TAPESTRY

Volume 7
Issue 1 *Winter 2015*

Article 2

2015

Using DIF to Monitor Equivalence of Translated Tests in Large Scale Assessments: A Comparison of Native Speakers in their Primary and the Test's Source Language

Jorge Carvajal Espinoza

The Tapestry Journal Winter 2015, Volume 7, No. 1
ISSN 1949-8268 pp. 14-21

Using DIF to Monitor Equivalence of Translated Tests in Large Scale Assessment: A Comparison of Native Speakers in Their Primary and the Test's Source Language

Jorge Carvajal Espinoza, Ph.D., University of Costa Rica

Abstract

This study utilized a Differential Item Functioning (DIF) methodology for examining translated tests wherein all the examinees have the same native language -the target language of the translation- in order to provide information about the psychometric equivalence of the original and translated versions of the tests at item level. This study has also found possible explanations for translation DIF in certain types of items.

Keywords: Large Scale Assessment, Translation DIF, English Language Learners

Administering tests to English language learners (EL) students presents challenges given their diverse cultural background and language development (Solano Flores, 2010). If the items used to assess their achievement measure factors other than the targeted knowledge and skills ELs could be constrained in showing what they know (Sato et al., 2010). ELs by definition have not achieved English proficiency; therefore, assessments that are administered in English may impair their performance in the test (Robinson, 2010). A possible solution to these challenges is translating the test to the EL's native language. However, when tests are translated into one or more languages, the question of the equivalence of items across language forms arises (Price, 1998).

Differential Item Functioning (DIF) methodology has been used as a means to evaluate this equivalence (Sireci & Khaliq, 2002; Emenogu & Childs, 2003; Ulterwijk & Vallen, 2003; Sireci, Fitzgerald & Xing, 1998; Gierl, Rogers & Klinger, 1999; Robin, Sireci & Hambleton, 2003). DIF is said to exist when test takers of equal ability differ, on average, according to their group membership in their responses to a particular item (AERA, APA, NCME, 2014). DIF may be a threat to test fairness (Camili, 2006); therefore it is imperative to try to detect it and remove it from test items.

It is important to carry out analyses of equivalence at item level since "…consideration only to the results of scale-level methods as evidence of translation equivalence may be misleading because item-level DIF may not manifest itself in scale-level analyses" (Zumbo, 2003, p.146).

When DIF studies are conducted in the context of translated and non-translated tests, there will be confounding factors that can account for the DIF finding. The cause could be the translation itself (a poor quality translation or a translation at a different reading level, for example), or it could be rooted in factors associated with group membership (cultural differences or differences in curricula, for example).

The objectives of this study are: a) to propose and apply a DIF methodology examining translated tests wherein all the examinees have the same native language (the target language of the translation) in order to provide information about the psychometric equivalence of the original and translated versions of the tests at item level and b) to identify possible explanations for items showing DIF. Specifically, the examinees are Hispanic ELs who are required to take a grade 4 or 7 mathematics test in English or the Spanish translated version in a statewide assessment program.

We believe that the adequacy of a translation can be better evaluated combining information obtained from this comparison (common native speakers of the target language of the translation taking either the original language or the translated tests) with information from the more traditional comparison found in the literature of using DIF methodology between students with different native languages. It is important to note that no studies have yet carried out the proposed comparison. Duncan et al. (2002) mentioned this comparison but did not conduct the DIF study due to small sample size in both the reference and focal groups. Sireci & Khaliq (2002) studied DIF in Dual Language test forms but they

compared English native speakers and Spanish native speakers. This study seeks to contribute to address this gap in the translation DIF literature.

METHOD

This study analyzed the original and translated versions of two different mathematics tests for grades 4 and 7 in a statewide assessment program.

The translation procedure was carried out as follows. For each test, two translators independently made a translation into Spanish. Then they did a consensual validation of the translation. A third translator, expert and proficient in both Spanish and English, then compared the English and the consensual version and made suggestions. The first two translators prepared a final version based upon those suggestions. All three translators are native speakers of Spanish, one of the translators is a mathematics educator, and the remaining two are English/Spanish foreign language educators.

Statewide, ELs were assigned by schools, according to their English proficiency level, to take one of the following tests: Plain English test (group A), Spanish version test (group B) and General Assessment test (group C).

This study compared only groups A and B and included only Hispanic ELs. The Plain English test is a language-simplified version of the General Assessment and the Spanish version is a translation of the Plain English test. In the Plain English test, the construct measured is intended to remain the same as in the General Assessment test and the procedure to solve each item is the same but the language is simplified by eliminating irrelevant information and using shorter sentences. In this sense, it is likely that the Plain English test has less difficulty than the General Assessment and therefore it would not be appropriate to use the total score from the General Assessment as a matching score in a DIF analysis with any of the other two groups. Comparisons involving group C were, therefore, not adequate for the objectives of this study.

The number of subjects included in the study was as follows. In the 4^{th} grade, group A had 871 subjects and group B had 86. In the 7^{th} grade, group A had 334 subjects and group B had 67. Each test consists of 52 multiple choice items, but after scoring and analysis one item was dropped from the 7^{th} grade test in both the Plain English and Spanish versions.

For each grade, group A is considered the reference group whereas group B is considered the focal group for the DIF study. The Mantel-Haenszel (MH) technique was chosen due to the small sample size. The FORTRAN program written by Raju (1988) was used to compute the MH statistics. The reported MH statistics are those computed with the studied item included, as recommended by Clauser & Manzor (1998) for when an internal matching criterion is used, as was the case of this study. Items were identified as showing DIF if both the Chi-Square statistic was significant at the .05 level and the MH-Delta-DIF was greater than 1.5 in absolute value.

After identifying the DIF items we conducted a judgmental analysis reviewing the translation for the flagged items and exploring possible causes of DIF.

RESULTS

Table 1: Math Assessment Descriptive Statistics

	Group	Subjects	Mean Score	Score SD	Alpha
Grade 4th	A	871	26.9	8.1	.84
	B	86	26.8	8.1	.84
Grade 7th	A	334	21.5	6.7	.78
	B	67	18.6	4.35	.47

Results revealed that in grade 4 descriptive statistics were very similar in groups A (Plain English) and B (Spanish). On the other hand, in grade 7 the mean score in group B was about 3 points lower than that in group A and the standard deviation was smaller in group B.

Table 2: Grade 4 DIF Results

Item	Chi-square	MHDelta	Item	Chi-square	MHDelta	Item	Chi-square	MHDelta
1	0.2	-0.44	19	0.49	-0.49	37	0.39	0.52
2	0.02	0	20	0.29	-0.37	38	0.08	-0.23
3	2.01	-0.85	21	3.2	-1.23	39	0.57	0.57
4*	6.89	1.66	22	0.01	0.02	40	0.34	-0.41
5	5.95	1.48	23	0	-0.07	41	1.64	0.88
6	0.55	-0.51	24	0.14	-0.29	42	0	-0.09
7	0.05	0.2	25	1.87	0.85	43	0.69	-0.5
8	0.25	0.39	26	0.4	-0.48	**44***	7.78	-1.61
9	0.38	-0.53	27	0.09	0.28	45	0.23	-0.57
10	1.5	0.82	28	0.73	-0.57	46	0.15	0.29
11	0.01	-0.17	29	0.02	0.17	47	2.06	-1.08
12	0.01	0.14	30	0	-0.1	48	0.02	-0.15
13	1.23	0.79	31	0.13	0.31	49	2.47	1.1
14	1.55	-1	32	0.24	0.41	50	0.08	-0.27
15	1.39	0.81	33	0.23	-0.34	51	1.71	-0.81
16	0.07	-0.24	34	0.23	0.36	52	1.61	0.86
17	0.05	-0.21	35	1.63	-0.86			
18	0.92	0.67	36	0.78	0.59			

* flagged items

In the 4th grade test, two items were flagged as showing DIF: Item 4 (favors focal group - B) and Item 44 (favors reference group - A). Item 4 involves a direct translation and we did not find any judgmental reason for DIF. The same applies to Item 44.

Table 3: Grade 7 DIF Results

Item	Chi-square	MHDelta	Item	Chi-square	MHDelta	Item	Chi-square	MHDelta
1	0.43	-0.57	18	0.12	0.33	35	0.23	-0.42
2	0.7	-1.25	19	0.29	-0.48	36	0.66	0.78
3	2.88	1.4	**20***	5.55	1.78	**37***	9.22	-2.4
4	0	-0.11	21	1.27	0.9	38	0.16	-0.41
5	0.7	0.63	22	2.7	1.36	39	0.21	-0.4
6	0.02	0	23	0.03	0.3	40	0.32	0.49
7	0.84	0.83	24	0.62	0.87	41	0.76	0.71
8	0.12	-0.34	25	0.51	-0.82	42	1.87	1.01
9	0	0.09	26	1.12	-1.1	43	0.01	0.06
10	0.03	0.2	**27***	15.44	-3.33	44	0.01	-0.08
11	0	0.13	**28***	4.43	-1.69	45	0.01	0.04
12	1.54	-1.03	**29***	4.89	-1.62	46	0.07	-0.32
13	0.36	-0.55	30	0.22	-0.43	47	3.58	1.38
14	0.12	-0.37	31	2.33	1.15	48	2.23	1.05
15	0.56	0.64	**32***	10.88	-2.72	49	0.77	1.06
16	1.08	-0.83	33	1.34	1.19	**50***	4.92	1.79
17	1.42	-1.39	34	0.05	-0.26	51	3.51	1.34

* flagged items

In the 7th grade test, seven items were flagged as showing DIF: Items 20 and 50 (favor focal group – B) and Items 27, 28, 29, 32 y 37 (favor reference group – A). Items 20 and 50 were direct translations and we did not find any judgmental reason for DIF. The rest of the flagged items favor the students taking the Plain English version and are discussed below.

Item 27: the term “mean” was translated as “media” which is a correct translation. However it is possible that Hispanic students are not familiar with the word “media” as a technical statistics term. The term “media” has another meaning, “sock,” which is a more common meaning than the technical one.

Items 28 and 29: we did not find any judgmental reason for DIF.

Item 32: the term “mode” was translated as “moda” which is a correct translation. However, as in item 27, Hispanic EL2 students might not be familiar with the technical meaning of “moda.” There is a more common meaning for “moda” which is “fashion”.

Item 37: the term “range” was translated correctly as “rango.” Even though the technical meaning of “rango” in Spanish is closer to its common meaning, again it might be that Hispanics students are not familiar with its technical meaning.

DISCUSSION

This study conducted a DIF analysis comparing subjects who are native speakers of the target language of two translated tests and who were required to take the tests in either the source or target language in a statewide assessment program. This type of comparison has been called for in the DIF literature, yet no research to date has reported any findings associated with this methodology. We think that the proposed comparison can generate important information about factors producing translation DIF.

We are proposing that this comparison be used in the future, when feasible, in combination with the more traditional approach of comparing source language native speakers with target language native speakers. We combined in this study a statistical method of detecting DIF with a judgmental analysis. These two types of analyses do not always coincide but both are important in DIF studies in the process of collecting evidence towards test validation (Clauser & Manzor, 1998). The results of this study are, of course, limited to two particular tests and more studies are needed to evaluate the contribution of the proposed comparison.

Other studies have found an improvement in performance when ELs are administered translated tests as opposed to English versions of the test (Robinson, 2010, Sato, 2010). This study utilized a comparison that can contribute to a better understanding of the effects of administering a translated test to ELs and therefore to improve such translations. These improvements can help to increase the validity of the translations.

Due to the fact that these are live tests we cannot provide further disclosure of the items. However, these findings suggest that items that involve terms that have both technical and more colloquial meanings could tend to produce DIF, especially if ELs are taught in English. DIF studies such as this one are then helpful for detecting undesirable differences in performance related to traits other than those intended to be measured by the test.

This study was conducted in a setting in which it was possible to compare native speakers of the target language taking the test in either the source or the target language in the context of a statewide assessment. Results from this kind of comparison have not been reported in the translation DIF literature and can generate a better understanding of the causes of DIF. In addition, the study has found possible explanations for translation DIF in certain types of items. These contributions can help to increase the validity of future translated tests.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education*, 16, 55-73.

Allalouf, A., Hambleton, R., & Sireci, S. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, 36, 185-198.

Camili, G. (2006). Test Fairness. In R. L. Brennan (Ed), *Educational Measurement (4th ed.)(p. 221-256)* Westport,CT: American Council on Education/Praeger.

Clauser, B., & Mazor, K. (1998) Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17, 31-44.

Duncan, T., Parent, L., Chen, L., Ferrara, S., & Johnson, E. (2002). *Study of a dual language test booklet in 8th grade Mathematics.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*; 2, 199-215.

Emenogu, B., & Childs, R. (2003). *Curriculum and translation differential item functioning: a comparison of two DIF detection techniques.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Gierl, M., Rogers, T., & Klinger, D. (1999). Using statistical and judgmental reviews to identify and interpret translation differential item functioning. *Alberta Journal of Educational Research. 45, 353-376.*

Hambleton, R., & Patsula, L. (2000). *Adapting tests for use in multiple languages and cultures.* (Laboratory of Psychometric and Evaluative Research, Report No. 304). Amherst: University of Massachusetts, School of Education.

Holland, P., & Wainer, H. (Eds.). (1993). *Differential item functioning. Hillsdale, NJ: Erlbaum Publishers.*

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language testing,* 18, 89-114.

Price, L. (1999). D*ifferential functioning of items and tests versus the Mantel-Haenszel technique for detecting differential item functioning in a translated test.* Paper presented at the meeting of the American Alliance of Health, Physical Education, Recreation, and Dance, Boston, MA.

Price, L., & Oshima, T. (1998). *Differential item functioning and language translation: a cross-national study with a test developed for certification.* Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

Muñiz, J., & Hambleton, R. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing,* 1, 115-135.

Robin, F., Sireci, S., & Hambleton, R. (2003). Evaluating the equivalence of different language versions of a credentialing exam. *International Journal of Testing*, 3, 1-20.

Robinson, J. P. (2010). The effects of test translation on young English learners' Mathematics performance. *Educational Researcher*, 39(8), 582-90

Sato et al. (2010). Accommodations for English language learner students: the effect of linguistic modification of Math test item sets. Institute of Education Sciencies.

Sireci, S., & Khaliq, S. (2002). An analysis of the psychometric properties of dual language test forms. (Center for Educational Assessment, Report No. 458). Amherst: University of Massachusetts, School of Education.

Sireci, S., Fitzgerald, C., & Xing, D. (1998). *Adapting credentialing examinations for international uses.* Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

Solano Flores, G. (2012). Adaptacion Linguistica y cultural de pruebas de rendimiento académico (66-68). In INEE: Una década de evaluacion 2002-2012. IEPSA

Ulterwijk, H., & Vallen, T. (2003). Test bias and differential item functioning: a study of the suitability of the CITO primary education final test for second generation immigrant students in the Netherlands. *Studies in Educational Evaluation*, 29, 129-143.

Zenisky, A., Hambleton, R., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: a study evaluating a two-stage approach. *Educational and Psychological Measurement*, 63, 51-64.

Zumbo, B. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing* 20, 136-147.