Psicothema

Psicothema 2015, Vol. 27, No. 3, 296-302
doi: 10.7334/psicothema2014.282

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Partial scalar invariance and observed differences across gender in a reasoning test battery

Paula Elosua and Josu Mujika
Universidad del País Vasco

Abstract

Background: The substantive basis of the Reasoning Test Battery (BPR) is the theory of the hierarchical organization of cognitive abilities and therefore, it combines a general cognitive factor and specific factors associated with abstract, numerical, verbal, practical, spatial and mechanical reasoning. The battery has three forms, covering an age range from 9 to 22 years. **Method:** The present study analyzes the internal structure of the Basque version of the battery using exploratory and confirmatory factor analyses. Factorial invariance studies across gender were performed and partial differences observed were analyzed in a sample of 1,923 students. **Result:** The results concluded: (a) the presence of one general reasoning factor in each of the forms, (b) partial scalar invariance across gender affecting mechanical reasoning and numerical reasoning, (c) no differences in the general reasoning factor, and (d) negligible observed differences in partial scales. **Conclusions:** Tests for measurement invariance indicate differences in factor intercepts, cautioning that comparisons of observed g scores across gender are not appropriate.

Keywords: Reasoning tests, factorial structure, factorial invariance, adaptation, gender differences, BPR, CHC.

Resumen

Invarianza escalar parcial y diferencias observadas entre sexos en una batería de tests de razonamiento. **Antecedentes:** construida sobre la teoría de la organización jerárquica de las habilidades cognitivas, la Batería de Pruebas de Razonamiento (BPR) combina un factor de razonamiento general y factores específicos asociados con el razonamiento abstracto, numérico, verbal, práctico, espacial y mecánico. La batería tiene 3 Formas que cubren un rango de edad entre 9 y 22 años. **Método:** se analizó la estructura interna de la versión en euskera de la batería por medio de análisis factoriales exploratorios y confirmatorios. Se llevaron a cabo estudios de invarianza factorial en función del sexo y se analizaron las diferencias observadas en las escalas parciales en una muestra de 1.923 estudiantes. **Resultados:** los resultados concluyeron: (a) la presencia de un factor de razonamiento general en cada una de las formas, (b) la invarianza escalar parcial que afectan a las escalas de razonamiento mecánico y razonamiento numérico, (c) la no diferenciación entre sexos en el factor general, y (d) diferencias mínimas en las escalas parciales. **Conclusiones:** los resultados del test de invarianza factorial apuntaron la presencia de valores interceptales diferentes, lo cual desaconseja la comparación de puntuaciones observadas g en función del sexo.

Palabras clave: test de razonamiento, estructura factorial, invarianza factorial, adaptación, diferencias de género, BPR, CHC.

Since the birth of scientific psychology, intelligence has been and continues to be one of the most widely studied psychological variables. Recent studies on the use of tests carried out both in Spain and other European countries have shown that the assessment of intellectual capabilities in educational, organizational and clinical contexts occupies a central place in professional practice (Elosua & Iliescu, 2012; Muñiz & Fernández-Hermida, 2010).

In intelligence modeling, psychology has played a role in the debate between defending the general factor (*g*) (Spearman, 1927) and considering aptitudes as differentiated autonomous structures (Thurstone, 1938; Guilford, 1967). Between the two divergent positions, a conciliatory model has gained ground: the hierarchical organization of cognitive abilities (Carroll, 2003; Cattell, 1963, 1971; Horn & Noll, 1997; McGrew, 2005; Schneider & McGrew, 2012; Vernon, 1961). The Cattell-Horn-Carroll (CHC) model proposes a three-stratum model of intelligence. Stratum I is defined by narrow-spectrum abilities that are related and grouped together to form the second strata, consisting of at least nine abilities: (a) fluid reasoning, Gf ; (b) comprehension-knowledge or crystallized intelligence, Gc; (c) cognitive processing speed, Gs; (d) reacting or decision-making speed, Gt; (e) short-term memory, Gsm; (f) long-term memory storage and retrieval, Glr; (g) reading and writing ability, Grw; (h) quantitative reasoning, Gq; (i) visual-spatial processing, Gv, and (j) auditory processing, Ga. These would be used to form the third stratum or general factor, g.

Using the hierarchical model, the Reasoning Test Battery ("Bateria de Provas de Raciocinio", BPR; Almeida & Lemos, 2006) was created. The BPR combines the assessment of a general reasoning with components associated with skills which are usually assessed in specific multifactorial intelligence batteries. The scales that make up the BPR measure: (a)

Received: December 16, 2014 • Accepted: May 17, 2015
Corresponding author: Paula Elosua
Facultad de Psicología
Universidad del País Vasco
20018 San Sebastián (Spain)
e-mail: paula.elosua@ehu.es

abstract reasoning associated with fluid intelligence, Gf; (b), verbal reasoning associated with comprehension knowledge or crystallized intelligence, Gc; (c) numerical reasoning, Gq; (d) special reasoning related to visual processing, Gv; (d) mechanical reasoning and practical reasoning associated with fluid reasoning, Gf, and reading comprehension, Grw. In terms of reasoning, the battery contains analogies, completion series and troubleshooting tasks. The item content consists of meaningless geometric figures (figurative-abstract), word meanings (verbal), number sequences (numerical), movement of cubes (spatial) and practical situations (concrete-mechanical).

Versions of the BPR have been adapted to Brazil (Primi & Almeida, 2000) and Spain (Elosua & Mujika, in press). The reliability coefficients reported for the scales are greater than.70 and, in most cases, the values are above .80. Studies conducted on the dimensionality of the BPR on the partial scales concluded the presence of a general factor which explains between 40% and 60% of the variance of the scores.

Since the early twentieth century, together with the hierarchical organization of cognitive abilities, analyzing the possible differences as a function of gender has been a recurring theme in psychology research (Fryer & Levitt, 2010; Johnson, Carothers, & Deary, 2008; Lemos, Abad, Almeida, & Colom, 2013; Lohman & Lakin, 2009), and the accumulation of evidence has kept the discussion going. The study of differences in observed means between men and women concludes the presence of differences in specific dimensions (Halpern et al., 2007). Girls tend to score higher in verbal abilities (Halpern, 1997; Lynn, Raine, Venables, Mednick, & Irwing, 2005; Johnson & Bouchard, 2007), and boys do better in spatial, abstract reasoning and numerical abilities (Colom, Quiroga, & Juan-Espinosa, 1999; Geiser, Lehman, & Eid, 2008; Linn & Petersen, 1985; Voyer, Voyer, & Bryden, 1995). However, there is less agreement when exploring the differences between sexes in the general reasoning factor. The conclusion that men achieve higher g factor scores than women (Lynn, 2002; Jackson & Rushton, 2006) contradicts studies that defend no sex differences in general intelligence (Dolan et al., 2006; van der Sluis et al., 2008).

The lack of agreement on differences in the general factor is accompanied by differences in methodological approach. The conclusions from research on gender differences are customarily based on observed score comparison; however, from a methodological perspective, comparing scores without accounting for the invariance of the hierarchical factorial structure of the questionnaire can lead to erroneous decisions. The lack of equivalence can mask differences and lead researchers to identify discrepancies where there are none (Finch & French, 2012). However, studies based on factor models continue to shed light on the subject. Irwing (2012) compared factor score estimates after adjusting a hierarchical factor model to the WAIS-III in which he concluded the presence of differences between males and females, with the males exhibiting higher values. But using the same test and studying the factorial invariance of the model in a Spanish and Dutch sample, a number of studies report no difference in the g factor (Dolan, Colom, Abad, Wicherts, Hessen, & van de Sluis, 2006; Van der Sluis et al., 2006). In this regard, and with respect to the BPR, the studies of factorial invariance published have concluded no sex differences in the general reasoning factor in Portuguese and Spanish samples (Lemos, Abad, Almeida, & Colom, 2013; Elosua & Mujika, in press).

Therefore, the objective of this work is twofold: the first is to analyze the internal structure of the BPR in a Basque sample, and the second, after studying the invariance of the factorial structure across sexes, is to take a closer look at the gender differences.

Methods

Participants

The sample comprised 1923 students, 985 females and 938 males. The age range of the students was 9 to 22 years. The mean age of the participants completing Form-1 was 10.32 years (SD = 0.98), Form-2, 13.51 years (SD = 0.97), and Form-3, 16.28 years (SD = 1.07). The distribution of the sample is shown in Table 1.

Instrument

The current version is a battery of tests (Table 2) in three different forms designed to assess a wide range of ages: BPR Form-1 consists of four scales aimed at students in grades 4, 5 and 6 (ages 9-12); BPR Form-2 comprises 5 scales and covers the first three years of secondary education (ages 12-15); BPR Form-3, with 5 scales, is designed for students enrolled in the fourth year of secondary education and the 2 pre-university years, known as the Spanish Baccalaureate or *Bachillerato* (ages 15-20).

The process of adapting the BPR to Basque followed the recommendations of the International Test Commission (Muñiz, Elosua, & Hambleton, 2013). The battery was adapted to the Basque using a double forward translation by two independent translators and reconciliation by a multidisciplinary team comprised of two psychometricians, two professional translators and two teachers of primary and secondary education. The expert committee reviewed the product through an iterative review process until a final version was adopted by consensus. Some items of the verbal section were modified to maintain semantic equivalence in terms of familiarity and difficulty: 5 items from Form-1, 5 items from Form-2 and 6 items from Form-3.

Table 1
Sample composition

Form	Females	Males	Total
BPR-1	320	343	663
BPR-2	331	351	682
BPR-3	334	244	578
Total	985	938	1923

Table 2
Structure of the reasoning tests nattery

		Abstract reasoning	Verbal reasoning	Spatial reasoning	Numerical reasoning	Practical reasoning	Mechanical reasoning
BPR-1	Items	20	20	–	15	15	–
BPR-2 BPR-3	Items	25	25	20	20	–	25
Tasks		Figurative analogies	Verbal analogies	Cube rotation	Numerical series	Problem solving	Problems

Procedure

The data were collected from an incidental sample comprising fifteen schools in the Basque Autonomous Community which signed cooperation agreements for the purposes of this research. The participants, or where applicable their legal guardians, signed an informed consent document. The questionnaires were administered by personnel specifically trained for this project.

Data analysis

The study of the internal structure was based on three complementary approaches: (a) the internal consistency of each scale was assessed with ordinal alpha; (b) the presence of the dominant factor for each partial scale was assessed using item factor analysis on the tetrachoric correlation matrix; and (c) the factorial structure for each form was analyzed by confirmatory factor analysis.

The study of gender differences was carried out in two stages: (a) factorial invariance across gender was analyzed for each of the forms. The measurement invariance examined the equality of factor pattern matrices (configural invariance), the equality of loading matrices (measurement invariance), the equality of intercepts (scalar invariance), and the equality of factor means; (b) differences in each of the partial scales were examined as a function of gender using the Student's *t*-test and Hedges' g to estimate the size of the effect.

The factorial analyses were conducted using the 'lavaan' package (Rosseel, 2012). Model fit was assessed with the chi-square statistic, the root mean square error of approximation (RMSEA), the standardized root mean square residual (SRMR) and the comparative fit index (CFI). Although Hu and Bentler (1999) suggested that RMSEA should be less than or equal to .06 for a good model fit, recent studies conclude that in models with small degrees of freedom, RMSEA too often indicates a poor fitting model (Kenny, Kaniskan, & McCoach, in press). The cut-off value for the CFI is usually fixed at .90 and for SRMR, at .08 (Hu & Bentler, 1999). Following the criteria proposed by Cheung and Rensvold (2002), the invariance is rejected if the value of the difference between the two nested models is higher than 0.01 in favor of the least strict model.

Results

Internal consistency

Internal consistency values of the scales ranged from .79 to .93 (see Table 3). The highest values in each form were obtained in the numerical scales (α_{Form1} = .92; α_{Form2} = .93; α_{Form3} = .88), and the lowest values were associated with the mechanical reasoning scales (α_{Form2} = .79; α_{Form3} = .81).

Unidimensionality of partial scales

Results of the item factor analyses in each of the scales (Table 3) showed that for numerical and practical reasoning scales, the percentages of variance associated with the one-dimensional factor were greater than .30. The mechanical reasoning scales showed slightly lower values; 16% of the variance was explained by the dominant factor in Form-2 and the explained variance was 19% in Form-3.

Factor structure

Table 3 presents the results of assessing the presence of a general reasoning factor. The values show the regression weights and their standard errors for each scale. The loadings in all three forms were statistically significant and greater than .55 except for the mechanical reasoning scale, which showed lower values (λ_{Form2} = .50; λ_{Form3} = .39). The SRMR and CFI indices (Table 4) met the criteria for a good fit in the three BPR test forms, and RMSEA showed values a little higher than the cut-off point in Form-2 ($RMSEA_{Form2}$ = .12). The extracted general factors explained 54% of the variance in Form-1, 38% in Form-2 and 37% in Form-3.

Gender factorial invariance

Results of the factorial invariance are shown in Table 5. The CFI indices obtained in evaluating baseline models had values greater than .93 in all three forms. The poorest fit was obtained by the male sample in Form-2 (CFI = .93; RMSEA = .13). Configural invariance models showed a reasonable fit to the data in the three forms with CFI values over .94. Metric invariance was held for Form-1, Form-2 and Form-3, where CFI difference values were lower than the cut-off point of .01. The results from the scalar invariance analysis were poor; changes in the CFI index in the three forms exceeded the critical value of .01 (CFI_{Form-1} = .998-.954 = .04; CFI_{Form-2} = .94-.91 = .03; CFI_{Form-3} = .98-.95 = .03). Following a sequential process of freeing parameters based on examining the modification indices (Elosua & Muñiz, 2010), the conditions for partial scalar invariance were defined for the three BPR forms. In Form-1, the parameter associated with the numerical reasoning scale was freed (ΔCFI_{Form-1} = -.001), and in Form-2 and Form-3, the intercepts of the mechanical reasoning test (ΔCFI_{Form-2} = 0; ΔCFI_{Form-3} = 0) were also freed. The parameters from the final models are shown in Table 6. The intercept values in the numerical and mechanical reasoning scales were systematically higher for the male group. After adjusting the partial invariance models, the

Table 3
Internal Structure of the BPR

	Form 1			Form 2			Form 3		
	α_{ORD}	Var%	λ	α_{ORD}	Var%	λ	α_{ORD}	Var%	λ
Abstract R.	.88	.29	.65	.89	.29	.55	.86	.25	.63
Verbal R.	.85	.23	.79 (.05)	.86	.22	.63 (.06)	.87	.22	.71 (.06)
Numerical R.	.92	.45	.70 (.05)	.93	.45	.65 (.06)	.88	.31	.54 (.05)
Practical R.	.92	.53	.79 (.05)						
Spatial R.				.88	.28	.75(.07)	.85	.25	.73(.06)
Mechanical R.				.79	.16	.50(.06)	.81	.19	.39 (.05)

Note: estimation errors in parentheses; α_{ORD} = Ordinal alpha

Table 4
Fit indexes of the confirmatory unidimensional model

	χ^2	df.	p	SRMR	CFI	RMSEA	$IC_{90\%}$ RMSEA
Form-1	3.80	2	.15	.01	.998	.037	.00 - .09
Form-2	52.87	5	<.01	.05	.93	.12	.09 - .15
Form-3	15.01	5	.01	.03	.98	.059	.02 - .09

Table 5
Models for gender factorial invariance

	Form-1				Form-2				Form-3			
Model	χ^2	**df**	**CFI**	**RMSEA**	χ^2	**df**	**CFI**	**RMSEA**	χ^2	**df**	**CFI**	**RMSEA**
Females. Base model	1.33	2	.998	.00	17.24	5	.96	.09	6.80	5	.99	.03
Males. Base model	1.57	2	.998	.00	3.39	5	.93	.13	6.21	5	.99	.03
Configural Invariance	2.90	4	.998	.00	47.62	10	.94	.11	13.01	10	.99	.03
Metric Invariance	6.67	7	.998	.00	53.28	14	.94	.10	22.66	14	.985	.05
Scalar Invariance	52.98	10	.954	.11	73.35	18	.91	.10	47.67	18	.949	.07
Partial Scalar Invariance (nr free)	11.81	9	.997	.03								
Partial Scalar invariance (nr and mr free)					53.49	16	.94	.08	27.76	16	.980	.05
Equal latent means	18.62	10	.991	.05	59.91	17	.935	.09	28.40	17	.980	.049

Note: nr = numerical reasoning; mr = mechanical reasoning

Table 6
Parameter estimates for final models

	BPR Form-1			BPR Form-2					
Scale	**λ**	**ν**	ν_{male}	**λ**	**ν**	ν_{male}	**λ**	**ν**	ν_{male}
Abstract R.	.079 (0.05)	12.81 (0.13)		1.00	15.01(0.14)		1.00	13.09(0.13)	
Verbal R.	1.00	11.98 (0.14)		1.24 (0.09)	14.40 (0.16)		1.39 (0.14)	16.04(0.16)	
Numerical R.	.903 (0.05)	7.82 (0.17)	9.12 (0.17	1.04 (0.09)	7.83 (0.23)	8.74 (0.23)	1.29 (0.12)	9.55 (0.18)	10.18 (0.21)
Practical R.	.825 (0.04)	8.99 (0.11)							
Spatial R.				1.36 (0.10)	11.52 (0.17)		1.57 (0.14)	10.77 (0.16)	
Mechanical R.				0.63 (0.07)	9.56 (0.18)	10.35 (0.21)	0.95 (0.10)	8.97 (0.15)	10.18 (0.22)

Note: Standard errors in parenthesis

equality of general factor means was analyzed. The results showed that once the mechanical and numerical reasoning parameters were freed, sex differences were absent in the latent factor ($\Delta CFI_{Form\text{-}1}$ = -.006; $\Delta CFI_{Form\text{-}2}$ = -.005; $\Delta CFI_{Form\text{-}3}$ = 0)

Observed mean scores by gender

Table 7 shows the observed means and comparison between each of the groups in the partial scales, as well as a measurement of the size of the effect.

Among the group of youngest students, aged 9-11, results revealed significant differences in the numerical reasoning scale (M_{Male} = 8.90; M_{Female} = 8.07; t(659) = -3.02; p = .003), and in the practical reasoning scale (M_{Male} = 8.55; M_{Female} = 9.33; t(651) = 3.42; p = .001). In the first case, mean scores were higher for males than for females; in the second case, the females reported better performance. However, based on Hedges' g, the size of the differences was small in both cases (g_{verbal} = -.23; $g_{practical}$ = .26). In the numerical reasoning scale, gender differences were no longer significant in Form-2, (M_{Male} = 8.44; M_{Female} = 8.05; t(653) = -1.16; p = .24) but were again significant among the older students (M_{Male} = 10.22; M_{Female} = 9.52; t(570) = -2.33; p = .02); the size of the effect was negligible in both cases (g_{Form2} = .09; $g_{Form\text{-}3}$ = .-.19).

As for verbal abilities, comparisons were significant in favor of girls in Form-2 (M_{Male} = 13.76; M_{Female} = 14.66; t(669) = 2.82; p = .005), although the size of the effect was small (g = .21). The rest of the comparisons were statistically non-significant; however, as expected, the values were higher for the female groups in both

Table 7
Observed scale means

BPR	Females		Males		Student's t	Hedges' g 95% CI
	Mean	**SD**	**Mean**	**SD**		
Abstract reasoning						
Form-1	13.00	3.29	12.59	3.60	t(654) = 1.52; p = .12	.11 [-03,.27]
Form-2	15.10	3.76	14.64	3.66	t(674) = 1.59; p = .11	.12 [-.02,.27]
Form-3	13.10	3.02	13.04	3.61	t(570) = 0.19; p = .84	.01 [-.14,.18]
Verbal reasoning						
Form-1	12.18	3.71	11.70	3.61	t(652) = 1.67; p = .09	.13 [-.02,.28]
Form-2	14.66	4.06	13.76	4.18	t(669) = 2.82; p = .005	.21 [.06,.37]
Form-3	15.39	3.71	15.21	3.97	t(565) = 0.54: p = .58	.05 [-.11,.28]
Spatial reasoning						
Form-2	11.75	4.27	11.02	4.47	t(664) = .15; p = .03	.16 [.01,.31]
Form-3	10.54	3.53	11.19	4.11	t(573) = -2.01: p = .04	-.16 [-.33,-.00]
Numerical reasoning						
Form-1	8.07	3.48	8.90	3.52	t(659) = -3.02; p = .003	-.23 [-.38,-.08]
Form-2	8.05	4.08	8.44	4.65	t(653) = -1.16; p = .24	-.09, [-.24,.06]
Form-3	9.52	3.49	10.22	3.57	t(570) = -2.33: p = .02	-.19 [-.36,-.03]
Mechanical reasoning						
Form-2	9.36	3.29	10.15	3.77	t(659) = -1.87; p = .06	-.14 [-.29,.00]
Form-3	8.98	2.94	10.18	3.59	t(567) = -4.34; p <.001	-.36 [-.52,-.19]
Practical reasoning						
Form-1	9.33	2.72	8.55	3.06	t(651) = 3.42; p = .001	.26 [.11,.42]

BPR forms ($M_{MaleForm1}$ = 11.70; $M_{FemaleForm1}$ = 12.18; $M_{MaleForm3}$ = 15.21; $M_{FemaleForm2}$ = 15.39).

In the mechanical reasoning scales, the boys scored higher than the girls in the two forms in which this skill is assessed ($M_{MaleForm2}$ = 10.15; $M_{FemaleForm2}$ = 9.36; $M_{MaleForm3}$ = 10.18; $M_{FemaleForm-2}$ = 8.98). The greatest differences were associated with the older students (g_{Form3} = -.36).

In the spatial reasoning test, which contains items requiring students to mentally rotate cubes, gender differences were observed ($p<.05$). Whereas the mean scores were higher among females in the 12-15 year-old group (M_{Male} = 11.02; M_{Female} = 11.75; $t(664) = 2.15$; $p = .03$), the results were the opposite among students in the older group, with males scoring higher than females (M_{Male} = 11.19; M_{Female} = 10.54; $t(573) = -2.01$; $p = .04$). In both cases, the differences were not of practical significance (g_{Form2} = .16; g_{Form3} = -.16).

In the abstract reasoning tests, the girls scored higher in all cases than the boys ($M_{MaleForm1}$ = 12.59; $M_{FemaleForm1}$ = 13.00; $M_{MaleForm3}$ = 14.64; $M_{FemaleForm2}$ = 15.10; $M_{MalesForm3}$ = 13.04; $M_{FemaleForm3}$ = 13.10), but the differences were of neither statistical ($p > .05$) nor practical significance (g_{Form1} = .11; g_{Form2} = .12; g_{Form3} = .01).

Discussion

Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests (AERA, APA, & NCMEA, 2014); given that the BPR, in its different forms, was constructed to assess general cognitive ability, the validation of the scores should provide arguments for the conceptual framework that supports the BPR, that is, the structure for reasoning.

In studying the internal structure of the three BPR forms, homogeneity was first assessed for each partial scale. The consistency indices ranged from .79, which was obtained for Mechanical Reasoning Form-2, to the highest reliability coefficients for the numerical reasoning scales in the three BPR forms (α_{Form1} = .92; α_{Form2} = .93; α_{Form3} = .88). These results agreed with the reliability studies carried out with the BPR (Almeida & Lemos, 2006; Baumgartl & Primi, 2006; Elosua & Mujika, in press). The high values of the numerical reasoning scales may be explained by the fact that the scales are the only ones that require a constructed response instead of multiple choice items. As some authors noted (Primi, Couto, Almeida, Guisande, & Miguel, 2012; Primi, Rocha da Silva, Rodríguez, Muniz, & Almeida, 2013), numerical reasoning scales also contain items that combine two numerical sequences, which could require visualization to identify both sequences.

The hypothesis of the presence of one dominant factor for each partial scale was assessed by using item factor analysis on the tetrachoric correlation matrices. The lowest values were related to the mechanical reasoning scales, with percentages of 16% for Form-2 and 19% for Form-3. The percentage of variance explained by the rest of the dominant factors ranged from .28 in the abstract reasoning scale Form-2 to .48 in the practical reasoning test Form-1. The values shown in the mechanical reasoning test were somewhat lower than the 20% usually adopted to define scale unidimensionality. Recent studies on BPR have suggested that the heterogeneity of situations presented by the items may allow students to respond through practical intuitive or tacit knowledge and through a process of visualization (Amaral, Almeida, & Morais, 2014), which could generate the presence of a specific factor associated with visual capacity (Lemos, Abad, Almeida, & Colom, 2013; Primi, Rocha da Silva, Rodrigues, Muniz, & Almeida, 2013).

According to the theoretical model on which the BPR was built, a common general factor was predicted which would reflect the importance of reasoning in the resolution of any of the test tasks. The confirmatory factor analyses for each of the BPR forms confirmed this hypothesis. The general factor explained a percentage of variance of 60%, 44% and 42% for Form-1, Form-2 and Form-3, respectively. As noted by Almeida, Guisande, Primi and Lemos (2008), the percentage of explained variance decreased slightly as the students' grade level increased and as the mechanical reasoning scales, which lower coefficients in the general factor, were included in the model. Similar results were confirmed by Elosua and Mujika (in press) in a Spanish sample.

Gender differences in the structure of each form of the BPR were assessed through a factorial invariance study. Configural and metric invariance models provided a good fit to the data in all of the BPR forms, but the results of the scalar invariance models were not good. The analyses showed that none of the BPR forms demonstrated scalar invariance across gender. Only after freeing the numerical and mechanical reasoning parameters were adequate adjustment indices obtained. In these cases, the estimated parameters were higher in the male group. Focusing on BPR Forms-2 and Form-3 (age range 13-22), it is interesting to note that the results concur with earlier research conducted in a Portuguese sample (Lemos, Abad, Almeida, & Colom, 2013), but differ somewhat from the findings of a Spanish sample (Elosua & Mujika, in press). The Spanish version of the battery confirmed the gender invariance associated with mechanical reasoning, but also found different intercept parameters in the abstract reasoning test in Form-3. In studying the origin of the differences between versions, it would be interesting to extend this research with an in-depth analysis of the differential item functioning between the three language groups in order to evaluate the existence of possible biases related to translation or curriculum. It is important to point out that despite divergences found in the partial scales, the three samples (Spanish, Portuguese, Basque) concluded that there were no differences in the general reasoning factor.

The results of the study of the observed differences in the partial scales between male and female scores were consistent with meta-analytic studies of gender differences. These studies concluded: (a) higher scores for males in numerical abilities (Hyde, 2005; Spelke, 2005), and (b) higher averages in verbal abilities among females (Halpern, Benbow, Geary, Gur, Hude, & Gernsbacher, 2007), but with smaller effect sizes. The results were not as clear, however, for abstract reasoning and spatial reasoning scales: (a) No differences were found in abstract reasoning in any of the BPR test forms, and the average scores were slightly higher among the group of females; (b) In spatial reasoning, the differences found between boys and girls were negligible ($g<.16$) and showed no homogeneous pattern. The differences favored the females in Form-2 ($M_{MaleForm2}$ = 11.02; $M_{FemaleForm2}$ = 11.75), but were reversed in Form-3, with the males obtaining higher mean scores ($M_{MaleForm3}$ = 11.19; $M_{FemaleForm2}$ = 10.54).

The mechanical reasoning tests warrant more detailed attention. On the one hand, their unidimensionality was called into question and on the other, these are the scales that produced the greatest differences between genders (g = -.36). The mechanical and spatial

tasks included in the BPR can be considered part of the group of visuospatial abilities at which literature has found males to perform better (Voyer, Voyer, & Bryden, 1995) and with different effect sizes depending on the task. Meta-analytic studies on spatial visualization ability conclude that the greatest gender differences are concentrated in mental cube rotation tasks; however, although this is the type of task included in the spatial reasoning test, the effects found in our study were negligible. This result, together with the behavior of the mechanical reasoning test, turn the focus back to two points of interest: (a) a more in-depth intercultural analysis that enables us to explore the differences among populations where the BPR is used, (b) the possibility that, together with the spatial reasoning test, a factor associated with visual ability is created.

In sum, the empirical evidence supports the presence of a general reasoning factor and is consistent with earlier studies on different samples which suggest that mechanical and spatial reasoning tests may define a group-specific factor. Differences between Spanish, Portuguese and Basque samples affecting these results should be further analyzed for differential item functioning. Although the presence of a general factor was confirmed in the three samples, the structure of the battery shows variations between samples; the main aim of BPR was to construct items with the least possible curricular baggage. The results reported open the door to a possible source of differentiation.

Acknowledgements

This research was supported by the Spanish Ministry of Economy and Competitiveness (PSI2011-30256, PSI2014-54020-P) and by the University of the Basque Country (GIU12/32).

References

Almeida, L.S., Guisande, M.A., Primi, R., & Lemos, G. (2008). Contribuciones del factor general y de los factores específicos en la relación entre inteligencia y rendimiento escolar [Contributions of the general factor and specific factors in the relation between intelligence and scholar achievement]. *European Journal of Education and Psychology, 1*, 5-16.

Almeida, L.S., & Lemos, G. (2006). *Batería de provas de raciocínio: manual técnico* [Reasoning test batttery: Technical manual]. Braga: Universidade do Minho, Centro de Investigação em Psicologia.

Amaral, A.O., Almeida, L.S., & Morais, M.J. (2014). Raciocínio e rendimento escolar: Estudo com adolescentes moçambicanos da 8ª à 10.ª clase [Reasoning and scholar achivement in a sample of mozambican adolescents]. In *Atas do 1º Congresso "Cognição, Aprendizagem & Rendimento"*. Braga: Universidad de Minho: Centro de Investigación en Educación.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington: AERA.

Baumgartl, V.O., & Primi, R. (2006). Evidences on the validity of the Battery of Reasoning Tests (BPR-5) for employment selection. *Psicologia: Reflexão e Crítica, 19*, 246-251.

Carroll, J.B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about 10 brad factors. In H. Nyborg (Ed.), *The Scientific Study of GeneralIintelligence: Tribute to Arthur R. Jensen* (pp. 5-21). Amsterdam: Pergamon.

Cattell, R.B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1-22.

Cattell, R.B. (1971). *Intelligence: Its structure, growth and action*. Boston, MA: Houghton Mifflin.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233-255.

Colom, R., Quiroga, M.A., & Juan-Espinosa, M. (1999). Are cognitive sex differences disappearing? Evidence from Spanish populations. *Personality and Individual Differences, 27*, 1189-1195.

Dolan, C.V., Colom, R., Abad, F.J., Wicherts, J.M., Hessen, D.J., & van de Skyusm S. (2006). Multi-group covariance and mean structure modeling of the relationship between the WAIS-III common factors and sex and educational attainment in Spain. *Intelligence, 34*, 193-210.

Elosua, P., & Iliescu, D. (2012). Tests in Europe. Where we are and where we should to go. *International Journal of Testing*, *12*, 157-175.

Elosua, P., & Mujika, J. (in press). Internal structure and gender invariance of the Spanish version of the reasoning test battery. *Spanish Journal of Psychology*.

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.

Finch, W.H., & French, B.F. (2012). The impact of factor noninvariance on observed composite score variances. *International Journal of Research and Reviews in Applied Sciences*, *10*, 1-13.

Fryer, R.G., Jr. & Levitt, S.D. (2010). An empirical analysis of the gender gap in Mathematics. *American Economic Journal-Applied Economics, 2*, 210-240.

Geiser, C., Lehmann, W., & Eid, M. (2008). A note on sex differences in mental rotation in different age groups. *Intelligence, 36*, 556-563.

Guilford, J.P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Halpern, D. (1997). Sex differences in intelligence: Implications for education. *American Psychologist, 52*, 1091-1102.

Halpern, D.F., Benbow, C., Geary, D.C., Gur, R.C., Hyde, J.S., & Gernsbacher, M.A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8*, 1-51.

Horn, J., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D.P. Flagnagan, J.L. Genshaft & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues*. New York, NY: The Guilford Press.

Hyde, J.S. (2005). The gender similarities hypothesis. *American Psychologist, 60*, 581-592.

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

Irwing, P. (2012). Sex differences in g: An analysis of the US standardization sample of the WAIS-III. *Personality and Individual Differences, 53*, 126-131.

Jackson, D.N., & Rushton, J.P. (2006). Males have greater g: Sex differences in general mental ability from 100,000 17- to 18-year-olds on the Scholastic Assessment Test. *Intelligence, 34*, 479-486.

Johnson, W., & Bouchard, T. (2007). Sex differences in mental abilities: g masks the dimensions on which they lie. *Intelligence, 35*, 23-39.

Johnson, W., Carothers, A., & Deary, I.J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science, 3*, 518-531.

Kenny, D.A., Kaniskan, B., & McCoach, D.B. (in press). The performance of RMSEA in Models with Small Degrees of Freedom. *Sociological Methods Research*.

Lemos, G., Abad, F.J., Almeida, LS., & Colom, R. (2013). Sex differences on g and non-g intellectual performance reveal potential sources of STEM discrepancies. *Intelligence, 41*, 11-18.

Linn, M.C., & Petersen, A.C. (1985). Emergence and characterisation of gender differences in spatial abilities: A meta-analysis. *Child Development, 56*, 1479-1498.

Lohman, D.F., & Lakin, J. (2009). Consistencies in sex differences on the cognitive abilities test across countries, grades, test forms, and cohorts. *British Journal of Educational Psychology*, *79*, 389-407.

Lynn, R. (2002). Sex differences on the Progressive Matrices among 15-16 year olds: Some data from South Africa. *Personality and Individual Differences, 33,* 669-673.

Lynn, R., Raine, T.A., Venables, P.H., Mednick, S.A., & Irwing, P. (2005). Sex differences on the WISC-R in Mauritius. *Intelligence, 3*, 527-533.

McGrew, K.S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D.P. Flanagan & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136-182). New York, NY: Guilford Press.

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,* 149-155.

Muñiz, J., & Fernández-Hermida, J.R. (2010). La opinión de los psicólogos españoles sobre el uso de los tests [The opinion of Spanish psychologists on use of the tests]. *Papeles del Psicólogo*, *31*, 108-121.

Primi, R., & Almeida, L.S. (2000). Estudo de validação da Bateria de provas de raciocínio (BPR-5) [Validation of the Reasoning test battery (BPR-5)]. *Psicologia: Teoria e Pesquisa*, *16*, 165-173.

Primi, R., Couto, G., Almeida, L.S., Guisande, M.A., & Miguel, F.K. (2012). Intelligence, age and schooling: Data from the Battery of Reasoning Tests (BRT-5). *Psicologia: Reflexão e Crítica*, *25*, 79-88.

Primi, R., Rocha da Silva, M.C., Rodrigues, P., Muniz, M., & Almeida, L.S. (2013). The use of the bi-factor model to test the uni-dimensionality of a battery of reasoning tests. *Psicothema*, *25*, 115-122.

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

Schneider, W.J., & McGrew, K.S. (2012). The Cattell-Horn-Carroll Model of Intelligence. In D.P. Flanagan & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 99-144). New York, NY: Guilford Press.

Spelke, E.S. (2005). Sex differences in intrinsic aptitude for mathematics and science? A critical review. *American Psychologist, 60*, 950-958.

Spearman, C. (1927). *The abilities of man*. New York: Macmillan.

Thurstone, L.L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Van der Sluis, S., Posthuma, D., Dolan, C.V., de Geus, E.J.C., Colom, R., & Boomsma, D.I. (2006). Sex differences on the Dutch WAIS-III. *Intelligence, 34,* 273-289.

Vernon, E. (1961). *The Structure of Human Abilities*. London: Methuen.

Voyer, D., Voyer, S., & Bryden, M.P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250-270.