Actualidades en Psicología, 29(119), 2015, 131-139
http://revistas.ucr.ac.cr/index.php/actualidades

ISSN 2215-3535
DOI: http://dx.doi.org/10.15517/ap.v29i119.19244

Using Bilingual Examinees to Evaluate the Comparability of Test Structure across Different Language Versions of a Mathematics Exam

Evaluación de la comparabilidad de la estructura de la prueba por medio de examinados bilingues en versiones en diferentes idiomas de un examen de Matemática

Tia Sukin[1]
Pacific Metrics, Inc., United States

Stephen G. Sireci[2]
University of Massachusetts Amherst, United States

Saw Lan Ong[3]
Universiti Sains Malaysia, Malaysia

Abstract. Malay- and English-language versions of a mathematics exam were analyzed for structural equivalence by administering both versions to a group of Malay-English bilingual students. The analysis and comparison of test structure was determined using both DIMTEST and weighted multidimensional scaling. The assessment was found to be unidimensional and to possess similar structure across the two language versions. Implications of this study suggest bilingual examinees can be used to evaluate the invariance of test structure across translated test forms. Future research should explore situations where bilingual examinees can be used to link different language versions of assessments for monolingual populations.

Keywords. Bilinguals, cross-lingual assessment, dimensionality, invariance, test structure, validity.

Resumen. Se analizó la equivalencia estructural entre las versiones de un examen de matemáticas en lengua malaya e inglesa mediante la administración de ambas versiones a un grupo de estudiantes bilingües en ambas lenguas. El análisis y comparación de la estructura del test fue realizada utilizando DIMTEST y escalamiento multidimensional ponderado. Se encontró que la evaluación es unidimensional y posee una estructura similar en las dos versiones. Las conclusiones de este estudio sugieren que se pueden utilizar personas bilingües para evaluar la invarianza de la estructura del test utilizando formas traducidas de un test. Las investigaciones futuras deberían explorar situaciones donde se puedan utilizar personas bilingües para conectar distintos idiomas en las evaluaciones de las poblaciones monolingües.

Palabras clave. Estructura de la prueba, validez, evaluación en varios idiomas, examinados bilingues, dimensionalidad, invariancia.

[1]Tia Sukin. Pacific Metrics, Inc., United States. Postal Address: Lower Ragsdale Drive Suite 1150 Monterey , California 93940. United States. Email: info@pacificmetrics.com

[2]Stephen G. Sireci. School of Education, University of Massachusetts. Postal address: 156 Hills South, Amherst, MA 01003, Massachusetts, United States. Email: sireci@acad.umass.edu

[3]Saw Lan Ong. School of Educational Studies Universiti Sains Malaysia, Malaysia, George Town, Penang, Malaysia. Email: osl@usm.my

Introduction

Diversity in the language spoken by students within and across countries has necessitated the process of adapting educational tests for use across multiple languages (Hambleton, Merenda, & Spielberger, 2005). International assessments such as the Trendsa in International Mathematics and Science Study (TIMSS; Mullis, Martin, & Foy, 2008), Program for International Student Assessment (PISA; Organization for Economic Cooperation and Development (OECD), 2006), Progress in International Reading Literacy (PIRLS; Baer, Baldi, Ayotte, Green, & McGrath, 2007), and the Program for the International Assessment of Adult Competencies (PIAC; Statistics Canada & OECD, 2005) are examples of large-scale tests that are administered in multiple languages so that comparisons can be made across examinees who function in different languages.

Within many countries, cross-lingual assessment is also necessary. Adapted tests based on test translation are used in Canada (Gierl & Khaliq, 2001), the United States (Sireci & Khaliq, 2002), and many other countries.

Measurement of educational or psychological constructs across languages typically involves translation. The process of translating a test from one language to another is known as adaptation because the intent is to reproduce the meaning and intent of each item in the target language, as opposed to a literal word-by-word translation (Hambleton, 2005). Although test adaptation facilitates the assessment and comparison of students who operate in different languages, that different language versions of a test are equivalent with respect to psychometric properties cannot be assumed. Adapting tests for use across multiple languages may result in differences in difficulty across the different language versions of a test or in the different versions measuring different constructs altogether (International Test Commission, 2010; Sireci, 1997; Sireci, Rios, & Powers, in press; van de Vijver & Poortinga, 2005).

The degree to which adapted versions of tests are equivalent across languages is an important issue in considering the validity of tests used across different language groups. The Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014), the Guidelines for Translating and Adapting Tests (International Test Commission, 2010), and many researchers (e.g., Hambleton, 2005; Sireci, 2011; Van de vijver & Poortinga, 2005) argue that empirical evidence must be put forward to support the validity of inferences derived from cross-lingual assessments, especially when comparisons of test performance are made across different language groups.

However, providing data to support the validity of cross-lingual assessments is difficult because one cannot assume that the items on the different language versions of the test are equivalent, and one cannot assume the different groups of examinees to be equivalent. Thus, there is nothing to anchor a true comparison of test difficulty or construct equivalence across languages (Sireci, 1997).

One way around this problem is to administer different language versions of an assessment to a sample of examinees who are proficient in both languages (Sireci, 2005). Bilingual examinees may represent a common group upon which comparisons of tests and items can be made. In this paper, the authors explore the utility of bilingual examinees for evaluating the factorial invariance (i.e., structural equivalence) of two different language versions of a ninth-grade math test administered in Malaysia. Both English and Malay versions of the test were administered in counterbalanced order to English-Malay bilingual students. Bilingual students have been used to evaluate cross-lingual invariance of survey items (Sireci & Berberoglu, 2000) and to link educational tests across languages (Boldt, 1969; CTB, 1988). However, the use of bilinguals for these purposes is rare, and there has been little

study of the invariance of test structure across languages using a bilingual group.

In this study, the authors analyze data from Malay-English bilingual ninth-grade students in Malaysia who received math instruction in English even though they were native speakers of Malay. These students took both English and Malay versions of a math test in counterbalanced order. Ong and Sireci (2008) evaluated the relative difference in difficulty across the English and Malay versions of the exam by using the bilingual group to equate the two test forms. Equating based on both classical test theory and item response theory (IRT) was conducted, and they concluded a one-point adjustment was needed to put the scores on the same scale (the Malay form was slightly easier according to the equating results).

The results of Ong and Sireci (2008) supported the use of bilingual examinees for adjusting for differences in difficulty across translated test forms. However, such a conclusion assumes the tests are invariant with respect to dimensionality (Millsap, 2007). If different dimensions are needed to account for the item variation within each language version of the test, to equate them would not make sense. If structural differences are observed, it might indicate that one or more items were perceived differently based on the language in which it was presented. Evaluation of whether the structure of the Malay and English versions of the exam are the same provides evidence regarding the degree to which scores on the two different language versions of the assessment are comparable.

The present study represents a new analysis of the data from Ong and Sireci (2008). In addition to evaluating an untested assumption in the earlier study, the authors demonstrate how bilingual examinees can be used to evaluate the structural equivalence of different language versions of an assessment.

Method

Data

Data from the 2005 Lower Secondary School Achievement Mathematics Test for ninth grade Malaysian students were used in this study. This test was administered in both English and Malay. Examinees typically see both language versions of the items in a dual-language test booklet when responding (i.e., both English and Malay versions of the items are printed on facing pages in the same booklet). However, as part of a special study (Ong & Sireci, 2008), the test booklets were designed such that only items for one language were presented during a given testing occasion.

The only difference between the two test forms was the language in which the items were written (English or Malay). The mathematics exam consisted of 40 dichotomously scored multiplechoice items and covered the content areas of algebra, measurement, geometry, and statistics.

A total of 505 examinees took both the English and Malay versions of the exam. The administration design was counterbalanced such that 255 examinees took the English version first and 250 students took the Malay version first. The interval between testing occasions was three weeks.

To evaluate the sampling variability of our dimensionality investigation, the authors randomly split the data from each language administration into two separate samples of approximately 250 for each version of the test. These two random samples were created for each language version so that "within language" variability could be assessed. An analysis of variance (ANOVA) was preformed to assess total test score differences between the four groups (2 English samples and 2 Malay samples), with the expectation that there would be no differences within each language version of the exam.

Data Analysis

Multidimensional scaling (MDS) and DIMTEST were used to evaluate unidimensionality and the similarity of dimensionality across the English and Malay versions of the test. The purpose of the DIMTEST analysis was to evaluate it the structure of the data for each group was essentially unidimensional. The MDS analyses were designed to see if there was any variation in dimensionality across the groups and if any secondary dimensions were detectable.

Multidimensional scaling analyses

MDS is a data analytic procedure that fits dimensions to proximity data so that the underlying structure of the data can be uncovered. In evaluating the structure of an educational assessment, distances or correlations can be computed among items using an MDS analysis. A separate matrix of inter-item Euclidean distances for each group was computed. These distance matrices served as the input data for the MDS analyses. Ordinal (non-metric) MDS was implemented, which means the input distances were subject to a monotonic transformation that preserved the rankorder of the original distances, but allowed for improved fit to the MDS solution. MDS computes coordinates for the items on a pre-specified number of dimensions to minimize the discrepancy between the transformed distances and the distances among the items in the MDS space. The classical (one-matrix) MDS model is

$$d_{jj'} = \sqrt{\sum_{r=1}^{R} (x_{jr} - x_{j'r})^2} \qquad (1)$$

where $d_{jj'}$ is the distance between item j and j' in the MDS space, x_{jr} is the coordinate of item j on dimension r, and R is the maximum number of dimensions specified in the model.

In multi-group MDS analyses (weighted MDS), there is more than one input matrix corresponding to multiple individuals or multiple groups. In the present study, four inter-item distance matrices were used—two derived from the two random samples from the English version of the exam, and two derived from the random samples from the Malay version of the exam. The equation for weighted MDS (Carroll & Chang, 1970) is

$$d^{k}_{jj'} = \sqrt{\sum_{r=1}^{R} w^{k}_{r} (x_{jr} - x_{j'r})^2} \qquad (2)$$

where w^{k}_{r} corresponds to the weight associated with dimension r for group k, and the remaining terms are as defined in equation 1.

The end result of a weighted MDS analysis is (a) a multidimensional configuration of stimuli (in this case, test items) that best fits the data for all groups when considered simultaneously, and (b) a matrix of group weights (with elements w^{k}_{r}) that represent how the group stimulus space should be adjusted to best fit the data for a particular group (k). The weights on each dimension for each group can be used to "stretch" or "shrink" a dimension from the simultaneous solution to create a solution that best fits the data for a particular group. Thus, the weights (w^{k}_{r}) contain the information regarding structural differences across groups.

A finding of similar dimension weights across all groups would suggest structural equivalence of the test data across the groups, while differences in group weights would indicate a lack of structural equivalence. Using simulated data, Sireci, Bastari, and Allalouf (1998) found that when structural differences existed across groups, one or more groups have weights near zero on one or more dimensions relevant to at least one other group. In the present study, all MDS analyses were conducted in SPSS 16.0 using the PROXSCAL algorithm (SPSS, 2007).

DIMTEST analyses

DIMTEST (Stout, 1987; Stout, Douglas, Junker, & Roussos, 1993) can be used to test the hypothesis that a set of items are "essentially" unidimensional. The DIMTEST analysis involves creating three

subsets of items, (a) Assessment Subtest 1 (AT 1), (b) Assessment Subtest 2 (AT 2), and (c) Partitioning Test (PT). AT 1 is made up of items that are most likely to be dimensionally different from one another. AT 2 is made of items that are as similar in difficulty as possible to those of AT 1. The PT is made up of the rest of the items and is used for stratifying examinees into *K* proficiency groups. While conditioning on the scores of PT, the covariation of the item scores on AT 1 and AT 2 are examined by computing two T-statistics. Each involves the calculation of two variance components; the first is based on observed subtest scores and the second is based on expected subtest scores, when unidimensionality is assumed. The equation for the observed variance component is,

$$T_L = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \frac{\hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2}{S_k} \qquad (3)$$

where the first variance estimate ($\hat{\sigma}_k^2$) is based on observed subtest scores and the second ($\hat{\sigma}_{U,k}^2$) is based on expected subtest scores given an unidimensional model. The variance estimate differences are then standardized. A second T-statistic is used to correct for the statistical bias associated with examinees and item difficulty. Thus, a final T-statistic is calculated and used to interpret whether 'essential' unidimensionality exists using conventional statistical significance values for the t distribution. A t value associated with a significance level of $p < .05$ was taken as an indication of a lack of essential unidimensionality. The default options in DIMTEST were used for selecting the subsets of items to be used in the AT 2 and PT subtests.

The default option in DIMTEST for selecting AT 2 item s involves performing a principal components analysis on the matrix of inter-item tetrachoric correlations and then identifying the items with the largest loadings on the second component (Stout et al., 1993).

Results

Equivalence of Samples

The ANOVA confirmed no statistically significant differences between total score means for each of the two random samples taken from each language administration of the exam ($F(3,994) = 1.58$, $p = 0.19$). Table 1 presents total score means, standard deviations (SD), and standard error of the mean (SEM) along with coefficient alpha (α) for the four groups. There is little variability in any of these descriptive statistics within and across languages.

Dimensionality

DIMTEST Results. The DIMTEST results (using the Malay test items) revealed that items were 'essentially' unidimensional (T' = 0.92, $p = 0.18$). 'Essential' unidimensionality was also confirmed using the responses to English test items (T' = 0.92, $p = 0.18$). These results suggest a dominant dimension can be used to account for the item variation in both the Malay and English versions of the exam.

Weighted MDS Results

Determining the number of dimensions underlying the data was based on the MDS fit values of SSTRESS and dispersion accounted for (DAF). SSTRESS is a badness of fit index and represents the normalized squared residual variance of the monotonic regression of the MDS distances on the transformed item distance data. Lower values of SSTRESS, and higher values of DAF, indicate better fit of an MDS model.

Table 1

Descriptive Statistics for Total Test Scores

Group	N	Mean	SD	SEM	α
Malay Sample 1	250	31.5	7.1	2.25	0.90
Malay Sample 2	249	31.8	7.1	2.25	0.90
English Sample 1	250	30.9	7.4	2.22	0.91
English Sample 2	249	30.5	7.5	2.25	0.91

Note. SEM=standard error of measurement.

Table 3
Dimension Weights by Group, Two-Dimensional Solution

Group	MDS Weights	
	Dimension 1	Dimension 2
Malay Sample 1	.48	.48
Malay Sample 2	.42	.31
English Sample 1	.30	.22
English Sample 2	.46	.50

Table 2
SSTRESS and DAF for 2-6 Dimensional Solutions

Dimensions	SSTRESS	DAF
1	0.23	0.85
2	0.14	0.93
3	0.09	0.95
4	0.07	0.97
5	0.06	0.97
6	0.05	0.98

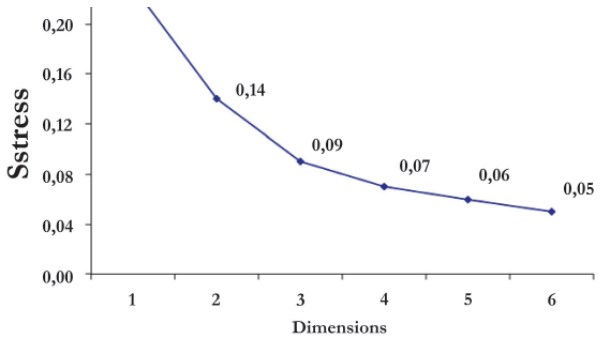

Figure 1. SSTRESS Elbow Plot. This SSTRESS value was obtained by assuming similar structure across groups and performing a replicated MDS analysis.

The results of the MDS fit analyses are presented in Table 2. The fit values for the unidimensional solution indicated the presence of a strong first dimension (SSTRESS = 0.23, DAF = 0.85); however, the two-dimensional solution led to a noticeable improvement in fit with about 8% additional variation in the data accounted for by the second dimension. After two dimensions, the improvement in fit tapered off (see Figure 1). These results suggest that evaluating structural equivalence using the two-dimensional solution should be sufficient for capturing any potential lack of meaningful variation in structure across the Malay and English versions.

The weights for each sample on each of the two dimensions are reported in Table 3 and displayed in Figure 2. There was little variation in weights across the groups. In fact, the greatest difference was found across the two English-language random samples (E1 and E2) on the first dimension. These results support the conclusion of structural invariance across the English and Malay versions of the test.

Discussion

The results of this study suggest that the dimensional structures of the English and Malay versions of this mathematics exam are similar. It is likely that, in general, the translation of these items retained their general difficulty. This finding supports the results of Ong and Sireci (2008) in that it supports the assumption of invariance of test structure across test forms and is congruent with their results that only a small adjustment in test difficulty (one-point) was needed.

Methodologically, the results suggest that weighted MDS is a useful procedure for evaluating the similarity of test structure across different language versions of a test administered to a common group of examinees. Although other studies have investigated similarity of test structure using different, monolingual groups of examinees, this study may be the first to use weighted MDS on a bilingual sample. DIMTEST confirmed the intended unidimensionality of the test data, but the MDS analysis suggested the presence of an additional secondary dimension, which allowed us to evaluate any

differences across the English and Malay versions with respect to the dominant and secondary dimensions. Had a greater improvement in fit from one to two dimensions been observed in the MDS analyses, it is likely the DIMTEST results would also have suggested multidimensionality. The degree to which DIMTEST and MDS provide similar conclusions regarding test dimensionality deserves further study, preferably using simulated data.

It is important to note that the examinees in this study are especially unique in that they were highly proficient in both the Malay and English language as instruction was delivered in both languages. Therefore, performance differences between the Malay and English versions of the exam were attributable to difficulty differences between the forms and not differences between language proficiency. The present study revealed that the 9th grade Malaysian mathematics exam was 'essentially' unidimensional and the structural composition of the Malay and English versions of the test were similar, which indicates the same dominant dimension was being measured. These results support the use and comparison of translated and adapted assessment forms among bilingual or multilingual populations. Additionally, these results provide some support for the use of bilingual examinees as the linking group between two language versions of an assessment intended to also assess monolingual examinees.

The question remains whether the different language assessments in this study are equivalent not only for Malay-English bilingual students, but for monolingual English and monolingual Malay students. The results of the study were consistent with the hypothesis that the different language versions are equivalent for all populations, but of course making that conclusion is generalizing too far from the present results, given the uniqueness of the bilingual sample. Thus, future research should consider including monolingual

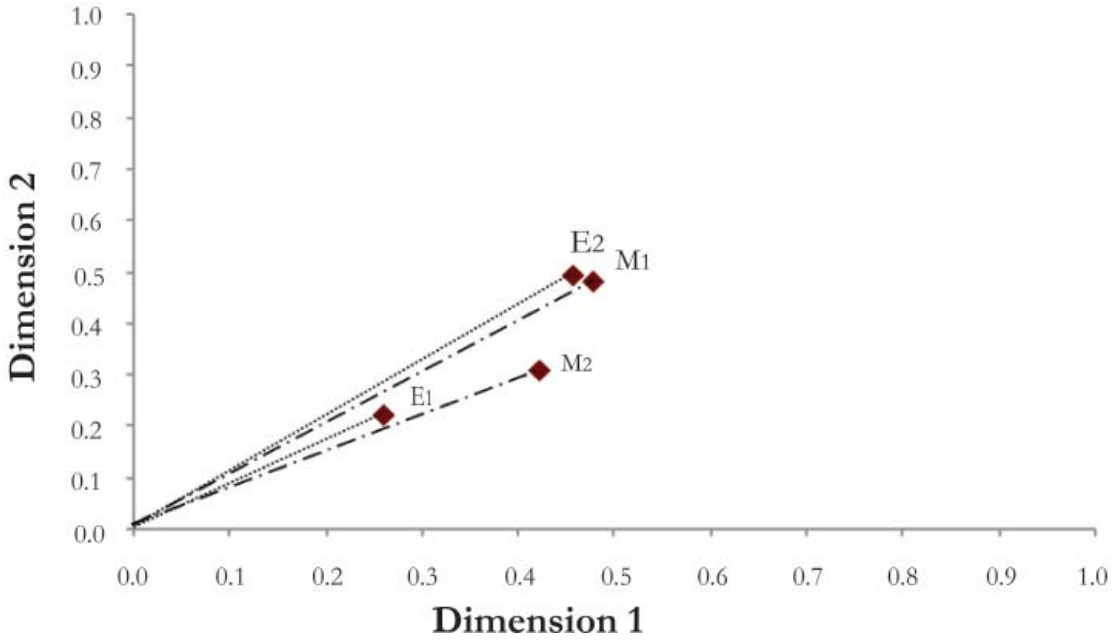

Figure 2. Dimension Weight Vectors, Two-Dimensional Solution. E1= Sample 1 from English version, M1= Sample 1 from Malay version, E2= Sample 2 from English version, M2= Sample 2 from Malay version.

groups along with bilingual groups in the analysis of the structural invariance of different language forms of a test. The multigroup MDS procedure used in the present study could accommodate additional groups, and it would be interesting to explore the similarity of the dimension weights not only across language versions of the test, but across monolingual and bilingual populations. *Multi-group confirmatory factor analysis* (CFA) can also be used to simultaneously evaluate the invariance of test structure across multiple groups. Previous research has shown multi-group CFA and WMDS provide similar decisions regarding invariance of test structure (e.g., Sireci & Wells, 2010), but clearly more research in this area is needed.

Conclusion

In this study, the authors analyzed the dimensionality of students' responses to test items to evaluate the similarity of the dimensionality of these data across groups of students who responded to English and Malay versions of the items. Our analyses of structural invariance provided some evidence that the different language versions of the exam were comparable, at least from the perspective of validity evidence based on test structure—one of the five sources of evidence stipulated by the Standards for Educational and Psychological Testing (AERA et al., 2014). Our analyses also show the utility of DIMTEST and WMDS for evaluating underlying dimensionality and the invariance of that dimensionality across different language versions of an assessment. Our design featured bilingual examinees, but future research could include both monolingual and bilingual examinees, and could involve additional statistical analyses such as CFA.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, D.C.: American Educational Research Association.

Baer, J., Baldi, S., Ayotte, K., Green, P. J., & McGrath, D. (2007). *The reading literacy of U.S. fourth-grade students in an international context: Results from the 2001 and 2006 Progress in International Reading Literacy Study (PIRLS).* (NCES-2008-017). Washington, DC: U.S. Department of Education. http://nces.ed.gov/pubs2008/2008017.pdf

Boldt, R. F. (1969). *Concurrent validity of the PAA and SAT for bilingual Dade County high school volunteers.* (College Entrance Examination Board Research and Development Report 68-69, No. 3). Princeton, NJ: Educational Testing Service.

Carroll, J. D., & Chang, J. J. (1970) Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. *Psychometrika, 35*, 283-319.

CTB/McGrawBHill (1988). *Spanish assessment of basic education: Technical report.* Monterey, CA: McGraw Hill

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting test into multiple languages and cultures. In R. K. Hambleton, R. Merenda, & C. Spielberger (Eds.) *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R. K., Merenda, P.F. & Spielberger, C.D. (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Hillsdale, NJ: Lawrence Erlbaum.

International Test Commission (2010). Guidelines for translating and adapting tests. Retrieved from http://www.intestcom.org.

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72*(4), 461-473.

Mullis, I.V.S., Martin, M.O., & Foy, P. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Ong, S. L., & Sireci, S. G. (2008). Using bilingual students to link and evaluate different language versions of an exam. *US-China Education Review, 5*, 37-46.

Organisation for Economic Co-operation and Development. (2006). *Literacy skills for the world of tomorrow—further results from PISA 2003*. Paris: Author.

Prieto, A. J. (1992). A method for translation of instruments to other languages. *Adult Education Quarterly, 43*(1), 1-14.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16*(1), 12-19.

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117-138). Hillsdale, NJ: Lawrence Erlbaum.

Sireci, S. G. (2011). Evaluating test and survey items for bias across languages and cultures. In D. Matsumoto and F. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 216-243). Oxford, UK: Oxford University Press.

Sireci, S. G., Bastari, B., & Allalouf, A. (1998, August). *Evaluating construct equivalence across adapted tests*. Invited paper presented at the annual meeting of the American Psychological Association (Division 5), San Francisco, CA.

Sireci, S. G. & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education, 35* (2), 229-259.

Sireci, S. G., & Khaliq, S. N. (2002, April). *An analysis of the psychometric properties of dual language test forms.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flawed items in the test adaptations process. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-115). Hillsdale, NJ: Lawrence Erlbaum.

Statistical Package for Social Sciences (SPSS). (2007). *Multidimensional Scaling (PROXSCAL)*. SPSS Categories™ 16.0 (pp. 64-78). Chicago, IL: SPSS, Inc. Available online at: http://support.spss.com/ProductsExt/SPSS/Documentation/SPSSforWindows/index.html#16

Statistics Canada & Organization for Economic Co-operation and Development (OECD) (2005). *Learning a living: First results of the adult literacy and life skills survey.* Paris: Minister of Industry, Canada, and OECD. http://www.oecd.org/edu/highereducationandadultlearning/41529631.pdf.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52(*4), 589-617.

Stout, W., Douglas, J., Junker, B., & Roussos, L. (1993). *DIMTEST Manual.* University of Illinois at Urbana-Champaign: Department of Statistics.

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-63). Hillsdale, NJ: Lawrence Erlbaum.

Received: May 5th, 2015
Accepted: September 16th, 2015