussed.

Background

Yes/No checklists of English vocabulary knowledge (YN tests) can be used as a quick, psychometrically reliable self-report method of checking students' knowledge of large numbers of words in a relatively short period of time (Meara & Buxton, 1987). However, an obvious pitfall of such an approach is that tests reliant on self-reporting may produce results that differ from reality. Indeed, recent research by Stubbe and Yokomitsu (2012) reveals that students who completed a YN test of English words and were then asked to provide Japanese definitions for words they had reported knowledge of could on average only provide correct definitions for half. In many cases, overestimation of vocabulary size was due to confusion of an unknown word (e.g., *root*) with a known, but untested word with similar orthography (e.g., *route*). Such a disparity between YN test scores and translation tests scores is not uncommon. Waring and Takaki (2003) reported a 70% drop on scores from a recognition test to a similar (L2 to L1) translation test.

Pseudowords (e.g. "steck" or "noof") are commonly used in YN tests to detect false reports (Read, 2000); if a student claims to know a word that does not actually exist, it calls their reports of knowledge on real words into question. SLA literature contains a number of different formulas for YN tests that use the number of pseudoword checks, called false alarms (FA), to adjust the proportion of words considered to be known. For example, Huibregtse, Admiraal and Meara (2002) studied the following four correction formulas: *h - f* (Anderson & Freebody, 1983), *cfg* (*correction for guessing*, Meara & Buxton, 1987), *Δm* (Meara, 1992) and *Isdt* (Huibregtse, et al., 2002).

However, it remains unclear which, if any, of these scoring formulas is preferable for improving the predictive power of YN tests. Mochida and Harrington (2006) proposed a research design for examining and comparing the efficacy of these various formulas:

- Give learners a self-report YN test.
- As a follow up, give them a conventional test of the same words.
- Score the self-report YN test by a variety of scoring formulas that use pseudoword false alarms to calculate the true number of words known, and see which formula results in the highest correlation to scores on the conventional test.

Mochida and Harrington examined the above formulas by correlating resulting YN test scores to scores on the Vocabulary Levels Test (Nation, 1990), but did not find substantial differences in correlations regardless of the adjustment formula used. Eight years following Huibregtse, et al., (2002), Schmitt noted that "it is still unclear how well the various adjustment formulas work" (Schmitt, 2010, p. 201). Possibly due to such uncertainties about the efficacy of these formulas, some researchers simply disregard individuals' YN test results when FAs exceed a given threshold. For example, Schmitt, Jiang and Grabe (2011) discarded any YN test forms reporting more than 3 false alarms (10% of the total of 30 pseudowords used in that study).

Can information gleaned from pseudowords be of use in adjusting YN test self-reports to better reflect the number of words a student is likely to know if they are tested on meanings of each word individually? In this research report, we will detail a statistical model built using experimental data that demonstrates that this is likely the case. Steps necessary for its creation can be described as follows:

- Follow Mochida and Harrington's (2006) research design, so that the resulting data set contains: a) a list of words students claim to know via a YN test; b) a count of pseudowords students claim to know (false alarms); and, c) a list of the same words that students actually know as determined by a conventional test.
- In addition to running a simple bivariate correlation between words self-reported as known (the predictor variable) and words actually known (the dependent variable), we can go a step further: Using multiple regression, add the count of pseudoword false alarms as a second predictor variable used to predict words actually known.

In addition to reporting an R^2 value indicating how well the two variables contained on the YN test predicted actual knowledge together, the model can also return a prediction expression, for example (Equation 1):

$$True\ number\ of\ words\ known = 8 + (0.7)(YN\ Score) - (2.4)(False\ Alarms) \qquad (1)$$

Here, 8 words is the intercept. For every word reported known on the YN test, we add 0.7 words truly known. For every false alarm, we subtract 2.4 words. Such a prediction expression is effectively an optimized "scoring formula", at least in regards to the sample of students and the particular YN test it was derived from. Although it is unlikely the formula will be optimal for another sample or YN test without modification, such models provide an empirical basis for the creation of scoring schemes. If the prediction expression can be shown to hold well over multiple samples, it can be used with confidence with this same YN test when testing comparable demographics.

This short paper will detail preliminary steps taken to develop such a scoring formula using data from one of the authors' previous studies (Stubbe & Yokomitsu, 2012), and propose methods for examining the validity of self-reports on individual words and the predictive power of individual pseudowords. Further development will be detailed in future studies.

Method

Participants were second year students enrolled in a compulsory English course at a private Japanese university, with *TOEIC® Bridge* scores ranging from 100 through 140. Unfortunately, two of the students reported 10 and 11 false alarms, each. As they would have considerable leverage on scoring formulas arrived at in this study they were removed, reducing the sample size to 69. Participants first took a YN test of 120 words with 32 pseudowords, and afterwards provided definitions for those same words in a translation test (L2 to L1). YN tests were marked using a scanner, while the translation test was hand marked by two native Japanese speakers, with an inter-rater reliability of .97

Preliminary analysis

Descriptive statistics for both the YN and translation tests are presented below.

Table 1. YN and Tr test descriptive statistics

Test	Mean	SD	Range	Low	High	Reliability
YN Items	59.56	15.49	71	20	91	0.93
YN FAs	0.96	1.56	7	0	7	0.65
Tr Items	29.94	9.48	38	11	49	0.89

Note: Reliability = Cronbach's alpha.

The correlation between the YN test form and the translation test was quite low, at only 0.60. To determine if a scoring formula using the experiment's pseudoword data could explain more variance, we ran a standard least squares model in the statistical software package JMP 8 (SAS, 2009) using false alarms and YN test scores as separate variables. Residuals were random and normally distributed, and there was not substantial collinearity between the two predictor variables (R = .23).

The mean response score was 30.55 (of the 120 items). The overall regression model was significant (F (2, 66) = 27.2, p < .0001), with an R^2 of 45.2%. The intercept of this model and both independent variables were significant (*p* = 0.0201 and ≤ 0.0011, respectively). Resulting variable weights (beta p) are presented in Table 2.

Table 2. Regression model variable weights and p-values

Variable	weight	p
Y-intercept	8.14	.0201
YN scores	0.41	.0001
FA scores	-1.94	.0011

R = .67

This suggests that a scoring formula that adjusts scores using false alarms counts could indeed be useful in estimating true vocabulary size. The optimal scoring formula for the YN test suggested by this initial model can be expressed as (Equation 2):

$$True\ number\ of\ words\ known = 8.14 + (0.41)(YN\ Score) - (1.94)(False\ Alarms) \quad (2)$$

However, while this represents a sizeable improvement on a prediction model that does not use pseudoword data ($R^2 = 35.6\%$, Table 6), the accuracy of prediction of vocabulary knowledge given self-reports on the YN test remains underwhelming. An R^2 of 45.2% is only equivalent to a correlation of about 0.67.

Improving predictive power with item analysis

To improve the predictive power of the YN test, we decided to look at specifically which words could be reliably predicted as known given self-reports on the YN test. We did this by examining phi (dichotomous) correlations between students' self-reports on given words and whether or not the same word was confirmed as known on the subsequent translation test.

The words in Table 3 had high phi correlations between self-report of knowledge and demonstrated knowledge on the translation test, meaning the self-report test appears to be a fairly valid predictor of actual knowledge of them.

The words in Table 4 had very *low* point phi correlations to the translation test results, meaning that there was little correlation between claims that these words were known and actual demonstrations of knowledge on the translation test.

Table 3. Real words with high phi correlations

Item	Phi correlation
salmon	0.57
chapel	0.56
crystal	0.51
narrow	0.51

Table 4. Real words with low phi correlations

Item	Phi correlation
maker	-0.11
concerto	-0.11
overall	-0.11
convenience	-0.23

It is easy to see why these words were falsely reported as known. For example, *convenience* was confused with the loan word *konbini* (convenience store), and *overall* was taken to have a literal meaning. This suggests that removing words that students have difficulty self-assessing knowledge on should improve the predictive power of the self-report test.

It is also possible to examine the efficacy of pseudowords used. Much in the same way language testers examine the point-biserial correlations of test items to overall test score to choose effective questions for a test, we can examine the degree to which pseudowords have *negative* point-biserial correlations with total scores on the translation tests in which students actually provide definitions of words reported as known. By doing this, we can design YN tests with pseudowords that have empirically validated predictive power.

The YN data was revised to include only the 40 words with the highest phi correlations to translation test results, and the nine pseudowords with the strongest negative point-biserial correlations to overall translation test scores. These nine point-biserials (on pseudowords such as *curify, lannery, noot,* and *skene*) ranged from -0.22 to -0.42, while the average for the other 23 pseudowords was -0.03 (essentially, no predictive power whatsoever). The regression model was re-run to see if this item analysis improved its predictive power. The mean response score became 13.07 (of the 40 items). The overall regression model was significant (F (2, 66) = 47.7, $p < .0001$), with an R^2 of 59.1% and the following variable weights (beta p):

Table 5. Revised regression model variable weights and p-values

Variable	weight	p
Y-intercept	3.26	.0214
YN scores	0.51	.0001
FA scores	-2.39	.0001
R = .77		

The R^2 was improved, and is now equivalent to a correlation of 0.77, and both the intercept (p = 0.0214) and the predictor variables ($p < 0.0001$) are significant. The scoring formula the model implies for this revised YN test is:

True knowledge of tested words = 3.26 + 0.51 × YN Score - 2.39 × False Alarms

The overall effectiveness of adding FA scores to the model can be seen below.

Table 6. R^2 values before and after entry of pseudoword predictor variable

Predictor variables	Original item list R^2	Revised item list R^2
YN scores only	35.6%	47.8%
YN scores + FA scores	45.2%	59.1%

Conclusions

This study has two findings of potential interest. First, much in the same way item analysis can aid in identifying items that perform well or poorly on a conventional test, they can assist in the creation of YN tests with greater predictive power. Self-reports of knowledge correlated to true knowledge more for some words than for others, and although most pseudowords used in this study had very little predictive power, a few had sufficiently negative correlations to true vocabulary knowledge to be of use in the model. For these reasons, we recommend that researchers subject the words and pseudowords they include on YN tests to item analysis, as is commonly done with more conventional test formats.

Second, although preliminary, this research indicates multiple regression may be of use in determining scoring formulas for self-report YN tests with pseudowords that can be empirically demonstrated to improve prediction of actual word knowledge, as measured by a separate test in which knowledge is confirmed by a human rater.

Of course, this finding is entirely preliminary. There are a number of remaining concerns. The formula must be tested on other samples to determine if it is generalizable to groups of learners other than the one examined for this particular experiment. Future directions will include finalizing such a model, and pitting it against existing scoring formulas. It should be noted that in all likelihood, any useful scoring formulas derived will only be applicable to the particular YN test used to develop it, and with demographics similar to the sample used in the research (i.e., Japanese university students). How generalizable such scoring formulas are to other populations has yet to be determined.

Finally, although item analysis and the scoring formula suggested by the model greatly improved the predictive power of the YN test, an R of 0.77 still seems fairly low. One possibility is that contrary to the findings of prior research, Japanese university students can overestimate their vocabulary knowledge, due to complicating factors such as English loanwords which have different usages in Japanese. Another possibility is that the sample examined in this data set gave less reliable self-reports due to their lower level of proficiency. For this reason further research must be conducted on learners with a wider range of proficiency.

References

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

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

Meara, P. 1992: *New approaches to testing vocabulary knowledge.* Draft paper. Swansea: Centre for Applied Language Studies, University College Swansea.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*(2), 142–154. doi:10.1177/026553228700400202

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

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

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

SAS Institute Inc. (2009). JMP Version 8. SAS Institute Inc., Cary, NC, 1989- 2009.

Schmitt, N. (2010) *Researching Vocabulary: A Vocabulary Research Manual.* NY: Palgrave Macmillan. doi:10.1057/9780230293977

Schmitt, N., Jiang, X., & Grabe, W. (2011) The percentage of words known in a text and reading comprehension. *Modern Language Journal, 95*, 26-43. doi:10.1111/j1540-4781.01146x

Stubbe, R. & Yokomitsu, H. (2012) English Loanwords in Japanese and the JACET 8000. *Vocabulary Education and Research Bulletin, 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.