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

# Second-Language Vocabulary Assessment Research: Issues and Challenges

Yo In'nami
*Shibaura Institute of Technology*
doi: http://dx.doi.org/10.7820/vli.v02.1.innami

## Abstract

The four papers on second-language vocabulary assessment reviewed below are exemplary works that merit close scrutiny. Therefore, this paper provides a brief summary of each study, followed by comments and suggestions, particularly in regard to the experimental designs and analyses used in the studies.

## 1 Introduction

I am honored to have the opportunity to closely read the four papers appearing in this issue, which each represent high-quality research in a second-language (L2) vocabulary assessment. Each of the papers makes an important contribution to the field. In what follows, I offer constructive criticism and suggestions regarding the experimental designs and analyses used in the studies.

## 2 Regression approach by Stubbe

Stubbe compares the relative effectiveness of regression methods versus correction formulas when applied to scores on a yes–no vocabulary test in order to predict scores on a passive L2-to-L1 recall test. The yes–no test included 96 real words and 32 pseudowords. The recall test measured L2-to-L1 translations of the same 96 real words. The two tests were administered to the same learners. The results show that the recall scores were best predicted by the regression-based yes–no scores for the real words and pseudowords (the latter of which are called false alarms), with an $R^2$ variance reaching 71%, when compared with the four correction formulas previously discussed in the literature. Stubbe concludes by advocating a regression approach to predicting a learner's passive recall of vocabulary from yes–no vocabulary tests.

The role of pseudowords in L2 has been researched since Meara and Buxton (1987), and Stubbe's current paper, along with Stubbe (2012a), is among the most recent works on this topic. In particular, four points regarding Stubbe's paper merit discussion. First, one strength of the study is that the sample size was large enough to split the data into two groups: A and B. Group A was first analyzed using the regression formula reported in Stubbe and Stewart (2012) and comparing those results with the application of the four correction formulas. Group A was then used to develop a new regression formula, and Group B was used to test whether the revised regression formula would work even better. This was a sensible approach given the relatively large sample size ($N = 431$) (see In'nami & Koizumi, 2012, for

Please cite this article as: In'nami, Y. (2013). Second-language vocabulary assessment research: Issues and challenges. *Vocabulary Learning and Instruction*, *2* (1), 74–81. doi: 10.7820/vli.v02.1.innami

such an approach). One plausible weakness – although not uniquely inherent in the study and not intended as a criticism of Stubbe's paper in particular – is that since the two groups were similar (at least in terms of the passive recall test scores), the high applicability of the new regression formula with Group B is not surprising. After all, Groups A and B were similar in many ways; thus, if the regression formula was derived from Group A, it could be predicted to work well with Group B. In other words, the study has high internal, ecological validity, but it does not guarantee a high generalizability of the finding beyond the present context. It would certainly be of great interest to investigate to what extent the current regression formula works with different samples.

A related issue in the comparison is that Stubbe and Stewart (2012) examined the proportion of variance in their translation test score explained by the yes–no test scores and false alarm scores, while retaining all items or only those with high discriminability. As Table 1 shows, prediction was most successful with the yes–no scores and false alarm scores were used as predictors when only high discriminatory items were included. As I understand, Stubbe's current regression includes the yes–no scores and false alarm scores as predictors while using all items. The predicted percentage of 70.56% in the study is higher than the figure of 45.2% in Stubbe and Stewart (2012). The prediction would have been even higher if only the high discriminatory items had been included in the analysis. This points to the effectiveness of regression-based approaches with high discriminatory items.

Second, an equally important issue is to conduct an item analysis both quantitatively and qualitatively on pseudowords to identify whether any particular kinds of pseudowords are more attractive to learners. The quantitative analysis may include an examination of item difficulty and discrimination, while the qualitative analysis may include an examination of the test-taking processes. As yes–no test formats simply require learners to signal dichotomously whether they know or do not know a particular word, the reasons for their responses remain unclear. To resolve this issue, we could ask them to think aloud their answering processes (e.g., Bowles, 2010; Ericsson & Simon, 1993) while engaging in the yes–no test. If collecting think-aloud data is not possible due to logistic constraints, we could instead create several small sets of pseudowords, randomly assign learners any one of these sets, and ask them to write the reasons for their judging the words as pseudowords. Alternatively, we could interview them about their responses. Stubbe and Stewart (2012) compared learners' responses to the same real words appearing in a yes–no test and translation test. Although some words (e.g., *salmon* and *chapel*) were correctly reported as known across the tests, others (e.g., *convenience* and *overall*) were falsely reported as known. According to the authors, this could be due to factors such as the loanword *konbini* "convenience store" and the multiple

Table 1. $R^2$ Values Before and After Entry of the Pseudoword Predictor Variable in Stubbe and Stewart (2012)

| Predictor variables | $R^2$ (all items) | $R^2$ (discriminating items only) |
|---|---|---|
| Yes–no scores only | 35.6% | 47.8% |
| Yes–no scores + false alarm scores | 45.2% | 59.1% |

meanings of the word *overall* (including everything; a piece of clothing), suggesting that the efficacy of pseudowords may change depending on learners' L1 and the number of dimensions of a word's meaning. These effects can be more deeply examined through an analysis of think-aloud protocols. The results will provide insight into the structure of learners' vocabulary knowledge and into a more sound construction and inclusion of pseudowords for yes–no tests while adjusting the efficacy of pseudowords as distractors.

Third, with regard to the use of pseudowords as a measure of false alarms, Stubbe (2012b) examined the score divergence between (1) a yes–no receptive vocabulary test and (2) a translation test of the same items. He reported that the cut-off point for false alarm in yes–no tests should be set at four (i.e., 12.5% of the 32 pseudowords in the study). In other words, if a learner reports knowledge of five or more pseudowords, s/he may be falsely claiming (i.e., overestimating) knowledge of real words. Impressive as setting the cut-off is, it is not clear why the cut-off is not used in Stubbe's current study. This may be due to the particular nature of the sample on which the cut-off is based (e.g., small sample size and test items). If this is the case, further research with different samples is warranted.

Finally, another issue regarding the use of pseudowords is the number of pseudowords in a yes–no test and its percentage of the total number of items. Stubbe (this issue) includes 96 real words and 32 pseudowords (3:1 ratio). Both Stubbe and Stewart (2012) and Stubbe (2012b) included 120 real words and 32 pseudowords (3.75:1). Given the central and ubiquitous role of vocabulary in almost all aspects of language learning, the representativeness of words taken randomly from the domain to which study findings are expected to generalize, and the relative easiness of administering a vocabulary test, few would argue against including, for example, over 100 real words in a test. However, things are less clear for pseudowords because, unlike real words, they are not part of language learning and do not have a population to which they can be generalized. It remains unknown how many and what kind of pseudowords should be included in a test, according to my reading of recent books such as Nation and Webb (2011). Schmitt (2010) argued that pseudowords usually account for 25–33% of the total items (p. 200), but this percentage may change depending on the aforementioned efficacy of pseudowords and the cut-off point for false alarms (the second and third points above). The use of yes–no tests with pseudowords as a proxy of passive recall tests is valuable and merits further research.

## 3. Differential item functioning (DIF) analysis by Stoeckel and Bennett

Stoeckel and Bennett attempt to clarify the sources of DIF across learners of different L1 backgrounds on a 90-item, multiple-choice, receptive written test of L2 English vocabulary. The test consisted of 30 randomly sampled words from each of the first and second 1,000 words of the General Service List and another 30 from the Academic Word List. The test was administered to Japanese and Korean university students. A total of 21 of the 90 items (23%) were flagged as DIF; 10 were considered to display DIF due to their word frequency and usage as loanwords in each language – particularly, phonological and/or orthographic

similarities between the word and an English loanword in the L1 – and also due to the function of distractors. The authors conclude that the word frequency and usage as a loanword in the L1 are responsible for DIF, in addition to cognate status, as previously reported.

This study is a good contribution to the DIF literature, particularly given, as the authors claim, the availability of only a small number of L2 studies investigating sources of DIF in cross-cultural settings (for a recent review of DIF, see Ferne & Rupp, 2007; McNamara & Roever, 2006: Chapter 4). While the word frequency and usage as a loanword in the L1 can be causes of DIF for the 10 words, the causes of DIF for the other 11 words remain unknown. These words include *interval* and *rise*, favoring Japanese examinees, and *contemporary* and *compound*, favoring Korean examinees. These DIF items do not seem to be related to their frequency or usage as loanwords in the L1 or to cognate status, and an attempt to identify explanatory variables, for example, by using a think-aloud protocol or expert content analysis, is a good avenue for future research. In this regard, in conducting research to identify sources of DIF, Ercikan, et al. (2010) recommended, for example, (1) examining both content expert reviews and think-aloud protocols and (2) examining both concurrent and retrospective verbalizations.

## 4. L2 word recognition analysis by Coulson, Ariiso, Kojima, and Tanaka.

Coulson et al. seek to identify factors related to reading difficulty among Japanese learners of English as a foreign language (EFL), particularly focusing on L2 word recognition. Their test battery consisted of three tests: (1) 120 high-frequency real words mixed with 40 pronounceable nonwords, (2) 50 pseudo-homophones mixed with 150 pronounceable nonwords, and (3) a 10-item spooning test. The results show that the Basic (low) Class performed significantly worse than the Advanced Class with the 50-item pseudo-homophones, with an average number of mistakes of 19.7 and 11.1, respectively, and standard deviations of 11.1 and 2.2. Although statistical significance did not seem to be found in the remaining tasks, the Basic Class consistently scored lower, except with the 150 pronounceable nonwords, where the Basic Class scored slightly higher. Coulson et al. conclude that low-proficiency learners lack word-recognition ability but may be able to overcome this inability through extensive reading.

Coulson et al.'s study is unique in that it methodologically replicates Wydell and Kondo (2003). That is, it develops and uses tests designed to measure the same constructs studied by Wydell and Kondo. The results are similar across the studies: the Basic Class consistently scored lower than the Advanced Class in almost all tasks in Coulson et al., and the Japanese EFL students consistently scored lower than the native speakers of English in all tasks in Wydell and Kondo. The stability of the results across the two studies suggests the high quality of the tests developed in Coulson et al. and the robust effects of a lack of L2 word recognition for Japanese EFL learners. In particular, Coulson et al.'s replication can be viewed as an approximate (or systematic or operational) replication consisting of an exact duplication of some of the essential variables (e.g., experimental procedures) of the original study (for details see, e.g., Lykken, 1968). Approximate replication seeks to

examine whether one can duplicate a result using the same/similar methods reported in the original study. The results show how dependent a finding is on the particular research conditions. Given the long-standing paucity of replication studies in the field of language learning (Porte, 2012), Coulson et al.'s study is highly welcome.

As Coulson et al. discuss in the last paragraph of their paper, it would be of great interest to know more about the backgrounds of the five students who underperformed on the pseudohomophone task. For example, did they have any experience doing extensive reading in their prior English language learning? Assuming they went through Japanese secondary education, where extensive reading is not common, we could expect them to develop their word-recognition ability with the aid of extensive reading. If these students had done extensive reading and yet still underperformed on the pseudohomophone task, this suggests the need to examine the type of extensive reading they did, including the amount of exposure (e.g., number of books/words they read, book levels), duration, use of comprehension tasks, and instructor support availability.

Coulson et al. seem to assume that extensive reading helps learners develop word-recognition ability. However, this assumption itself deserves close empirical investigation. One recent – and perhaps the most comprehensive – study is Nakanishi's (2013) dissertation on a meta-analysis of extensive reading. It reports on a meta-analysis of 34 primary studies that provided 43 unique effect sizes with a total sample size of 3,942. Nakanishi finds that extensive reading improved university students' overall performance (Cohen's $d = 0.48$ [95% confidence interval = 0.22, 0.74] for between-group designs and 1.12 [0.23, 2.01] for pre-post designs). While studies on word recognition in relation to extensive reading seem to be under-researched and were not included in the meta-analysis, reading speed was found to improve for between-group designs (0.98 [0.64, 1.33]) but not for pre-post designs (0.61 [–2.79, 4.02]). The statistically nonsignificant effect for pre-post designs may be partly due to test–retest effects: The learners improved or felt demotivated by their second exposure to the same instrument. Also, the pre- and post-tests were not of equal difficulty. As Nakanishi describes, it was not possible to meta-analyze the effect of extensive reading on university students' reading speed, because further subdivision of the studies would have reduced the number of the studies in the meta-analysis, yielding unstable results. However, the tentative finding from the between-group designs described above shows that extensive reading leads to progress in reading speed. Since reading speed is considered closely related to word recognition, it is possible that extensive reading leads to improvement in word recognition. This speculation needs to be empirically studied. Nakanishi lists the individual studies included in the meta-analysis, with information on moderated variables coded. The list shows which studies measured the reading speed of university students after extensive reading. These studies merit close scrutiny.

## 5. Item response theory (IRT) analysis by Tseng

Tseng aims to develop a multiple-choice pictorial vocabulary test for primary and junior high school Taiwanese students based on the 1,200 word list compiled

by the Taiwanese Ministry of Education. A total of 180 words was selected from the list, subdivided into two forms (90 items each), and administered to students as vocabulary test items. The results from the IRT analysis showed that the three-parameter model, which includes item difficulty, item discrimination, and guessing parameters, best fit the data. A formula was constructed to convert the parameter estimates into vocabulary size estimates. Tseng concludes that developing a vocabulary test using IRT is theoretically (i.e., psychometrically) and pedagogically meaningful: Theoretically, it can benefit from the advantages of IRT over classical test theory, and pedagogically, it can bring good washback effects on vocabulary learning by helping diagnose learners' strengths and weaknesses.

The line of studies connecting assessment and learning is timely and essential, and, in this regard, Tseng's research deserves great attention. To best benefit from Tseng's findings, we should consider two important issues. First, given Tseng's dual focus on assessment and learning, it will be necessary to provide fine-grained pedagogical information on vocabulary learning in the Taiwanese context. For example, learners and teachers would be interested to know not only individual learners' vocabulary sizes but also the average vocabulary size of students in each school grade (e.g., up to 500 words by Grade 5), longitudinal change in vocabulary size, and feedback on how to learn unmastered words and expand vocabulary size. While having a better understanding of one's vocabulary size would enhance learning, this can only happen if we help students and teachers make the best use of the information from vocabulary tests. In many cases, it is highly likely that students and teachers merely glance at score reports without reflecting on the various implications. This may be because the score reports contain too much or too little information that is useful for learning. Therefore, educators should think carefully about designing score reports (see Kunnan & Jang, 2009; Roberts & Gierl, 2010) and also examine how students and teachers use them.

Second, related to the first point and particularly unique to Tseng's study is the need to consider how to conduct a validation study. In his elucidating commentary on Kane's (2013) argument-based approach to validation, Brennan (2013) summarized that Kane's approach consists of (1) specifying the intended purpose of a test and (2) evaluating empirical evidence and logical arguments. To the extent that the intended purpose of the test is clearer, this should be explicitly stated. For example, following this framework and the aims of Tseng's study, we can articulate that the pictorial vocabulary size test in Tseng's study was developed to measure and diagnose primary and junior high school Taiwanese students' vocabulary size in accordance with the 1,200 high-frequency word list compiled by the Ministry of Education. Since the list was created for textbook publishers to follow as guidelines on developing textbooks for primary and junior high school Taiwanese students, we can also articulate that the list (and the test results) is expected to assist textbook publishers in creating textbooks.

Brennan's (2013) second phase calls for conducting validation studies in accordance with various types of inferences: inferences consist of scoring, generalization, extrapolation, and decision rules. The reliability and item analysis as currently reported in Tseng belong to scoring. We also need to collect validity evidence for: a) generalization of the scores obtained to the well-defined target domain (for example by conducting a generalizability theory analysis of the items

and the test versions); b) extrapolation to a wider and less well-defined target domain (by correlating the scores to other criterion measures, for example), and c) decision rules (i.e., test uses and consequences, for instance by examining whether diagnostic feedback benefits learners and teachers as intended and whether the word list [and the test results] benefits textbook publishers). Washback effects may or may not come in intended forms (e.g., Cheng, Watanabe, & Curtis, 2004). These pieces of evidence are all needed to argue for the validity of interpretation and use based on Tseng's test scores.

## 6. Conclusion

While the four studies reviewed above have different focuses and purposes, they each make an important contribution to the literature on vocabulary assessment. Future research based on these works will allow us to gain a more detailed and sophisticated understanding of L2 vocabulary and the appropriate methods of assessing it.

## Acknowledgement

I would like to thank Rie Koizumi and Raymond Stubbe for their valuable comments on earlier versions of this paper.

## References

Bowles, M.A. (2010). *The think-aloud controversy in second language research*. New York, NY: Routledge.

Brennan, R.L. (2013). Commentary on "Validating the interpretations and uses of test scores." *Journal of Educational Measurement*, *50*, 74–83. doi:10.1111/jedm.12001

Cheng, L., Watanabe, Y.J., & Curtis, A. (2004). *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Erlbaum.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2010). Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice*, *29*, 24–35. doi:10.1111/j.1745-3992.2010.00173.x

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Ferne, T., & Rupp, A.A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*, 113–148.

In'nami, Y., & Koizumi, R. (2012). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing*, *29*, 131–152. doi:10.1177/0265532211413444

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. doi:10.1111/jedm.12000

Kunnan, A.J., & Jang, E.E. (2009). Diagnostic feedback in language assessment. In M.H. Long., & C.J. Doughty (Eds.), *The handbook of language teaching* (pp. 610–627). Oxford, UK: Wiley-Blackwell. doi:10.1002/9781444315783.ch32

Lykken, D.T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151–159. doi:10.1037/h0026141

McNamara, T., & Roever, C. (2006). *Language Testing: The Social Dimension*. Malden, MA: Blackwell.

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

Nakanishi, T. (2013). A meta-analysis of extensive reading research.

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

Porte, G. (Ed.). (2012). *Replication research in applied linguistics*. Cambridge, UK: Cambridge University Press.

Roberts, M.R., & Gierl, M.J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice*, *29*, 25–38. doi:10.1111/j.1745-3992.2010.00181.x

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Hampshire, UK: Palgrave Macmillan.

Stubbe, R. (2012a). 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. (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

Wydell, T.N., & Kondo, T. (2003). Phonological deficit and the reliance on orthographic approximation for reading: A follow-up study on an English-Japanese bilingual with monolingual dyslexia. *Journal of Research in Reading, 26*, 33–48. doi:10.1111/1467-9817.261004