Psicothema

Psicothema 2013, Vol. 25, No. 3, 390-395
doi: 10.7334/psicothema2013.46

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Invariance levels across language versions of the PISA 2009 reading comprehension tests in Spain

Paula Elosua Oliden and Josu Mujika Lizaso
Universidad del País Vasco

Abstract

Background: The PISA project provides the basis for studying curriculum design and for comparing factors associated with school effectiveness. These studies are only valid if the different language versions are equivalent to each other. In Spain, the application of PISA in autonomous regions with their own languages means that equivalency must also be extended to the Spanish, Galician, Catalan and Basque versions of the test. The aim of this work was to analyse the equivalence among the four language versions of the Reading Comprehension Test (PISA 2009). **Method:** After defining the testlet as the unit of analysis, equivalence among the language versions was analysed using two invariance testing procedures: multiple-group mean and covariance structure analyses for ordinal data and ordinal logistic regression. **Results:** The procedures yielded concordant results supporting metric equivalence across all four language versions: Spanish, Basque, Galician and Catalan. **Conclusions:** The equivalence supports the estimated reading literacy score comparability among the language versions used in Spain.

Keywords: PISA, reading comprehension, testlet, equivalence, language.

Resumen

Evaluación de la invarianza entre las versiones lingüísticas de las pruebas de comprensión lectora PISA 2009 en España. **Antecedentes:** el proyecto PISA es la base de estudios y comparaciones sobre diseño curricular y factores de eficacia educativa, que solo son posibles si se garantiza la equivalencia entre las versiones idiomáticas de las pruebas. La aplicación de PISA en comunidades autónomas con lengua propia extiende el cumplimiento de la equivalencia a las versiones lingüística utilizadas en España: español, gallego, catalán y vasco. El objetivo de este trabajo fue analizar la equivalencia de la Prueba de Comprensión Lectora (PISA 2009) entre las 4 versiones idiomáticas. **Método:** tras definir el testlet como unidad de análisis se analizó la equivalencia entre versiones utilizando dos procedimientos de estudio de la invarianza: las estructuras de medias y covarianzas multigrupo para datos ordinales y la regresión logística ordinal. **Resultados:** los procedimientos arrojaron resultados concordantes que permiten avalar la equivalencia métrica entre versiones idiomáticas tanto con referencia al español, como entre los idiomas vasco, gallego y catalán. **Conclusiones:** la equivalencia respalda la comparabilidad de las estimaciones de competencia lectora entre las versiones lingüísticas utilizadas en España.

Palabras clave: PISA, comprensión lectora, testlet, equivalencia, idioma.

The information provided by the *Programme for International Student Assessment* (PISA) is an avenue for reflection in education, with a view to policy implementation and curriculum development and to studying the factors that influence educational effectiveness in participating countries. The conclusions from the PISA study are analysed both at the national and international levels by comparing information from different countries about the factors that influence specific competencies. Starting in 2003, the PISA project was expanded to include information at the regional level (the 2009 assessment included 14 autonomous regions in Spain), thereby contributing regional target populations and locally-based studies.

The comparisons and inferences drawn from them, whether inter-national or inter-regional, are based on the hypothesis of estimated score comparability; this hypothesis is valid if and only if the estimates represent equal or invariant assessment of a competency in all participating countries or regions. The hypothesis of comparability is fundamental to a project which involves 67 countries, 45 languages and 101 different test versions (OECD, 2012). The literature on intercultural studies and test adaptation (Hambleton, Merenda, & Spielberger, 2005; Matsumoto & Van de Vijver, 2011) warns that comparability can be affected by factors related to cultural, linguistic and curricular diversity between countries or regions, and by problems encountered in test adaptation. This means that language of administration is an aspect of the context of assessment that cannot be ignored (Dorans & Middleton, 2012).

In the context of cross-linguistic comparability these factors threaten the validity of intergroup comparisons and can be the origin of bias. Basically, there are three main types of bias: construct bias, method bias and item bias (Van de Vijver & Hambleton, 1996; Van de Vijver & Leung, 2011). Construct bias occurs when the measured construct shows significant differences between the original language for which it was developed and the

Received: February 21, 2013 • Accepted: May 17, 2013
Corresponding author: Paula Elosua Oliden
Facultad de Psicología
Universidad del País Vasco
20018 San Sebastián (Spain)
e-mail: paula.elosua@ehu.es

adapted language. Method bias refers to factors or issues related to the administration of the test that may affect the validity of the test. Item bias or differential item functioning (DIF) means that the item/construct relation is different among languages or cultures, due to poor item translations, or to culture/linguistic specific elements (Hambleton & Zeniski, 2011).

PISA is not alien to the problems of deficient adaptation. Thus, one of the main priorities is to make sure that the information is equally reliable and comparable between countries. In order to accomplish this, PISA implemented a double translation from two different source languages (French and English), and reconciliation by a third person, as well as translation/adaptation verification procedures (OECD, 2012). The applied practices are based primarily on guidelines for test translation/adaptation developed by the International Test Commission (Muñiz, Elosua, & Hambleton, 2013), and on external checks designed to meticulously evaluate linguistic quality and the format of tests and test items.

Even so, the equivalence of the PISA questionnaires is an assumption that does not always hold true. Recent research shows that the degree of invariance between the different language versions is not equivalent in content area or among languages. A number of studies have reported a higher degree of invariance among mathematics tests than reading or science tests, and higher equivalence among countries with Indo-European languages than countries whose languages belong to different language families. (Grisay, de Jong, Gebhardt, Berezner, & Halleux-Monseur, 2007; Grisay & Monseur, 2007). Oliveri & von Davier (2011) concluded that the fit of test items to the item response model (IRT) applied in the calibration/estimation process varied across countries. A significant factor related to cultural and curricular differences has also been found, as has an effect associated with language differences in countries in which the test is administered in more than one language (Monseur & Halleux, 2009). The few studies that compare invariance with reference to Spain have reported high degrees of equivalence, although some problematic items have been found in comparisons with the United Kingdom, (Elosua, 2006), the United States (Elosua, Hambleton, & Zenisky, 2006) and Mexico (Bully, Elosua, & Mujika, 2011).

There are no studies, however, that compare the structure of the different language versions of the test used in Spain. In 2009 the PISA tests were administered in five languages: Catalan, Basque, Spanish, Galician and Valencian. With the exception of Basque, all of them are Indo-European languages. Thus, the purpose of this study is to analyse the item equivalence levels among language versions of the PISA 2009 reading comprehension test. We did not focus on construct bias or on method bias since Reading Literacy and test formats are applicable to Spanish students (OECD, 2012).

Method

Participants

The sample of participants of Spanish nationality for the PISA 2009 edition included 25,647 students, 12,626 females and 13,019 males, of fourth-year secondary education. The test was administered in Basque to 1,167 students, in Catalan to 2,566 students, in Galician to 1538, in Spanish to 20,376 (Table 1) and in Valencian to 156 students (the latter language was not analysed given the sample size).

Instrument

PISA uses a matrix design in which items are arranged in clusters and placed in 13 different booklets. The priority competency for PISA 2009 was reading literacy (OECD, 2009). The reading comprehension tests consist of groups of items related to a single content area. Reading literacy was assessed via 29 reading units and a total of 101 questions related to the units. Some of the reading units were continuous (narration, description, exposition, argumentation, etc.) and others were discontinuous (charts, graphs, tables, diagrams, maps, forms, etc.). The items followed a multiple-choice format with dichotomous coding (*Correct*/*Incorrect* – 0/1), except for seven open-response items, which were coded on scores ranging from 0 to 2. As one of the original items was not administered to the Catalan population, it was removed from the study. The reading literacy scale has a mean of 500 and a standard deviation of 100.

Testlet. In the context of reading comprehension tests, dependent items that share a common text are reorganized as polytomous items, with scores ranging from 0 to a maximum equivalent to the number of items in the testlet. Each dependent items group is one testlet.

Procedure and data analyses

The first step in this study was to evaluate the local independence among items in order to define the unit of analysis. Differential item functioning methods were then applied.

Local independence and unit of analysis. The presence of groups of items related to a single content area can violate the principle of local item independence and yield misleading results in the application of psychometric models (Wainer & Lukhele, 1997); local independence must therefore be assessed prior to any analysis. Local independence was examined using the χ^2 statistic (Chen & Thissen, 1997; Hambleton, Swaminathan, & Rogers, 1991), which compares observed and expected response frequencies under the hypothesis of local independence. The analysis was conducted for each pair of items within each of the 29 groups of items, with responses based on the eight levels of reading literacy as measured by PISA (OECD, 2009).

Testlet definition. A testlet is a set of dependent items which are analysed as a unit (Wainer & Kiely, 1987; Wainer & Lewis, 1990; Wainer, Sireci, & Thissen, 1991). In this study, before forming the testlets, the seven open-response items were dichotomized, assigning a 1 to the 2-point scores, and a 0 to the 0- and 1-point scores. The dichotomization was used for two reasons: first, because the number of items affected was minimal (7 out of 101; 6% of the items) and second, because all items were thus given the same weight.

Item level equivalence. Item level equivalence was assessed according to the language of administration. Two differential item functioning (DIF) detection procedures were used (Mean and Covariance Structure Analysis (MACS) and Ordinal Logistic Regression). The first one is based on the linear factor model and evaluates factorial invariance. Factorial invariance means that the same measurement model fit across samples. The second one basically applies different regression models to each of the testlets to evaluate the effect of grouping variable (language) and the interaction of language/testlet after conditioning on the reading competence. Both methods were chosen to combine the advantages and avoid the disadvantages of each (Elosua & Wells, 2013). Theoretically, the MACS model is preferred because it

directly compares the factorial structure of the data. However, the MACS method has strong assumptions which are sometimes difficult to meet. The ordinal logistic regression is a less restrictive model-based method; it is flexible and overall works well spotting DIF items, but the parameter values are difficult to interpret. The reference sample included all of the students who took the test in Spanish, and the focal groups were defined by test language: Basque, Catalan and Galician.

Ordinal Logistic Regression. Ordinal logistic regression, or cumulative logistic regression, is an extension of the dichotomous logistic regression introduced by Swaminathan and Rogers (1990) (French & Miller, 1996). The dependent ordinal variable was defined as the score obtained in the testlet, and the predictor variable was defined as the reading literacy expected a posteriori (EAP). Two models were assessed for each testlet. The first is the baseline model, which includes only one independent predictor. The second adds two more parameters, the language of administration and the interaction between language and reading competency. After estimating both models, the difference is calculated between the -2log likelihood, which follows a χ^2 distribution with 2 degrees of freedom. An effect size measure is also found by computing the difference between the estimated R^2 for the two models. As a guideline for interpreting this measure, Jodoin and Gierl (2001) proposed a cutoff value of .07 for severe lack of invariance, and .03 for moderate differential functioning. Differential item functioning is concluded if the chi-square value is significant and the R^2 difference is great enough.

Multiple-group mean and covariance structure. Firstly, data for each sample was independently analysed using confirmatory factor analysis (CFA) in order to establish baseline unidimensional models and to estimate the reliability of the scores. Secondly, various levels of invariance were assessed progressively (Byrne, 2008) and jointly across the four language groups (Elosua & Muñiz, 2010). According to the linear factor model equation ($y = \upsilon + \lambda F + \sigma^2_e$), the measurement model for an observed variable (testlet; y) includes factor loadings (λ), error variances (σ^2_e) and intercepts (υ). Depending on the parameters which holds the invariance condition different levels of invariance can be defined. The simplest model is the configural invariance or equality of factor pattern matrices. By adding constraints to this model, it is possible to assess the equality of the loadings (metric invariance) and the equality of the intercepts (scale invariance). After assessing the configural invariance, the invariance of the intercept parameters was estimated; lastly, the invariance restriction was imposed on the response thresholds. The analysis model took into account the ordinal nature of the variables (Elosua, 2011) and used the robust weighted least squares estimator with adjustment for means and variance (WLSMV; Muthén, du Toit, & Spisic, 1997) employed in Mplus-6 (Muthén & Muthén, 2010). To compare the nested models two criteria were used simultaneously: the statistical significance of the likelihood ratio test ($p<.01$) and the changes in CFI values (Cheung & Rensvold, 1999).

Results

Descriptive statistics

The highest average reading scores were earned by the students who took the test in Galician ($M_{reading}$ = 486.77, *SD* = 83.89). The lowest average scores were found among the students who completed the test in Valencian ($M_{reading}$ = 452.22, *SD* = 71.93). Assessment of the statistical significance of differences concluded that the hypothesis of equality of the competency means related to testing language, $F_{reading}$ (4, 25830) = 7.09, $p<.001$, cannot be accepted. The sample size of students who completed the test in Valencian was too small to include it in subsequent analyses.

Local independence

Local item independence was examined using 1104 two-way contingency tables. The hypothesis of local independence was rejected in 49% of the cases ($p<.01$). Accordingly, *testlets* were designed for each of the 29 reading units. The number of items in each testlet ranged from 1 to 5. One of the testlets contained only one item, two testlets had two items, 12 contained three items, 11 had four items, and three testlets contained five items.

Unidimensionality and reliability

Internal consistency was tested using the ordinal alpha coefficient (Elosua & Zumbo, 2008). The goodness-of-fit indexes (CFI) for the Catalan (CFI = .923), Galician (CFI = .962) and Spanish (CFI = .961) samples were greater than .9. The Basque sample showed a slightly lower index (CFI = .88). The RMSEA values were optimal across all groups; none of them exceeded the cutoff point of .06 (Hu & Bentler, 1999). The internal consistency coefficients were greater than .9 in the four samples assessed (Table 1).

Ordinal logistic regression

Logistic regression models were estimated for each of the 29 reading units; the Spanish reference sample was compared with the Basque, Catalan and Galician focal groups. Although the chi-square values obtained for some of the comparisons were significant (Table 2), the effect size associated with the language did not reach the preset limit ($R^2_{Mod2-Mod1}$ = .07) in any of the comparisons.

Multiple group mean and covariance structure

Progressive assessment of invariance began with the configural invariance model. The goodness-of-fit values (CFI = .958; RMSEA = .031) supported the baseline invariance model. With restrictions added on the regression coefficients, the data was tested against the metric invariance hypothesis. Although the difference in chi-square values between the configural and metric models was statistically significant, $\chi^2(65) = 122$, $p<.001$, the CFI did not change substantially. The scale invariance was assessed by placing restrictions on the response thresholds. The difference in

Table 1
Descriptive statistics, unidimensionality and internal consistence

Group	N	M	SD	χ^2	df	CFI	RMSEA	ordinal α
Spanish	20401	485.71	87.80	3007*	168	.961	.029	.958
Basque	1168	481.83	75.74	448*	137	.879	.044	.942
Catalan	2570	482.31	84.58	722*	139	.923	.040	.954
Galician	1540	486.77	83.89	322*	131	.962	.031	.958

* significant values p<.01

chi-square values between this model and the previous one was significant, $\chi^2(148) = 939$, $p<.001$; However, the CFI value showed that the differences across the four versions were scale invariant.

Table 2
Ordinal logistic regression

Testlet	Spanish/Basque		Spanish/Catalan		Spanish/Galician	
	$G^2_{Mod2-Mod1}$	$R^2_{Mod2-Mod1}$	$G^2_{Mod2-Mod1}$	$R^2_{Mod2-Mod1}$	$G^2_{Mod2-Mod1}$	$R^2_{Mod2-Mod1}$
R055	9.25	.0008	7.83	.0006	9.11	.0008
R067	23.47*	.0026	5.10	.0005	2.95	.0003
R083	7.25	.0007	1.22	.0001	1.76	.0002
R101	33.99*	.0031	1.37	.0001	5.48	.0005
R102	24.21*	.0026	.53	.0001	6.52	.0007
R104	1.54	.0002	15.46	.0019	4.47	.0006
R111	2.37	.0002	17.39*	.0015	1.15	.0001
R219	5.16	.0009	4.31	.0007	10.26	.0017
R220	5.31	.0004	1.93	.0002	.24	.0000
R227	127.83*	.0125	33.02*	.0031	4.05	.0004
R245	8.15	.0009	5.38	.0006	.38	.0000
R404	31.73*	.0025	.05	.0000	11.57	.0009
R406	7.09	.0007	37.72*	.0037	1.13	.0001
R412	26.54*	.0028	4.55	.0005	10.63	.0011
R414	12.56	.0011	4.09	.0003	13.50	.0011
R420	63.04*	.0057	2.91	.0002	3.97	.0003
R424	7.88	.0008	2.64	.0003	8.41	.0009
R432	3.73	.0003	20.26*	.0017	4.68	.0004
R437	.24	.0000	53.49*	.0060	4.06	.0005
R442	10.73	.0008	2.61	.0002	2.89	.0002
R446	18.93*	.0025	3.67	.0005	.52	.0001
R447	1.35	.0001	2.09	.0002	.08	.0000
R452	71.89*	.0060	3.60	.0003	5.60	.0005
R453	1.70	.0001	41.23*	.0034	9.84	.0008
R455	18.59*	.0018	40.29*	.0036	9.56	.0009
R456	5.79	.0008	14.98	.0018	13.24	.0017
R458	2.62	.0002	1.70	.0001	3.79	.0003
R460	32.55*	.0033	24.72*	.0023	12.48	.0012
R466	2.36	.0002	47.56*	.0040	5.16	.0004

* $p<.01$

Table 3
Progressive assessment of factorial invariance

Model	Goodness-of-fit indexes				Difference test	
	χ^2	df	CFI	RMSEA	χ^2	df
Configural invariance	4069*	569	.956	.030		
Metric invariance	3444*	542	.961	.029	122*	65
Scale invariance	4049*	637	.959	.029	939*	148

* $p<.01$

Discussion

In a multilingual context in which autonomous regions enhance the PISA study by contributing their own sample groups, the aim of this research was to study one of the basic hypotheses underpinning the comparability of PISA results: item level equivalence. Given that in Spain PISA is administered in five languages, the purpose of this work was to assess the equivalence among the language versions used in the 2009 edition of PISA to assess reading literacy. The reference sample was the group that completed the test in Spanish. The focal groups consisted of students who took the test in the Basque, Catalan, and Galician language versions. The peculiarity of the reading comprehension tests, in which a set of dependent items was designed for each reading unit, made it necessary to first assess local item independence. After the hypothesis of independence was rejected, the testlet was defined as the unit of analysis. The items designed for each of the 29 texts in the reading comprehension test were then converted to 29 polytomous variables. Two methods to assess invariance were applied, ordinal logistic regression and multiple-group mean and covariance structure models. By using more than one procedure, cross information can be gathered to support the results obtained. Ordinal logistic regression was applied to pairs, using the Spanish language as the reference sample. Equivalence across the four versions could be assessed simultaneously with multiple-group mean and covariance structure models, offering information for all possible comparisons. This characteristic extends the generalization of results to inter-linguistic comparisons. The results obtained using both procedures were congruent and positive, supporting the hypothesis of estimated reading literacy score comparability among the Spanish, Basque, Catalan and Galician language versions, and between the Spanish version and the rest of the official languages.

The complexity and linguistic wealth attached to our social environment makes the testing language a variable to be controlled in every educational assessment process. The adaptation of tests and the verification of equivalence means that a check must be performed to ensure that no bias can invalidate comparisons between scores obtained in different language versions of the same test. If the internal structure of the tests was not equivalent in the different language groups, students with the same level of competence would obtain different scores. This would lead to erroneous conclusions in studies based on the hypothesis of equivalence between scores.

The relevance of inter-regional studies in a country made up of 17 regions, each with its own legislative autonomy, executive powers, and social, economic and even linguistic peculiarities, is clear. Among other aspects, the regions differ in terms of per capita income, gross domestic product, spending on education and even language. For example, in 2009 the per capita income in the Basque Country was 32,133 euros, while in Andalusia, the figure stood at only slightly more than half that amount: 18,507 euros (INE, 2009). In this differential context, comparisons tied to the PISA results or to any other educational assessment project are only valid if no bias is present in the instrument used. Few studies have been conducted in Spain which compare PISA results as a function of autonomous region. Ferrer, Valiente and Castel (2010) provide a description of the PISA 2006 results by region with reference to indicators related to the wealth index, the socioeconomic and cultural index, type of school and educational resources. Their study concludes that above and beyond regional differences, characteristics concerning type of school and directly associated with the students' social profile have a significant impact on results. Elosua (2013) analyses the relationship of the individual Index of Socioeconomic and Cultural Status (ESCS) and the regional ESCS

on reading comprehension in PISA 2009. The results showed that the estimation of variance components in reading comprehension is determined in part by the students' ESCS. They also showed that the different regional averages of this indicator of wealth do not significantly affect reading comprehension results and that the regression slopes are equivalent across the regions analysed, except for Ceuta and Melilla. Substantive studies such as those cited here are both important and basic for improving the education system; however, they rely on the measurement equivalence, a condition that must be evaluated.

The lack of studies such as those carried out in this work could affect any project which compares school effectiveness among regions with different languages; the effect would be even more extreme in the case of any regional study in which more than one language is spoken and different language versions of the same test are used. The 2009 PISA test was administered in two languages in the Basque Country, Navarre and the Balearic Islands. The Spanish and Basque versions were used in the Basque Country, and the Balearic Islands administered the test in Spanish and Catalan. In the presence of bias, the comparisons carried out in each of these regions could be called into question if statistical procedures are not used to adjust for differences between scores.

Considering these circumstances, it is important to have studies, such as the one presented in this article, which provide an in-depth analysis of the psychometric structure of the tests used. This kind of work delves into the origin of the differences and experts to develop measurement instruments that meet the conditions required by the goals of any assessment project.

Acknowledgements

This research was funded in part by the Spanish Ministry of Economy and Competitiveness (PSI011-30256) and by the University of the Basque Country (GIU12-32).

References

Byrne, B.M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20*, 872-882.

Bully, P., Elosua, P., & Mujika, J. (2011, marzo). Procedimientos de juicio y métodos empíricos en el estudio de la equivalencia psicométrica en PISA entre México y España [Judgemental procedures and empirical methods in the study of the psychometric equivalence in PISA between Mexico and Spain]. Paper presented at *6th International Congress on Psychology and Education*, Valladolid.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Cheung, G.W., & Rensvold, R.B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25,* 1-27.

Dorans, N.J., & Middleton, K. (2012). Addressing the extreme assumptions of presumed linkings. *Journal of Educational Measurement*, *49*, 1-18.

Elosua, P. (2006). Funcionamiento diferencial del ítem en la evaluación internacional PISA. Detección y comprensión [Differential item functioning on the international PISA assessment. Detection and understanding]. *Relieve, 12*, 247-259.

Elosua, P. (2011). Assessing measurement equivalence in ordered-categorical data. *Psicológica, 32*, 403-421.

Elosua, P. (2013). Diferencias individuales y autonómicas en el estatus socioeconómico y cultural como predictores en PISA 2009 [Individual and regional differences in socioeconomic and cultural status as predictors on PISA 2009]. *Revista de Educación*. Doi: 10.4438/1988-592X-RE-2013-361-236.

Elosua, P., Hambleton, R.K., & Zenisky, A. (2006). Improving the methodology for detecting biased test items. Paper presented at *5th International Conference of the International Test Commission*, Brussels, Belgium.

Elosua, P., & Muñiz, J. (2010). Exploring the factorial structure of the Self-Concept: A sequential approach using CFA, MIMIC and MACS models, across gender and two languages. *European Psychologist, 15*, 58-67.

Elosua, P., & Wells, C.S. (2013). Detecting DIF in polytomous items using MACS, IRT and ordinal logistic regression. *Psicológica, 34,* 327-342.

Elosua, P., & Zumbo, B.D. (2008). Coeficientes de fiabilidad para escalas de respuesta categórica ordenada [Reliability coefficients for ordered categorical response scales]. *Psicothema, 20*, 896-901.

Ferrer, F., Valiente, O., & Castel, J.L. (2010). Los resultados PISA-2006 desde la perspectiva de las desigualdades educativas [PISA-2006 results from the perspective of the educational inequalities]. *Revista Española de Pedagogía*, *245*, 23-48.

French, A.W., & Miller, T.R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315-332.

Grisay, A., de Jong, J.H.A.L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, *8*, 249-266.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the varios versions of an international test. *Studies in Eductional Evaluation, 33*, 69-86.

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

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. London: Sage publications Inc.

Hambleton, R.K., & Zenisky, A.L. (2011). Translating and adapting tests for cross-cultural assessments. In D. Matsumoto & F.J.R. van de Vijver (Eds.), *Cross-cultural research methods in psychology*. New York: Cambridge University Press (pp. 46-70).

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1-55.

Jodoin, M.G., & Gierl, M.J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.

Matsumoto, D., & Vijver, F.J.R. van de (2011). *Cross-cultural research methods in psychology*. New York: Cambridge University Press.

Monseur, C., & Halleux, B. (2009). *Translation and verification outcomues: National versions quality*. In OECD (Ed.), OECD Technical Report. PISA: OECD Publishing.

Muñiz, J., Elosua, P., & Hambleton, R.K. (2013). Segundas directrices de la ITC para la adaptación de tests [Second ITC guidelines for test adaptation]. *Psicothema*, *25*, 149-155.

Muthén, B., du Toit, S.H.C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Retrieved from URL:http://pages.gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf.

Muthén, L.K., & Muthén, B.O. (1998-2010). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

OECD (2009). *PISA 2009 assessment framework - key competencies in reading, mathematics and science*. PISA: OECD Publishing.

OECD (2012). *PISA 2009 Technical Report*. PISA: OECD Publishing.

Oliveri, M.E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Journal of Psychological Test and Assessment Modeling, 53*, 315-333.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Van de Vijver, F.J.R, & Hambleton, R.K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, *1*(2), 89-99.

Van de vijver, F.J.R., & Leung, K. (2011). Equivalence and Bias: A review of concepts, models and data analytic procedures. En D. Matsumoto & F.J.R. van de Vijver (Eds.), *Cross-cultural research methods in Psychology* (pp. 17-45). Cambridge: Cambridge University Press.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometric for testlest. *Journal of Educational Measurement, 27*, 1-14.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational & Psychological Measurement, 57*, 741-758.

Wainer, H., Sireci, S.G., & Thissen, D. (1991). Differential Testlet Functioning: Definitions and detection. *Journal of Educational Measurement, 28*, 197-219.