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

Sources of Differential Item Functioning between Korean and Japanese Examinees on a Second-Language Vocabulary Test

Tim Stoeckel and Phil Bennett
Miyazaki International College

doi: http://dx.doi.org/10.7820/vli.v02.1.stoeckel.bennett

Abstract

The use of item response theory in equating or creating computer-adaptive tests relies on the assumption of invariance of item parameters across populations. This assumption can be assessed with an analysis of differential item functioning (DIF). The purpose of this study was (a) to ascertain whether DIF between two native language groups was present on a 90-item multiple-choice English vocabulary test and (b) to explore the causes of DIF, should it exist. Participants were 184 Korean and 146 Japanese undergraduate students learning English as a foreign language in their home countries. A separate calibration *t*-test approach was used to identify DIF, with the criteria set at $p < 0.01$ and effect size >1 logit, calculated as the difference in Rasch item-difficulty between the two groups. Twenty-one items displayed DIF. The causes of DIF in nine of those items were tentatively identified as relating to their status as loanwords in the L1. When a tested word was a loanword in both Korean and Japanese, differences in both the frequency and range of use of the loanword in the two languages predicted the direction of DIF. Similarly, phonological/orthographic overlap between two separate English loanwords in the L1 was found to be a possible cause of DIF. Implications for test development and further research in this area are discussed.

1 Background

In the area of second-language vocabulary test development, item response theory (IRT) has been used in attempts to create equivalent forms of well-known tests such as the Vocabulary Levels Test (VLT) (Beglar & Hunt, 1999) and for the calibration of item banks for computer-adaptive tests such as the Lexxica Word Engine (Browne, 2008). In both of these IRT applications, an underlying assumption is the invariance of item parameters. That is, the difficulty of test items should remain stable in relation to other items for any sample of the intended test population (Hambleton & Jones, 1993). It is this assumption that enables test developers to balance equivalent test forms with a range of items of similar difficulties, and in computer-adaptive tests, it enables statements regarding the probability of correctly answering un-administered test items once a person's ability is estimated.

The degree to which test items meet the IRT assumption of invariance can be assessed with an analysis of differential item functioning (DIF). DIF is present when two or more groups of examinees perform differently on a test item even after

Please cite this article as: Stoeckel, T., & Bennett, P. (2013). Sources of differential item functioning between Korean and Japanese examinees on a second-language vocabulary test. *Vocabulary Learning and Instruction*, *2* (1), 47–55. doi: 10.7820/vli.v02.1.stoeckel.bennett

differences in ability have been accounted for (Zumbo, 1999). When DIF analysis identifies such items, the source of DIF is investigated, and items displaying DIF thought to be caused by true differences in the latent construct (construct-relevant causes) can be retained while those with clear sources of bias such as differences resulting from gender, race, or ethnicity are either discarded or revised (Zieky, 2006).

Though there is a growing body of literature describing DIF analyses in second-language testing (e.g., Abbott, 2007; Pae, 2012; Park, 2008), few studies have investigated DIF between different native language groups in ESL vocabulary testing, and to our knowledge only two of those explored the causes of DIF when it was detected. Chen and Henning (1985) and Sasaki (1991) both compared native speakers of Spanish and Chinese and found that items testing English words which were cognates in Spanish were more difficult for the Chinese group.

The existence of cognates (words with a common etymological origin) or loanwords (words borrowed into a language) may lead to DIF, but this is due to natural language evolution and should not be considered item bias. Japanese and Korean are two languages which have adopted a large number of English-derived loanwords into their lexicons (Baker, 2008b; Daulton, 2008). Daulton has shown, however, that multiple factors complicate loanword status, including the degree of conceptual overlap between the loanword and its English equivalent (e.g., the Japanese *feminisuto* referring to both *feminist* and *gentleman*) and the effects of shortening (e.g., *department store* becoming *depaato*) or compounding (e.g., *personal computer* becoming *pasocon*) that may occur when a word is borrowed into another language. These factors are likely to affect whether the presence of a loanword in the first language (L1) will lead to knowledge of its counterpart in a second language (L2). As the frequency with which a learner encounters a word is a powerful predictor of word knowledge (Milton, 2009), the frequency of a loanword in the L1 may also affect the likelihood of knowing the corresponding word in the L2.

2 Purpose of the study

The objective of this study was to ascertain whether DIF was present on a multiple-choice English vocabulary test administered to Japanese and Korean learners. If DIF were found, the study would then explore factors which could explain the differing response patterns between the two native-language subgroups, including (but not limited to) the frequency and range of use of loanwords as well as construct-irrelevant difficulty leading to item bias.

3 Method

The (convenience) sample included 184 Korean and 146 Japanese college students residing in their home countries at the time of the study. The instrument was a 90-item multiple-choice test designed to assess receptive written knowledge of general and academic vocabulary; it took approximately 30 minutes to complete. To ensure that the test was representative of the content domain (Messick, 1989), it was constructed by randomly sampling 30 words from each of the first and second 1,000 words of the General Service List (available at http://www.lextutor.ca/freq/

lists_download/) and 30 from the Academic Word List (Coxhead, 2000). The results of the test did not affect participants' course grades.

A separate calibration *t*-test approach was used to identify items displaying DIF between the Korean and Japanese groups. To account for any initial differences in group ability, Rasch-based person measures were calculated from the entire data set, and these values were anchored when items were calibrated separately for each group. The criteria for DIF was set at $p < 0.01$ and effect size >1 logit, calculated as the difference in Rasch item-difficulty between the two groups. This design was considered appropriate given the sample size and unknown levels of group ability (Linacre, 1994; Riley, 2011).

For each DIF item, the frequency with which the tested word occurs as a loanword in each of the Korean and Japanese languages was estimated. This count included all related inflected and derived forms of the word as defined by Bauer and Nation's (1993) level seven criteria for word family membership. It also included loaned compound words which retain a clear sense of the tested word. This approach was taken because research has suggested that the combined frequency of all members of a word family is a better predictor of word recognition than the frequency of only the word itself (Nagy, Anderson, Schommer, Scott, & Stallman, 1989).

Possible loanword forms were initially identified using bilingual English–Korean (Baker, 2008a) and English–Japanese (Daulton, 2008) loanword lists and Google Translate (http://translate.google.com/). One Korean and two Japanese informants, all highly English-proficient, confirmed loanword status and provided information regarding the range of use of each loanword in their native language as compared to English. Frequencies of these loanword forms were then retrieved from a Japanese list based on a 253-million word web-crawled corpus (University of Leeds, Center for Translation Studies, n.d.) and a Korean list based on a 3-million word corpus (The National Institute of the Korean Language, as cited by Schmitt Youngberg, 2006). Low-frequency loanwords encountered primarily in such contexts as media titles and blogs were not included in these counts if they were unknown to our informants and occurred less than once per million words of running text. Examples include loanwords in both languages for *collector*, used in movie titles, and *globalism*, found principally in blogs or minor online news venues.

The informants were also shown each DIF item and asked to identify possible sources of confusion and to speculate on whether any aspect of the item would make it particularly easy or difficult for members of their L1 group. These informants were not shown the direction of DIF. Finally, the authors reviewed each of these items for possible sources of DIF.

4 Results

Preliminary Rasch analysis indicated satisfactory person fit, item fit, and construct dimensionality. Twenty-one of the 90 test items displayed DIF. Ten of these were classified as loanwords in either Korean or Japanese and are shown with their frequency counts and DIF orientations in Table 1. The words tested in the remaining 11 DIF items are shown together with their DIF orientations in Table 2.

Table 1. Loanword DIF Items with Loanword Frequency in the L1

	Frequency[a]	
Tested word	Japanese	Korean
Easier for Japanese		
Link	276.6	<1.0
Zero	28.6	2.5
Hall	27.5	18.6
Globe	20.4	<1.0
Room	**15.2**	**29.2**
Sheet	12.8	6.7
Collection	12.3	1.8
Device	11.5	<1.0
Mixed	6.1	2.5
Easier for Koreans		
Pose	**5.9**	**4.6**

Note. Bold font denotes test items whose difficulties were not predicted by frequency.

[a]Frequency (per million words of running text) of loanwords which are family members or compound derivations of the tested word.

Table 2. Non-Loanword DIF Items

Tested word
Easier for Japanese
Interval
Rise
Easier for Koreans
Contemporary
Compound
Dominant
Erode
Solution
Crucial
Formula
Complain
Bias

In examining causes of DIF, the frequency of the tested word's loan equivalent in the native language predicted the direction of DIF in all of the words in Table 1 except *room* and *pose.*

Three additional and overlapping causes of DIF were tentatively identified. First, the range of use of the tested word's loan equivalent in each L1 helped explain the DIF orientation of *room* and provided additional insight into response patterns for *hall*. In Japanese, loan forms of *room* and *hall* appear by themselves and in numerous compounds (e.g., bedroom, living room; concert hall, wedding hall), and the range of use of these loanwords is similar to that of English. By contrast, the Korean loanwords for *room* and *hall* are used in fewer compounds, and their meanings are more restricted. Specifically, the Korean loan for *room* refers nearly exclusively to a room in a hotel or inn. Similarly, the Korean loan for *hall*, though retaining the English sense of "large room," is associated with something akin to a lobby or vestibule. This limited range of use of the Korean loanwords for *room* and *hall* may have contributed to item difficulty. In fact, despite the high frequency of these words in English, the Rasch-based difficulty estimates for the items testing these words were higher than for any other items for the Korean examinees.

Knowing these disparate usage patterns for *room* and *hall*, it became apparent that the wording of distractors was another possible source of DIF. For *room*, distractor b, *a place to play* (shown in Table 3), attracted many high-ability Korean respondents. Perhaps these individuals associated *room* with the loanword meaning *hotel room*, which in turn was associated with leisure. The interpretation is less clear for *hall*, but our Korean informant suggested that perhaps distractor b, *a place under a building*, was attractive because lobbies are typically on the first floor, under the rest of the building. Though it is possible that DIF would not have been present in these items if different distractors had been used, it is difficult to conclude that these items are biased because the difficulty that Korean respondents appear to experience cannot be considered construct irrelevant.

A final potential cause of DIF was phonological/orthographic overlap between the tested word and another English loanword in the L1. The loanword for *hall* in both Korean and Japanese is identical in written and spoken form to the loanword for *hole*. This could explain the relatively high number of respondents in both groups who chose distractor b, *a place under a building* (Table 3).

No cause of DIF was evident for *pose* in Table 1 or for any of the words listed in Table 2. Neither the informants nor the authors detected sources of bias in item

Table 3. Items with DIF from Multiple Causes

	Responses			
	Korean		Japanese	
Item	*n*	Ability *M*	*n*	Ability *M*
Room: How many rooms are there?				
a. Something to sleep on	18	0.55	6	−0.20
b. A place to play	84	2.90	9	0.28
c. Something for drawing with	4	−1.27	1	−1.44
d. A space in a building	75	2.56	130	1.21
Hall: We went into the hall.				
a. A large building	63	2.89	87	1.31
b. A place under a building	96	2.56	38	1.08
c. An office in a school	8	0.43	7	0.32
d. A house in a garden	9	0.77	5	0.19

Note. Bold font denotes the correct answer.

wording, and a post hoc analysis of non-uniform DIF resulted in no discernible pattern of differences between learners of high and low ability for these items.

In summary, linguistic (i.e., construct-relevant) sources of DIF were found for each loanword in Table 1 except *pose*, and no apparent sources of bias were found in the remaining items. The items for *room* and *hall* pose measurement-related problems in that one distractor in each item attracts examinees of unduly high ability, and for this reason, these items could be improved with revision.

5 Discussion

These findings reveal causes of difficulty in items assessing knowledge of words which are loanwords in the L1. Whereas previous research has found that yes/no cognate status could explain DIF between native language groups (Chen & Henning, 1985; Sasaki, 1991) and that items testing loanwords are generally easier than other words of similar frequency in the L2 (Beglar, 2010), we have found evidence that item difficulty may be mediated by both the frequency and the usage patterns of the L1 loanword as well as phonological/orthographic overlap with other loanwords in the L1.

The findings of this study have clear implications for development of second-language vocabulary tests for use in cross-cultural contexts. Even when bias is not present, test items which consistently demonstrate DIF require special attention if multiple versions of a test are purported to be of equivalent difficulty, or if item calibrations are used in computer-adaptive testing. In these cases, we concur with Brown and Iwashita (1996) in that best practice entails separately calibrating test items for each native language group. In the case of multiple versions, separate sets of equivalent forms could then be made for each native language group, and in the case of computer-adaptive testing, a prompt for examinees to enter their native language could be used to determine which set of item calibrations is applied.

The alternative approach would be to discard DIF items, but this is problematic for several reasons. First, there would be a risk of compromising construct representativeness in that it would disregard the principle of random sampling from word-frequency bands which is common in L2 vocabulary test construction. Second, it would overlook real linguistic differences between the groups. Third, it appears that DIF in L2 vocabulary items is relatively common between native language groups, making it impractical to discard such items in tests used with diverse populations.

This study was limited in that our native informants were those people available to us at the time of the study. As such, we had only one Korean and two Japanese informants, and none of them had expertise in language testing. A second limitation was that the Korean corpus was small. Though the texts in the corpus were representative of seven separate genres, and we needed only to determine whether loanword frequency was greater than that in Japanese, a larger corpus would have provided more certainty in those determinations.

Further research would help to clarify causes of DIF in L2 vocabulary testing. The present study has found that for items displaying DIF, the frequency and range of use of an L1 loanword equivalent are good predictors of the direction of DIF. However, an examination of the same variables for items *not* displaying DIF would also be useful, since it could offer insights as to whether loanword frequency can only *explain* DIF or whether it can also *predict* in which items it will occur.

Another research design that would help to clarify sources of DIF would be to engage speakers of Korean and Japanese in a think-aloud study. The informants used in this study provided some indication of how different L1 groups approach test items, and the possible sources of DIF described here might be confirmed if a greater number of informants explained their thought processes as they took each item. It would also be informative to conduct DIF analyses for popular instruments such as the VLT to ascertain whether multiple forms which have been found to be equivalent with one language group (e.g., Beglar & Hunt, 1999) remain so with others.

Author Note

Tim Stoeckel, Department of Comparative Culture, Miyazaki International College; Phil Bennett, Department of Comparative Culture, Miyazaki International College.

Correspondence concerning this article should be addressed to Tim Stoeckel, Department of Comparative Culture, Miyazaki International College, 1405 Kano, Kiyotake, Miyazaki 889-1605. E-mail: tstoecke@sky.miyazaki-mic.ac.jp,

References

Abbott, M.L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing, 24* (1), 7–36. doi:10.1177/0265532207071510

Baker, K. (2008a). English–Korean Transliteration List (v0.1). Electronic document. Retrieved from http://purl.oclc.org/net/kbaker/data

Baker, K. (2008b). *Multilingual distributional lexical similarity* (Doctoral dissertation). Retrieved from http://www.ling.ohio-state.edu/~kbaker/KirkBakerDissertation.pdf

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6* (4), 253–279. doi:10.1093/ijl/6.4.253

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27* (1), 101–118. doi:10.1177/0265532209340194

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing*, *16* (2), 131–162. doi:10.1177/026553229901600202

Brown, A., & Iwashita, N. (1996). Language background and item difficulty: The development of a computer-adaptive test of Japanese. *System*, *24* (2), 199–206. doi:10.1016/0346-251X(96)00004-8

Browne, C. (2008). A research-based approach to developing portable technology software for testing and teaching high frequency vocabulary. In I. Koch (Ed.), *CamTESOL Conference on English Language Teaching: Selected Papers*: Volume 4, 2008 (pp. 8–18). Retrieved from http://www.camtesol.org/Download/Earlier_Publications/Selected_Papers_Vol.4_2008.pdf

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, *2* (2), 155–163. doi:10.1177/026553228500200204

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34* (2), 213–238. doi:10.2307/3587951

Daulton, F.E. (2008). *Japan's built-in lexicon of English-based loanwords.* Clevedon, UK: Multilingual Matters.

Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12* (3), 38–47. Retrieved from http://ncme.org/publications/items/

Linacre, J.M. (1994). Sample size and item calibration (or person measure) stability. *Rasch Measurement Transactions, 7* (4), 328. Retrieved from http://www.rasch.org/rmt/contents.htm

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18* (2), 5–11. doi:10.3102/0013189X018002005

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Clevedon, UK: Multilingual Matters.

Nagy, W., Anderson, R.C., Schommer, M., Scott, J.A., & Stallman, A.C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24* (3), 262–282. doi:10.2307/747770

Pae, T.I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing, 29* (4), 533–554. doi:10.1177/0265532211434027

Park, G.P. (2008). Differential item functioning on an English listening test across gender. *TESOL Quarterly, 42* (1), 115–123.

Riley, B. (2011). Considering large group differences in ability in DIF analysis. *Rasch Measurement Transactions, 25* (2), 1326. Retrieved from http://www.rasch.org/rmt/contents.htm

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing*, *8* (2), 95–111. doi:10.1177/026553229100800201

Schmitt Youngberg, K. (2006). Vocabulary lists from the National Academy of the Korean Language. Available from: https://docs.google.com/file/d/0B1k1KSQCJJUKdFBvUGVqSzBLMVk/edit

University of Leeds, Center for Translation Studies. (n.d.). *Japanese frequency lists: Lemmas from the Internet corpus.* [Data file]. Retrieved from http://corpus.leeds.ac.uk/frqc/internet-jp.num

Zieky, M.J. (2006). Fairness review in assessment. In S.M. Downing, & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 359–376). Mahwah, NJ: Lawrence Erlbaum Associates.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense. Retrieved from http://educ.ubc.ca/faculty/zumbo/DIF/index.html