*Vocabulary Learning and Instruction*
Volume 2, Issue 1, September 2013
http://vli-journal.org

# Comparing Regression versus Correction Formula Predictions of Passive Recall Test Scores from Yes–No Test Results

Raymond Stubbe
*Kyushu Sangyo University*
doi: http://dx.doi.org/10.7820/vli.v02.1.stubbe

## Abstract

A novel form of scoring formula for self-report yes–no vocabulary tests was presented in Stubbe and Stewart, based on multiple regression models that use both real-word and pseudoword reports to predict subsequent scores on a test of passive recall knowledge (as measured by L2 to L1 translations). The aim of the present study is to determine how well passive recall test scores can be predicted from yes-no test results adjusted using two methods: (1) *regression* formulas versus (2) the four established *correction for guessing* formulas outlined in Huibregtse, Admiraal, and Meara: *h-f*, *cfg*, Δm and *lsdt*. After taking a yes-no test followed by a passive recall test of the same 96 real-words, the sample of Japanese university students ($N = 431$) was split into two groups of comparable proficiency (A and B). The original Stubbe and Stewart regression formula was compared to the four *correction* formulas by analyzing their application with the Group A. Despite having a lower correlation with passive recall test scores than one of the *correction* formulas, the predicted scores produced were significantly closer. A new regression formula was then created using the Group A's test results and this was used to predict translation test scores on Group B, along with the four *correction* formulas. As the resulting predictions were superior to those of any of the *correction* formulas, and not significantly different from the actual passive recall test scores, plus the correlation with these translation test scores was the highest (0.845), it was concluded that regression formulas produced the best predictions.

## 1 Background

There are two approaches to utilizing pseudowords, or non-real-words, which are included in yes–no (YN) vocabulary tests as a means of checking for overestimation of lexical knowledge by test takers (Schmitt, 2010). The first is to adjust the YN scores, using a *correction* formula. The test results from learners checking pseudowords are adjusted using a variety of formulae, to better reflect their actual vocabulary size. Four such formulas were studied in Huibregtse, Admiraal, and Meara (2002): *h-f, cfg*, *Δm* and *Isdt.* Subsequently Schmitt (2010, p. 201) noted that "it is still unclear how well the various adjustment formulas work".

The other common use of pseudowords is to set a maximum acceptable number, beyond which "the data are discarded as unreliable" (Schmitt, 2010, p. 201). Schmitt, Jiang, and Grabe (2011) set their limit at three (10% of their 30

Please cite this article as: 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. doi: 10.7820/vli.v02.1.stubbe

pseudowords), as did Barrow, Nakanishi, and Ishino (1999). Stubbe (2012b) demonstrated that a cut-off point of four (12.5% of the pseudowords) better suited those YN test results. Stubbe and Stewart (2012) present a novel usage for YN test pseudoword data (also referred to as false alarms, FAs) – the creation of a standard least squares (multiple regression) model and formula which can be used to predict PR test scores using self-reports of lexical knowledge on a YN test. PR is the second most advanced of the four degrees of word knowledge studied in Laufer and Goldstein (2004), along with (in descending order of difficulty) active recall, active recognition and passive recognition. The final two degrees of word knowledge, active and passive recognition are normally tested using a multiple-choice format (for example, the Vocabulary Levels Test, Nation, 1990; Schmitt, Schmitt, & Clapham, 2001). Finally, the use of regression analysis with YN test results, though not that common, is not unprecedented in the literature (Mochida & Harrington, 2006).

## 2 Aim

The aim of this study is to determine which better adjusts yes-no test results to predict passive recall test scores: regression formulas as outlined in Stubbe and Stewart (2012) or *correction* formulas, such as the four considered in Huibregtse et al. (2003).

## 3 Method

### 3.1 Test preparation

For this study a yes-no test was prepared consisting of 96 real-words [six loanwords and six non-loanwords from each of the eight JACET 8000 (2003) levels] plus 32 pseudowords. The nine best pseudowords as well as the top 40 real-words identified in Stubbe and Stewart (2012), plus four more non-loanwords from the pilot to this study (Stubbe & Yokomitsu, 2012) were included in this item pool. The remaining 52 real-words were randomly selected as required from the various JACET 8000 (2003) levels, and the 23 additional pseudowords were randomly selected from *Tests 101–106* of the *EFL Vocabulary Tests* (Meara, 2010). Both real words and pseudowords were randomly distributed throughout the test. A passive recall test (L2 to L1 translations) of the same 96 real-words, all randomly distributed, was also created.

### 3.2 Testing procedure and sample

Participants took the yes-no test at the beginning of one class in July or August 2012. 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 455 students took a paper passive recall test towards the end of that same class in order to maximize test pairings. The yes-no test was scored by means of an optical scanner; the passive recall test was hand-marked by three native Japanese raters. Inter-rater reliability was 92%, and *Facets analysis* (Linacre, 2012) revealed

that the raters were basically equal with overall measures of 0.02, 0.02 and −0.04 logits. Participants were all students enrolled in one of four Japanese universities, with Test of English for International Communication (TOEIC®), Educational Testing Service (ETS), scores ranging from about 200 through 450. Of the 455 participants, 24 had checked more than eight pseudowords each, which is greater than two standard deviations (SDs) above the original false alarm mean of 2.17. These 24 outliers will not be included in this study as they could skew the results of any regression formula application (Tabachnick & Fidell, 2007), leaving a sample size of 431.

### 3.3 The RF and analysis procedure

Stubbe and Stewart (2012, p. 6) improved their original RF by utilizing item analysis to select 40 of the 120 real-words on the YN test which had the highest *phi* correlations to PR test results, and the 9 pseudowords with the highest negative-point biserial correlations to overall translation test scores. The resulting RF was reported as "True knowledge of tested words = 3.26 + (0.51 × YN Score) – (2.39 × FAs)", hereinafter referred to as the "S&S RF" (Stubbe & Stewart Regression Formula). This formula was based on a sample size of 69 of the original 71, as two outliers with high FA counts (also more than two SDs above the FA mean) were removed because they could have seriously skewed the any application of a RF. Not reported in that study (because of space limitations) was that the four *correction* formulas had correlations with translation test scores ranging from 0.676 through 0.768, while the S&S RF had a correlation of 0.769.

The results of applying the S&S RF and the four *correction* formulas will be compared using half of our data set by first sorting the 455 participants by PR test scores (highest to lowest), and then by FAs (lowest to highest), and removing the bottom 24 false alarm outliers (discussed above) leaving the sample size of 431. This dual sorting removed the outliers and should ensure the creation of two groups of comparable ability and honesty levels. Finally every second participant was assigned into Group A, the remainder in Group B ($n = 215$ and 216, respectively). All five formulas were used to predict PR test scores based on YN results (real-word and pseudoword) for Group A to determine the most useful formula. To determine if S&S RF can be improved upon, a revised RF was generated based on regression analysis of Group A's results. The new RF and four *correction* formula will be applied to Group B's results in a manner replicating the above, to determine the most useful prediction formula.

### 3.4 The four correction formulas

The four *correction for guessing* formulas presented in Huibregtse et al. (2002) and used in this study are *h–f*, *cfg*, *Δm* 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 increasingly more complicated *correction* formulas are all based on this hit rate and FA rate (*h* and *f*):

a) *cfg* (correction for guessing: Anderson & Freebody, 1983; Meara & Buxton, 1989),
b) *Δm* (Meara, 1992),
c) *Isdt* (Huibregtse et al., 2002).

## 4 Results and discussion

Means and standard deviations (SDs) for the YN test and the PR test results are presented in Table 1. The nearly 43% drop in item mean (47.94 versus 27.42, YN and PR tests, respectively) can be interpreted as students overestimating their lexical knowledge on the YN test. However, PR versus recognition of word forms and meanings offers another explanation. Considering this recall versus recognition dichotomy, Laufer and Goldstein (2004, p. 408) explained that "recalling a word's meaning or form can be considered a more advanced degree of knowledge than recognizing it". Additionally, Nation and Webb (2011) suggest that partial word knowledge may also contribute to such a gap.

Table 1. Summary of Yes–No and Translation Tests

| Test | Mean | SD | Range | Low | High | Reliability |
|---|---|---|---|---|---|---|
| YN Items | 47.94 | 16.96 | 78 | 5 | 83 | 0.95 |
| YN FAs | 1.58 | 1.76 | 8 | 0 | 8 | n/a |
| PR Items | 27.42 | 12.14 | 58 | 3 | 61 | 0.92 |

*Note.* Reliability = Cronbach's alpha; $n = 431$; $k =$ ;96 real-words and 32 pseudowords on the yes–no test and 96 real-words on the translation tests.

Table 2 reports the results of applying the S&S RF and four correction formulas to Group A's YN test results ($n = 215$). Similar to the full 431 participants, the drop between the recognition YN test results and the PR test results was about 43%. YN results correlated quite strongly with PR test scores (r = 0.783) but false alarm counts did not (r = −0.059), suggesting that FA rates did

Table 2. Means, SDs, Correlations and Residuals of Applying the Regression Formula and the Four Correction Formula to Group A ($n = 215$)

| | mean | SD | *r* | residual |
|---|---|---|---|---|
| PR Score | 27.44 | 12.00 | 1 | – |
| YN Score | 48.45 | 16.71 | 0.783 | – |
| FA Total | 1.57 | 1.75 | −0.059 | – |
| S&S RF | **24.36** | 8.46 | 0.825 | **7.65** |
| *h-f* | 43.73 | 16.07 | **0.834** | 18.60 |
| *cfg* | 46.10 | 16.99 | 0.817 | 21.14 |
| *Δm* | 36.68 | 22.09 | 0.770 | 17.52 |
| *Isdt* | 51.73 | 12.11 | 0.789 | 25.52 |

*Note.* *r* = correlation (Pearson Product-Moment) with translation test scores. Residuals were calculated by squaring the differences between each PR test score and each of the five predictions, summing those squares, calculating the mean and finally acquiring the square root. Closest mean (to PR score mean), highest correlation and smallest residual are in bold.

not substantially increase as ability level decreased, a finding shared by Stubbe (2012a). The highest correlation with PR test scores was the *h–f correction* formula, with S&S RF placing second (0.834 and 0.825, respectively). Using Chen and Popovich's (2002, p. 23) *t* (*difference*) formula for paired *t*-tests as adapted by Field (2009, p. 192), the difference between the S&S RF and the *h–f* RF correlations were found to be not statistically significant ($t = 1.270$, df = 212, $p = 0.206$, two-tailed). Based solely on these correlations with translation test scores, it appears as if the *h–f* formula was a slightly better predictor than the S&S RF, and significantly better than any of the other three *correction* formulas (for example, the *h–f* and *cfg* correlations were significantly different ($t = 2.651$, $df = 212$, $p = 0.0086$, two-tailed).

The means and residuals in Table 2 suggest a totally different interpretation, however. The S&S RF mean of 24.36 when compared to the translation test mean of 27.64 suggests that any prediction based on S&S RF would be much closer than one based on any of the *correction* formulas. The residuals of 7.65 versus the higher residuals for the *correction* formulas (17.52–25.52) confirm this observation. The significance of the difference between the S&S RF mean and the mean of the closest *correction* formula, $\Delta m$ (23.89 and 36.68, respectively) was confirmed using a paired *t*-test ($t = 3.461$, $df = 214$, $p = 0.0006$, two-tailed). The effect size (Cohen, 1988) between the two means was almost large ($d = 0.765$; $r = 0.357$, respectively). As only 51 of the 128 real-words and pseudowords included in this YN test originated from Stubbe and Stewart's (2012) reduced item set, plus the sample was greatly increased in terms of numbers and English proficiency levels, the correlation of 0.825 and much closer mean and lower residual (compared to any of the four *correction* formulas) suggest that the S&S RF appears to predict translation test results reasonably well.

Group A was then analyzed using multiple regression analysis to create a revised regression formula (New RF). This was created by including the YN real-word scores and false alarm counts as the independent variables in a multiple regression analysis, with PR test scores as the dependent variable. Residuals appeared randomly distributed, and there was not substantial collinearity between the two predictor variables ($r = 0.277$). The new RF, which was used to predict Group B's PR test scores, became: "passive recall knowledge of tested words = $0.536 + (0.622 \times \text{YN}) - (2.044 \times \text{FA})$".

Group B appears to be a little weaker than Group A with lower YN and PR test means, while the FA mean remained constant. The correlations with PR test score are lower for the YN results, but higher for the FA counts (see Table 3). The New RF had the highest correlation, closest mean to the PR test scores as well as the smallest residual compared to any of the *correction* formulas. Using the *t* (*difference*) formula for paired *t*-tests again, the difference between the New RF and *h–f* correlations (0.845 and 0.842) was found to be statistically significant ($t = 2.382$, $df = 213$, $p = 0.0181$) because the collinearity between the two correlations approached 1 ($r = 0.999526$).

As the New RF and PR test means were so close (26.82 and 27.21, respectively), it was decided to run a one-way ANOVA comparing the Translation test scores, the New RF predicted scores and the lowest of the *correction* formula

Table 3. Means, SDs, Correlations and Residuals of Applying the Regression Formula and the Four Correction Formulas to Group B ($n = 216$)

| | Mean | SD | *r* | Residual |
|---|---|---|---|---|
| PR Score | 27.21 | 12.53 | 1 | – |
| YN Score | 47.63 | 17.56 | 0.769 | – |
| FA Total | 1.57 | 1.75 | −0.086 | – |
| New RF | **26.82** | 10.17 | **0.845** | **6.61** |
| *h–f* | 42.70 | 16.35 | 0.842 | 17.70 |
| *Cfg* | 45.05 | 17.41 | 0.816 | 20.41 |
| *Δm* | 35.05 | 23.78 | 0.780 | 17.68 |
| *Isdt* | 50.89 | 12.49 | 0.816 | 24.66 |

*Note.* $r$ = correlation (Pearson Product-Moment) with translation test scores. Residuals were calculated by squaring the differences between each PR test score and each of the five predictions, summing those squares, calculating the mean and finally acquiring the square root. Closest mean (to PR score mean), highest correlation and smallest residual are in bold.

predicted scores, *Δm*. Significant differences were found between the three [$F(2,642) = 16.51$, $p < 0.0001$]. Post hoc analysis (Tukey's HSD) revealed significant differences between the PR test scores and the *Δm* predicted scores, as well as between the New RF predicted scores and those of *Δm* ($t = 5.347$ and 5.302, $p < 0.0001$, respectively). The significance level for the tests was set at $alpha = 0.0397$, using a partial Bonferroni adjustment for multiple comparisons (3) on correlated scores ($r = 0.789$) (Uitenbroek, 1997). The Cohen's $d$ effect size between the two prediction formula means was small ($d = 0.450$, $r = 0.220$). However, a significant differences was not found between the translation test scores and the New RF predicted scores ($t = 1.317$, $p = 0.189$, in the same post hoc analysis), and the Cohen's $d$ effect size between these two means was negligible ($d = 0.034$, $r = 0.017$). These findings support the suggestion that the New RF is a significantly stronger predictor of PR scores than any of the *correction* formulas.

## 5 Conclusion

The ability to grade yes-no tests using an optical scanner makes them considerably more convenient than traditional translation tests, which must be hand-marked by the teacher or researcher. Developing the ability to reliably predict passive recall ability based upon yes-no test scores could be a valuable contribution to the field of lexical development. The correlations arrived at in this study (0.825–0.845) and especially the closeness of the new regression formula's predicted scores to the actual translation test scores in the second analysis (Group B) suggest that using yes-no test real-word and pseudoword results to predict passive recall ability holds promise. The improved prediction ability of the New RF ($R^2 = 70.56\%$) over the original S&S formula ($R^2 = 68.56\%$) confirms that testing similar populations with items similar to those from which the RF(s) were derived also appears to be important for predictions. It may also suggest that caution should be exercised when using this approach with different test items or sample populations.

Directions for future research include improving the YN test items using the item analysis approach outlined in Stubbe and Stewart (2012) and to continue testing and refining the prediction formulas with other samples. Regression analysis could also be used to determine whether the optimal the cut-off point for false alarm outliers is the *greater than two SDs above the false alarm mean* utilized in this study.

## Acknowledgments

I would like to thank Shintaro Hoke and the staff of the Language Education and Testing Center (LERC) of Kyushu Sangyo University for marking the translation tests. I also thank and acknowledge Jeffrey Stewart for his original idea of applying regression analysis to YN real-word and pseudoword results for predicting receptive vocabulary knowledge, as well as for his invaluable editing of this paper.

## 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.

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

Chen, P., & Popovich, P. (2002). Correlation: Parametric and non-parametric measures. In Lewis-Beck, M. (Ed.), *Sage university papers series on quantitative applications in the social sciences* (pp. 7–139). Thousand Oaks, CA: Sage.

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

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

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

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

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

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

Meara, P. (2010). *EFL vocabulary tests. Swansea: _lognostics second edition 2010*. 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). Yes/No test as a measure of receptive vocabulary. *Language Testing, 23*, 73–98. doi:10.1191/0265532206lt321oa

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

Nation, I.S.P., & Webb, S. (2011). *Researching and analyzing vocabulary.* Boston, MA: Heinle.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* New York, 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/j.1540-4781.2011.01146.x

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the vocabulary levels test. *Language Testing*, *18*, 55–88. doi:10.1177/026553220101800103

Stubbe, R. (2012a). Do pseudowords false alarm rates and overestimation rates in Yes/No vocabulary tests change with Japanese university students' English ability levels? *Language Testing*, *29* (4), 471–488. doi:10.1177/0265532211433033

Stubbe, R. (2012b). Searching for an acceptable false alarm maximum. *Vocabulary Education & Research Bulletin, 1* (2), 7–9. Retrieved from http://jaltvocab.weebly.com/uploads/3/3/4/0/3340830/verb-vol1.2.pdf

Stubbe, R., & Stewart, J. (2012). Optimizing scoring formulas for yes/no vocabulary checklists using linear models. *Shiken Research Bulletin, 16* (2), 2–7. Retrieved from http://teval.jalt.org/node/12

Stubbe, R., & Yokomitsu, H. (2012). English loanwords in Japanese and the JACET 8000. *Vocabulary Education & Research Bulletin, 1* (1), 10–11. Retrieved from http://jaltvocab.weebly.com/uploads/3/3/4/0/3340830/verb-vol1.1.pdf

Tabachnick, B., & Fidell, L. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education Inc.

Uitenbroek, D.G. (1997). SISA binomial. D.G. Uitenbroek. Retrieved from http://www.quantitativeskills.com/sisa/calculations/bonfer.htm