Psicothema

Psicothema 2014, Vol. 26, No. 3, 328-335
doi: 10.7334/psicothema2013.344

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Impact of family language and testing language on reading performance in a bilingual educational context

Paula Elosua Oliden and Josu Mujika Lizaso
Universidad del País Vasco

Abstract

Background: When different languages co-exist in one area, or when one person speaks more than one language, the impact of language on psychological and educational assessment processes can be considerable. The aim of this work was to study the impact of testing language in a community with two official languages: Spanish and Basque. **Method:** By taking the PISA 2009 Reading Comprehension Test as a basis for analysis, four linguistic groups were defined according to the language spoken at home and the test language. Psychometric equivalence between test forms and differences in results among the four language groups were analyzed. The comparison of competence means took into account the effects of the index of socioeconomic and cultural status (ISEC) and gender. **Results:** One reading unit with differential item functioning was detected. The reading competence means were considerably higher in the monolingual Spanish-Spanish group. No differences were found between the language groups based on family language when the test was conducted in Basque. **Conclusions:** The study illustrates the importance of taking into account psychometric, linguistic and sociolinguistic factors in linguistically diverse assessment contexts.

Keywords: PISA, testing language, family language, metric equivalence.

Resumen

Impacto del idioma familiar y del idioma del test sobre la comprensión lectora en un contexto educativo bilingüe. **Antecedentes:** en áreas en las que coexisten más de un idioma, o en aquellas en las que una persona habla más de un idioma, el efecto del lenguaje sobre los procesos de evaluación educativa o psicológica puede ser considerable. El objetivo del trabajo fue estudiar el impacto del idioma de administración de un test en una comunidad bilingüe con dos idiomas oficiales, el español y el euskera. **Método:** tomando como base del análisis la prueba Comprensión Lectora de PISA 2009, se definieron cuatro grupos lingüísticos en función del idioma familiar y del idioma del test. Se analizaron la equivalencia psicométrica entre las versiones idiomáticas y las diferencias entre los grupos. Los análisis tuvieron en cuenta el sexo y el índice de estatus socioeconómico y cultural (ISEC). **Resultados:** se detectó una unidad de lectura con funcionamiento diferencial. La comparación de los promedios de competencia lectora mostró resultados significativamente superiores en el grupo monolingüe español-español. Cuando el test se administró en euskera no se observaron diferencias entre los grupos lingüísticos en función de su idioma familiar. **Conclusiones:** el estudio ilustra la importancia de considerar los aspectos psicométricos, lingüísticos y sociolingüísticos en la evaluación en contextos de diversidad lingüística.

Palabras clave: PISA, idioma del test, idioma familiar, equivalencia métrica.

Educational and psychological tests differ in terms of content, format, scoring and the objectives and frameworks applied. However, one feature is common to all of them: the use of language. Language as a tool need not have a differential impact on assessment, and therefore, except in the assessment of verbal competence, the language level required for the test is assumed to be similar among students. This basic principle of linguistic homogeneity with regard to a command of the language may be violated in linguistic diversity contexts.

In linguistic diversity contexts in which the language of instruction and the family language differ, it is important to correctly define the testing language. In this respect, psycholinguistics refers to the need to draw a distinction between basic interpersonal language skills (BICS) and cognitive academic language proficiency (CALP; Abedi, 2009; Cummins, 1981, 2000; Hakuta, Butler, & Witt, 2000). Acquiring the former does not entail gaining a simultaneous command of the language on an academic level. The validity of the assessment may be compromised if cognitive academic language proficiency-related competences fail to attain the minimum level required to avoid any irrelevant variance (AERA, APA, & NCME, 1999). Studies conducted in the United States comparing the performance between students whose family language is English and immigrants for whom English is their second language (English Learning Students, ELS) have found problems in the assessment of the ESL students (Abedi, 2009, 2010; Abedi & Lord, 2001; Solano-Flores & Trumbull, 2003).

While it is important to take students' linguistic competence into account in linguistic diversity assessment contexts, it is

Received: December 19, 2013 • Accepted: March 19, 2014
Corresponding author: Paula Elosua Oliden
Facultad de Psicología
Universidad del País Vasco
20018 Donostia - San Sebastián (Spain)
e-mail: paula.elosua@ehu.es

also necessary to ensure psychometric equivalence between language versions. The scenario defined by conducting tests and questionnaires in contexts of language diversity raises the question of metric equivalence. It is known that adapting and using tests in different language contexts can result in changes to the psychometric properties (Muñiz, Elosua, & Hambleton, 2013; Hambleton, Merenda, & Spielberger, 2005; van der Vijver & Tanzer, 1997). Equivalence with the original test would be guaranteed by measurement invariance. When invariance exists, subjects belonging to different language groups with the same competence level obtain the same expected observed mean score.

Within this context of linguistic diversity, the aim of this work was to study in depth the variables related to the testing language. Taking into account that a) being multilingual is normal for most of the world today, b) linguistic diversity is a common and growing phenomenon, and c) safeguarding this diversity is one of the most urgent challenges facing our world (UNESCO, 2003), fairness in testing needs to consider this diversity. This study is focused on a historically bilingual environment in which Basque and Spanish coexist. Basque, or Euskera as it is known in the Basque language, is an isolated minority language spoken in the northern part of the Iberian Peninsula and the south-west of France, that is, in the Autonomous Community of Navarra (ACN), the Basque Autonomous Community (BAC), and the southern region of the Atlantic Pyrenees in France. Basque is the sole surviving non-Indo-European language in Western Europe. The number of speakers stands at around 700,000. Together with Spanish, it is an official language. The Basque Education System is bilingual, with Basque and Spanish as languages of instruction. Students may choose to be taught in either language.

In this bilingual environment education authorities need to decide in which language to conduct educational assessment programs: the language of instruction or the family language (mother tongue). Although it may seem trivial the decision is not simple. In international educational assessment programs, such as PISA or TIMMS, students in the BAC take the tests in their family language (ISEI-IVEI, 2004, 2009, 2011), but for regional diagnosis assessment programs the test language is the language of instruction. The results of comparing achievement in non-linguistic competences as a function of the language are inconclusive and even contradictory. Comparing the PISA results obtained by students whose family and instructional language is Basque with students whose family language is Spanish but language of instruction is Basque generally shows similar performance for both groups (ISE-IVEI, 2004, 2011). But in TIMMS 2007 (ISEI-IVEI, 2009) the mean proficiency of the students enrolled in the Basque educational system was higher for those who took the test in Basque. In general, the results show better performance when the students' home language and test language are the same.

However, given the complexity of the Basque-Spanish bilingual context, these studies can be confusing because they do not take into account the multidimensional nature of the assessment, and therefore the psychometric dimension (measurement invariance) could explain the differences in the results. In the framework of the OECD Programme for International Student Assessment (PISA) research headed by Grisay (Grisay & Monseur, 2007; Grisay, De Jong, Gebhardt, Berezner, & Halleux-Monseur, 2007) reveals problems related to psychometric equivalence between language versions and greater levels of non-equivalence in countries where non-Indo-European languages are spoken. Monseur and Halleux (2009) found an effect associated with language differences in countries where PISA is conducted in more than one language. Studies in equivalence in bilingual contexts carried out in Spain have evidenced differential item functioning associated with language (Elosua, López, Egaña, Artamendi, & Yenes, 2000; Elosua, López, & Egaña, 2000; Ferreres, González, & Gómez, 2000).

In this framework, the questions raised in this research focus on two points: (a) an analysis of psychometric equivalence between Spanish and Basque language versions in the Reading Comprehension tests used in PISA 2009, and (b) an in-depth study into the differences found in assessing reading competence according to family language and testing language. In order to achieve those objectives and given the relation between gender and reading comprehension and the relation between the index of socioeconomic and cultural status (ISEC) and performance, our analysis included these relevant variables (Chiswick & DebBurman, 2004; Coleman et al., 1996; Elosua, 2013; Feinstein & Symons, 1999; OECD, 2010).

Method

Participants

The sample was made up of 5,726 fifteen-year-olds (2,787 females and 2,939 males) from the PISA 2009 edition, who carried out the test in the Spanish communities in which Basque is the official language. Two factors were taken into consideration as criteria for inclusion: (a) that students should be Spanish, and (b) that Basque or Spanish should be spoken in their homes. Two groups of students were defined, depending on the responses gathered from the *student questionnaire*: those whose family language was Spanish (n = 4559) and those who spoke Basque at home (n = 1167). These students answered the PISA test either in Spanish or Basque. Table 3 shows the number of students in each of the language groups. The choice of testing language did not follow any consistent guideline applied to all students. In the region of Navarre, students sometimes answered in the language of instruction – which could be either Spanish or Basque – regardless of their family language (PISA databases contain no information about the language of instruction) and in the BAC the students answered in the family language.

Instruments

Reading Comprehension Test. PISA is an international study that was launched by the OECD in 1997. It aims to evaluate education systems worldwide every three years by assessing 15-year-olds' competencies in the key subjects: reading, mathematics and science. The priority competency for PISA 2009 was reading literacy (OECD, 2010). PISA 2009 used a matrix design in which items were arranged in clusters and placed in 13 different booklets. The Reading Comprehension tests consisted 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. 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. The reading literacy scale had a mean of 500 and a standard deviation of 100.

Data analysis

Two different methodological approaches were used to achieve the goals of this work; the first was psychometric, and the second was performed in the framework of the general linear model. The psychometric approach included the assessment of local independence, the definition of the testlet and the evaluation of the measurement equivalence. The aim of the second one was to perform a linguistic group comparison.

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 (Monseur, Baye, Lafontaine, & Quittre, 2011; Wainer & Lukhele, 1997). Local independence was examined using two different approaches: Yen's Q_3 statistic (Yen, 1984, 1993) and χ^2 statistic (Chen & Thissen, 1997; Hambleton, Swaminathan, & Rogers, 1991). In order to get the Q_3 matrix, the Generalized Partial Credit Model (GPCM) was fit to the data using an unconditional maximum likelihood factor analysis and the students' proficiency was estimated. A Yen's Q_3 correlation matrix was computed for each reading unit except for the R219 unit, which is associated with a single item. The χ^2 statistic, which is based on the co-variation of two-way contingency table, was estimated on each pair of items within each of the 29 reading units, with responses being conditioned on 8 levels of competence as reported by the PISA 2009 database (OECD, 2010).

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.

Measurement equivalence. Measurement equivalence was subsequently assessed according to the testing language, by using a model-based approach, multiple-group confirmatory factor analysis, and via ordinal logistic regression.

Multi-group-Confirmatory Factor Analysis (MG-CFA). Data for each sample was independently analyzed using confirmatory factor analysis in order to establish baseline models, and then measurement invariance was tested using MG-CFA. This model assesses factor invariance across groups by comparing the equality of parameters in the measurement model (Meredith, 1993; Sörbom, 1974). Different levels of invariance were defined depending on the number of parameters which hold the invariance condition across groups (same parameters). The simplest model was the configural invariance or equality of factor pattern matrixes. By adding constraints, the equality of the loadings (measurement invariance) and the equality of the intercepts (strong invariance) were assessed. The difference in the CFI indexes between two adjacent models was deemed to assess invariance; Cheung and Rensvold (2002) defined the .01 cutoff point for the difference between two 'nested' models. The analyses were carried out in the R environment (R Development Core Team, 2012) using the lavaan package (Rosseel, 2012).

Ordinal Logistic Regression (OLR) or cumulative logistic regression. Two different models are assessed within the context of DIF studies. The first one is the baseline model, which only includes one independent predictor, competence estimation. The second model adds two more parameters, the language parameter and the interaction between language and competence. After estimating both models, the likelihood ratio was evaluated. One measure of effect size was computed, the R^2 or generalized coefficient of determination. As a guideline for interpreting this measure of effect size, Jodoin and Gierl (2001) proposed a cutoff value of .07 for large DIF. Differential item functioning is concluded if the Chi-square value is statistically significant and the R^2 difference is great enough.

Group comparison. In the framework of the general linear model the mean performance in reading literacy was estimated by controlling the effects of gender and of the index of socioeconomic status. The assumption of variance homogeneity was evaluated and differences among linguistic groups based on the combination of test language and family language were estimated. The analyses were carried out in the R environment (R Development Core Team, 2012).

Results

Local independence

For each reading unit, the Yen's Q_3 correlation matrix was computed, resulting in a total of 138 pairwise correlations; 115 of the Q_3 values (83.33%) were positive; 79 values (57.25%) were between .00 and .09, and 26 (18.84%) of the correlation values were between .10 and .19. The remaining 10 values (7.25%) were greater than .20. Most of the estimated Q_3 values showed a degree of dependency between items. Moreover, the dependency level between item pairs in some reading units was substantial. The independence analyses based on χ^2 values were carried out by creating 1104 two-dimensional contingency tables. The local independence hypothesis was rejected on 45% of occasions (p<.01). Consequently, testlets were defined for each of the 29 sets of items. Each context dependent group of items was reorganized as a polytomous item, with scores on a testlet ranging from zero to the number of items in the group.

Proficiency estimator

Given that PISA 2009 student's proficiency is not based on testlets, a new performance indicator was estimated. The generalized partial credit model was fit to the testlet data using an unconditional maximum likelihood factor model, and the expected a posteriori estimators of each student's proficiency were obtained (Monseur, Baye, Lafontaine, & Quittre, 2011). The Pearson correlation between the PISA database EAP proficiency estimator and the testlet based proficiency estimator was .92 (p<.01).

Multi-group-Confirmatory Factor Analysis (MG-CFA)

Factor invariance was analyzed independently for each booklet (table 1). Given the sample size and the number of items, there were convergence problems in 4 of the 13 booklets when estimating the models, and invariance was not assessed. The means of the goodness-of-fit measures for the baseline model in the Basque sample were somewhat lower than those obtained in the Spanish sample (CFI_{Basque} = .90, $RMSEA_{Basque}$ = .08, $CFI_{Spanish}$ = .93 and $RMSEA_{Spanish}$ = .07

Table 1
Multi-Group confirmatory factor analysis

Model	N	χ^2	d.f.	RMSEA	CFI	Basque		Spanish	
						λ	υ	λ	υ
Booklet 2									
Basque baseline	86	385.51	78	.11	.76				
Spanish baseline	347	1417.72	78	.08	.88				
Configural Invariance		357.71	130	.09	.86				
Measurement Invariance		377.91	142	.09	.86				
Strong Invariance		431.97	154	.09	.83				
Free estimation R101		418.91	153	.09	.84		2.90		3.48
Free estimation R460		406.43	152	.09	.85		1.47		1.95
Booklet 4									
Basque baseline	80	322.85	66	.10	.84				
Spanish baseline	356	1157.45	66	.06	.94				
Configural Invariance		220.26	108	.07	.92				
Measurement Invariance		251.67	119	.07	.90				
Free estimation R227		237.14	118	.07	.91	1.48		.68	
Strong Invariance		303.71	129	.08	.87				
Free estimation R227		273.37	128	.07	.89		2.44		1.59
Free estimation R452		259.39	127	.07	.90		1.75		2.18
Booklet 5									
Basque baseline	82	181.27	28	.04	.99				
Spanish baseline	362	683.19	28	.07	.95				
Configural Invariance		73.96	40	.06	.96				
Measurement Invariance		89.07	47	.06	.95				
Strong Invariance		109.80	54	.07	.93				
Free estimation R442		99.98	53	.06	.94		3.47		3.08
Booklet 6									
Basque baseline	90	476.75	136	.04	.96				
Spanish baseline	340	1914.83	136	.04	.96				
Configural Invariance		327.77	238	.04	.96				
Measurement Invariance		347.61	254	.04	.96				
Strong Invariance		404.04	270	.05	.94				
Free estimation R452		382.92	269	.04	.95		1.04		1.53
Booklet 7									
Basque baseline	76	198.29	28	.14	.82				
Spanish baseline	360	711.45	28	.10	.90				
Configural Invariance		136.74	40	.11	.89				
Measurement Invariance		142.97	47	.10	.89				
Strong Invariance		203.25	54	.11	.83				
Free estimation R101		166.13	53	.10	.87		1.79		2.95
Free estimation R420		149.86	52	.09	.89		3.43		2.99
Booklet 8									
Basque baseline	86	154.39	28	.07	.94				
Spanish baseline	347	561.33	28	.10	.86				
Configural Invariance		121.46	40	.10	.88				
Measurement Invariance		125.48	47	.09	.88				
Strong Invariance		195.85	54	.11	.79				
Free estimation R227		143.72	53	.09	.86		2.79		1.89
Free estimation R420		131.67	52	.08	.88		3.10		2.65
Booklet 9									
Basque baseline	87	168.85	28	.05	.97				
Spanish baseline	357	668.35	28	.05	.98				
Configural Invariance		58.43	40	.05	.98				
Measurement Invariance		65.72	47	.04	.98				
Strong Invariance		84.33	54	.05	.96				
Free estimation R220		77.02	53	.05	.97		1.77		2.15
Booklet 11									
Basque baseline	94	274.78	36	.06	.96				
Spanish baseline	348	834.85	36	.06	.95				
Configural Invariance		99.14	54	.06	.96				
Measurement Invariance		106.34	62	.06	.96				
Strong Invariance		138.30	70	.07	.93				
Free estimation R227		124.71	69	.06	.95		2.28		1.84
Booklet 13									
Basque baseline	87	317.38	66	.10	.83				
Spanish baseline	362	1158.05	66	.07	.93				
Configural Invariance		234.33	108	.07	.91				
Measurement Invariance		249.88	119	.07	.90				
Strong Invariance		288.61	130	.07	.88				
Free estimation R227		273.15	129	.07	.89		2.18		1.70

Measurement invariance was achieved in all booklets except in number 4, in which the discriminatory parameter for item R227 had to be freely estimated in each group. Progressive assessment of invariance continued with strong invariance. The parameters from two testlets in booklet 2 (R101 and R460), in booklet 4 (R227 and R452), in booklet 7 (R101 and R420) and in booklet 8 (R227 and R420) were not constrained. One testlet was freely estimated in each group from 5 of the remaining booklets (see table 1). In total, intercepts from 13 testlets were estimated freely. Of these, 6 obtained higher estimates in the case of the Spanish version and 7 evidenced higher parameters in the Basque version. Among the 13 testlets, R227 and R452 were systematically detected in all the analysed booklets in which they appeared.

Ordinal Logistic Regression

Results from the application of ordinal logistic regression on 29 testlets are summarized in table 2. The language effect was statistically significant (p<.01) in 12 of the 29 units analyzed, although only one of them reached the size of the pre-set effect as the cutoff point (R227; $R^2_{\text{Mod2-Mod1}}$ = .05).

Results for linguistic groups

The homogeneity of the variances across the interaction of the variables included in the model (gender, home language, test language, and index of socioeconomic status) was assessed using Levene's test, $F(3780, 1904) = .69$, p>.05. The independence among the covariate ISEC and the groups based on the home language and on the test language, and between the ISEC and gender was also assessed. None of the main effects were statistically significant; all ps were bigger than .05, $F_{\text{gender}}(1, 5683) = 2.71$, p = .95; $F_{\text{testlang}}(1, 5683) = 2.59$, p = .10; $F_{\text{homelang}}(1, 5683) = 2.64$, p = .10. An ANCOVA including interaction term between home language and test language revealed no main effects of family language, $F(1, 5679) = .33$, p = .56, or test language, $F(1, 5679)$ = 0.27, p = .60, but did reveal interaction between these variables, $F(1, 5679) = 14.00$, p<.01. As predicted, the effects of gender, $F(1, 5679) = 332.70$, p<.01, and ISEC, $F(1, 5679) = 556.18$, p<.01, were statistically significant.

In order to evaluate the association between the linguistic variables and the outcome, a multiple regression model was fit to the data. The independent variables were ISEC, gender and a new variable consisting of a combination of family language and testing language. The parameter estimates of the model reflect a positive relationship between ISEC and reading literary (b_{ISEC} = 23.33, t = 23.58). The estimated parameter for gender (b_{gender} = -34.55, t = -18.24) showed a negative impact for males. Using the Basque-Basque group as a reference, the estimate coefficients were statistically significant for the Spanish-Spanish group (b = 23.06, t = 8.53), and they did not reach statistical significance for the Basque-Spanish group (b = -2.76; t = -0.58) or for the Spanish-Basque group (b = 2.42, t = .52). The post-hoc comparisons among linguistic groups were carried out using Tukey's honestly significant difference (HSD) test and the standardized mean differences between group-pairs were calculated (see table 4).

The monolingual Spanish-Spanish group systematically obtained higher means than the other bilingual groups and the monolingual Basque-Basque group (p<.01). The standardized mean differences values for these groups with regard to the rest of the groups were .26, .30 and .35; the highest one being between the Spanish-Spanish group and the group whose testing language was Basque and home language was Spanish (Basque-Spanish). No statistically significant differences were noted between the

Table 2
Results from analyses on DIF. Ordinal Logistic Regression

Testlet	G^2_{Mod1}	R^2_{Mod1}	G^2_{Mod2}	R^2_{Mod2}	$\Delta G^2_{\text{Mod2-Mod1}}$	$\Delta R^2_{\text{Mod2-Mod1}}$
R055	883.49	.42	890.87	.42	*7.38	<.01
R067	651.22	.34	651.52	.34	0.30	<.01
R083	798.70	.39	800.27	.39	1.56	<.01
R101	792.34	.38	825.38	.39	*33.04	.01
R102	661.86	.35	684.57	.36	*22.70	.01
R104	386.65	.23	389.11	.23	2.45	<.01
R111	969.85	.46	974.93	.46	5.07	<.01
R219	312.20	.25	317.01	.25	4.81	<.01
R220	998.04	.45	1005.03	.45	6.99	<.01
R227	610.12	.31	737.18	.36	*127.06	.05
R245	547.35	.32	552.27	.32	4.92	<.01
R404	1219.30	.51	1239.36	.52	*20.07	.01
R406	651.48	.33	653.98	.33	2.50	<.01
R412	682.46	.34	697.25	.35	*14.79	.01
R414	936.09	.44	948.34	.44	*12.25	<.01
R420	826.41	.40	868.10	.42	*41.70	.01
R424	709.16	.36	712.41	.36	3.25	<.01
R432	894.79	.44	896.18	.44	1.39	<.01
R437	424.02	.24	424.40	.24	0.39	<.01
R442	1217.15	.52	1232.80	.52	*15.65	.01
R446	341.59	.23	351.57	.23	*9.98	.01
R447	905.76	.43	906.07	.43	0.30	<.01
R452	970.79	.45	1038.71	.47	*67.92	.02
R453	830.24	.40	830.83	.40	0.59	<.01
R455	791.26	.38	802.43	.38	*11.18	<.01
R456	449.69	.28	456.29	.29	6.59	<.01
R458	732.20	.37	733.83	.37	1.62	<.01
R460	687.99	.36	714.11	.37	*26.12	.01
R466	885.85	.43	889.89	.43	4.04	<.01

Note: The numbers with an asterisk ("*") are statistically significant results (p<0.01)

Table 3
Groups of students according to family language and test language

Group	Test	Family	N	Mean	SD
1	Basque	Spanish	312	479.55	72.67
2	Spanish	Basque	332	486.51	76.91
3	Basque	Basque	835	484.01	75.86
4	Spanish	Spanish	4247	506.71	77.09

Table 4
Statistical significance of differences between linguistic groups

Group Comparison (Test Language - Family Language)		Tukey	t	p	Cohen's d
EspañoBasque - Spanish	Basque-Basque	-2.76	-0.58	.93	-.05
Spanish-Basque	Basque-Basque	2.42	0.52	.95	.03
Spanish - Spanish	Basque-Basque	23.06	8.53	<.01	.30
Spanish -Basque	Basque -Spanish	5.19	0.91	.78	.09
Spanish - Spanish	Basque -Spanish	25.83	6.15	<.01	.35
Spanish - Spanish	Spanish -Basque	20.67	5.05	<.01	.26

rest of the groups (p<.05). The standardized mean differences for those comparisons were close to 0.

Discussion

The aim of this work was to study the relationship between the testing language/family language and estimated reading competence in the PISA 2009 edition within a bilingual context in which Basque and Spanish coexist. Two different approaches were followed to achieve this goal. The first was a psychometric approach to assess the measurement invariance between the language versions of the test. The second was a statistical approach which, after controlling the effects of the gender and of the index of socioeconomic status, compared the mean performance of the linguistic groups defined in terms of family language and test language.

The psychometric study of the Spanish and Basque versions of the Reading Comprehension test showed a high level of equivalence. Given the characteristics of the Reading Comprehension test and the lack of the local independence among items depending on the same reading passage, testlets were defined as the unit of analysis. Multi-group confirmatory factor analysis was carried out on each of the PISA 2009 booklets. This study detected one testlet (R227) with different parameters in the four booklets in which it appeared. In all of them, the location parameter was greater in the case of the Basque-speaking sample, which in applied terms means that the testlet proved more difficult for those who did the test in Basque among students with the same level of reading competence. Testlet R452 also evidenced different parameters in the two samples and in the two analysed booklets in which it appeared. In both cases, the intercept parameter was greater in the case of the Spanish group. 13 parameters were freely estimated in each language version, although none of them had to be systematically freed in the booklets in which they appeared except for those described above. The ordinal logistic regression study as applied to each of the 29 testlet detected a single problematic testlet, with moderate differential item functioning (ΔR^2>.03). Special mention should be made of the concordance between both procedures in detecting this testlet. By adopting a conservative criteria in which one testlet is deemed to function differentially if it is simultaneously detected by more than one procedure (Fidalgo, Ferreres, & Muñiz, 2004), the conclusion drawn from this invariance study is that the R227 testlet evidences differential item functioning. It would therefore be of interest to pinpoint the origin of the problematic item; however, PISA databases have not released this information.

The study of differences in reading competence was carried out in the framework of the general linear model. Two important variables which impact the performance were statistically controlled: the ISEC and gender. As expected, their regression weights on the outcome variable were statistically significant (b_{ISEC} = 23.33, b_{gender} = -34.55). These results have been reported by previous studies (OECD, 2010, 2014). Therefore, it would not be correct to carry out group comparisons without accounting for those variables. In terms of linguistic group comparison it was observed that the results obtained from the monolingual Spanish-Spanish group, the group whose testing language and family language are Spanish, were significantly higher than the other bilingual groups and the monolingual Basque-Basque group. The standardized mean differences were about .30 for all of the comparisons. The results were consistent with previous studies, which drew the same conclusions (Elosua, López, Egaña, Artamendi, & Yenes, 2000; ISEI-IVEI, 2004). The relationship between family language and testing language within the bilingual Basque/Spanish context when reading competence is assessed has made it clear that only when the testing language is Spanish is the relationship between family language and competence statistically significant.

Attention should be particularly drawn to the fact that no differences between the bilingual samples have been noted according to the test language used. Within a context in which the language that maximizes student performance is discussed, this

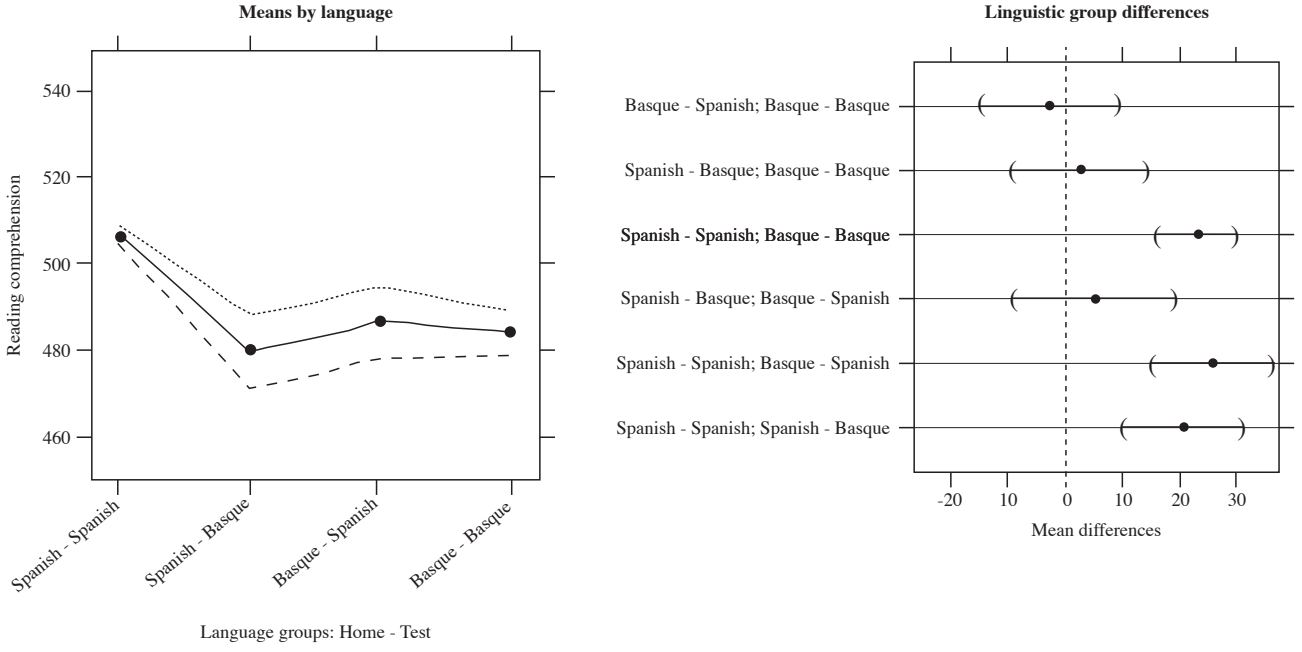

Figure 1. *Mean competence and linguistic groups differences (Test Language - Family Language)*

is a major result. Students taught in Basque within the Basque/Spanish educational landscape who do not speak this language at home achieve performance levels equivalent to those whose family language is Basque. Yet it has been explicitly shown that the means for the group whose family language is Basque but who does the test in Spanish is equivalent to that obtained by students who speak Basque at home and are tested in the same language. Equivalence between levels of competence obtained by the bilingual groups implies that students taught in Basque whose family language is Spanish end up gaining a command of the academic language in the sense defined by Cummins (1981) equivalent to that of their colleagues whose family language is Basque.

In order to put the results in context and to explain the differences found among the linguistic groups, it is important to remember that although no bias was found in the reading comprehension test, there are many factors that can explain the results. The linguistic characteristics of Spanish and Basque are different; they belong to different linguistic families, but there are also sociolinguistic differences between the two; language status and language prestige are not equal for Spanish and Basque; everyone speaks Spanish, but not everyone speaks Basque; Spanish has a long written history and Basque does not. Given these differences and in order to achieve score comparability, it would be important to model the effect of these variables in the assessment of performance.

The complexity and linguistic wealth attached to the actual social environment makes the testing language a variable to be controlled in educational assessment, either owing to the language proficiency required of students or to the problem of psychometric equivalence between versions and use. The use of questionnaires within contexts of linguistic diversity in which the family language and the testing language may differ demands that (a) a decision be taken as to testing language, (b) the tests be adapted to the language to maximize the validity of scores of each student being assessed, and (c) their psychometric equivalence be examined. None of the three questions should be trivialised.

The results shown in this work are important from an educational and psychometric standpoint: they reinforce the need to contextually study the impact associated with language in greater depth in the search for factors that, on an individual, school, community, psychometric or social level, may influence the indicators generated in these educational assessments.

It is important to remember that the results of this work have to be interpreted in the context of the linguistic diversity in which the study has been carried out. Linguistic diversity is a common phenomenon, but linguistic diversity contexts must be differentially analyzed. The sociological context of this work is defined by the existence of two official languages with different social status that belong to different linguistic families and have different literary traditions. This complexity of factors affecting performance should be considered in any educational assessment program in order to have reliable and valid outcomes in linguistic diversity contexts.

Acknowledgements

This work was financed by the Spanish Ministry of Economy and Competitiveness (PSI2011-30256) and by the University of the Basque Country (GIU12-32).

References

Abedi, J. (2009). Validity of Assessment for English Language Learning Students in a National/International Context. *Estudios sobre Educación, 16*, 167-183.

Abedi, J. (2010). *Performance Assessment for English Language Learners*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Abedi, J., & Lord, C. (2001). The language factor in mathematics. *Applied Measurement in Education*, *14*, 219-234.

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: APA.

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. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Chiswick, B.R., & DebBurman, N. (2004). Educational attainment: Analysis by inmigrant generation. *Economics of Education Review*, *23*, 361-379.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mod A.M., et al. (1996). *Equality of Educational Opportunity*. Washington, DC: US Department of Health, Education & Welfare.

Cummins, J. (1981). Four misconceptions about language proficiency in bilingual education. *NABE Journal, 5*(3), 31-45.

Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Clevedon, England: Multilingual Matters Ltd.

Elosua, P. (2013). Diferencias individuales y autonómicas en el estatus socioeconómico y cultural como predictores en PISA2009 [Socieconomic and cultural individuals and automous differenes on the socioeconomic and cultural index as predictors on PISA2009]. *Revista de Educación, 361*, 646-664.

Elosua, P., López, A., & Egaña, J. (2000). Idioma de aplicación y rendimiento en una prueba de comprensión verbal [Application language and performance on a verbal comprehension test]. *Psicothema, 12*(2), 201-206.

Elosua, P., López, A., Egaña, J., Artamendi, J.A., & Yenes. F. (2000). Funcionamiento diferencial de los ítems en la aplicación de pruebas psicológicas en entornos bilingües [Differential item functioning in the application of psychological tests in bilingual environments]. *Revista de Metodología de las Ciencias del Comportamiento, 2*(1), 17-33.

Ferreres, D., González, V., & Gómez, J. (2000). Comparación del estadístico Mantel-Haenszel y la Regresión Logística en el funcionamiento diferencial de los ítems en dos pruebas de aptitud intelectual en un contexto bilingüe [Comparison of Mantel-Haenszel statistic and Logistic Regression in differential item functioning on two tests of intellectual aptitude in a bilingual context]. *Psicothema, 12*(2), 214-219.

Fidalgo, A.M., Ferreres, D., & Muñiz, J. (2004). Liberal and conservative differential item functioning detection using Mantel-Haenszel and SIBTEST: Implications of the type I and type II error rate. *The Journal of Experimental Education*, *73*(1), 23-39.

Feinstein, L., & Symons, J. (1999). Attainment in secondary education. *Oxford Economic Papers*, *51*, 300-321.

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*(3), 249-266.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation, 33*(1), 69-86.

Hakuta, K., Butler, Y.G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* Santa Barbara: University of California Linguistic Minority Research Institute.

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

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

ISEI-IVEI (2004). *Influencia de la lengua de la prueba en los resultados de las evaluaciones [Influence of the test language on assessments results]*. Bilbao: ISEI-IVEI.

ISEI-IVEI (2009). *TIMMS 2007. Resultados en Matemáticas y Ciencias en el País Vasco [TIMMS 2007. Results in Mathematics and Science in the Basque Country]*. Bilbao: ISEI-IVEI.

ISEI-IVEI (2011). *PISA 2009 Euskadi. Informe de evaluación. Proyecto para la evaluación internacional de estudiantes de 15 años en lectura, Matemáticas y Ciencias [PISA 2009 Euskadi. Assessment Report. Project for international assessment of 15 years old students in reading, Math and Science]*. Bilbao: ISEI-IVEI.

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.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*(4), 525-543.

Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. In M. von Davier & D. Hastedt (Eds.), *IERI monograph series – Issues and methodologies in large scale assessments* (Vol. IV). Hamburg, Germany: IEA/ETS Research Institute (IERI).

Monseur, C., & Halleux, B. (2009). Translation and verification outcomes: National versions quality. In *OECD technical report* (pp. 96-104). Paris: OECD.

Muñiz, J., Elosua, P., & Hambleton, R.K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición [International Test Commission Guidelines for test translation and adaptation: Second edition]. *Psicothema, 25*, 151-157.

OECD (2010). *PISA 2009 results: What students know and can do - Student performance in reading, Mathematics and Science (Vol. 1)*. Paris: OECD.

OECD (2014). *PISA 2012 results: What students know and can do - Student performance in reading, Mathematics and Science (Vol. 1)*. Paris: OECD.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Viena, Austria: R Foundation for Statistical Computing.

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36.

Solano-Flores, G., & Trumbul, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English Language Learners. *Educational Researcher*, *32*(2), 3-13.

Sorböm, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.

UNESCO (2003). *Education in a multilingual world. UNESCO Education Position Paper*. Paris: UNESCO. Retrieved from: http://unesdoc.unesco.org/images/0012/001297/129728e.pdf.

van der Vijver, F.J., & Tanzer, N.K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 54*, 119-135.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometric for testlest. *Journal of Educational Measurement, 24*(3), 185-201.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational & Psychological Measurement, 57*(5), 749-766.

Wainer, H., Sireci, S.G., & Thissen, D. (1991). Differential Testlet Functioning: Definitions and detection. *Journal of Educational Measurement, 28*(3), 197-219.

Yen, W.M. (1984). Effects of local item dependence on the fit and equation performance of the three-parameter logistic model. *Applied Psychological Measurement, 2*, 125-145.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.