The Spanish Journal of Psychology (2015), **18**, e74, 1–7.

doi:10.1017/sjp.2015.75

CAMBRIDGE UNIVERSITY PRESS

Internal Structure and Partial Invariance across Gender in the Spanish Version of the Reasoning Test Battery

Paula Elosua and Josu Mujika

Universidad del País Vasco (Spain)

Abstract. The Reasoning Test Battery (BPR) is an instrument built on theories of the hierarchical organization of cognitive abilities and therefore consists of different tasks related with abstract, numerical, verbal, practical, spatial and mechanical reasoning. It was originally created in Belgium and later adapted to Portuguese. There are three forms of the battery consisting of different items and scales which cover an age range from 9 to 22. This paper focuses on the adaptation of the BPR to Spanish, and analyzes different aspects of its internal structure: (a) exploratory item factor analysis was applied to assess the presence of a dominant factor for each partial scale; (b) the general underlined model was evaluated through confirmatory factor analysis, and (c) factorial invariance across gender was studied. The sample consisted of 2624 Spanish students. The results concluded the presence of a general factor beyond the scales, with equivalent values for men and women, and gender differences in the factorial structure which affect the numerical reasoning, abstract reasoning and mechanical reasoning scales.

Received 13 November 2014; Revised 25 March 2015; Accepted 2 May 2015

Keywords: reasoning tests, factorial structure, adaptation, gender invariance.

The Reasoning Test Battery ("Bateria de Provas de Raciocionio", BPR) was developed out of the "Tests de Raisonnement Différentiel" built by Meuris (1969) in Belgium. The purpose of the battery is to assess reasoning competence among children and adolescents based on tasks containing different contents. The BPR assesses general aspects of intelligence together with components associated with skills usually assessed in specific multifactorial intelligence batteries. The substantive basis of the Reasoning Test Battery is the theory of the hierarchical organization of cognitive abilities (Cattell, 1963, 1971; Vernon, 1961) and therefore combines a general cognitive factor and specific factors (Horn & Noll, 1997). Drawing on the three-factor model or CHC theory (Cattell-Horn-Carroll), general reasoning is associated with the g factor (fluid intelligence), and particular contents with specific factors related to cognitive functions or types of information to be processed (oral, figurative, numerical) (Carroll, 2003).

In terms of reasoning, the battery contains analogies, series completion 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). The current version is a battery of tests with three different forms which were designed to assess a wide range of ages: BPR Form 1 consists of four scales for 4th, 5th and 6th grade students (9–12 years old); BPR Form 2 comprises 5 scales and covers the first three years of secondary education, (12–15 years old); and BPR Form 3, with 5 scales, is designed for students in the fourth year of secondary education and the two pre-university years, called Bachiller (15–20 years old).

Versions of the BPR have been adapted to Portugal (Almeida & Lemos, 2006) and Brazil (Primi & Almeida, 2000). The reliability coefficients reported for the scales are greater than .70, and in most cases the values are above .80. Studies on the dimensionality of the BPR have concluded the presence of a general factor which explains between 40% and 60% of the scale variance.

With the aim of extending the use of a reasoning test constructed in Europe, the purpose of this research was to adapt the three BPR test forms to Spanish and to analyze their internal structure through different approaches: exploratory item factor analysis, confirmatory factor analysis and factorial invariance across gender.

Including a gender related factorial invariance study of the BPR is crucial, since together with the hierarchical organization of cognitive abilities, analyzing the possible differences in reasoning as a function of gender is a recurring theme in psychology research (Deary, Penke, & Johnson, 2010; Johnson, Carothers, & Deary,

Correspondence concerning this article should be addressed to Paula Elosua. Universidad del País Vasco. Avda. Tolosa, 70. San Sebastian, Guipuzcoa (Spain). 20018.

E-mail: paula.elosua@ehu.es

This research was supported by the Spanish Ministerio de Economía y Competitividad (PSI2011-30256, PSI2014-54020-P) and by the University of the País Vasco (GIU12/32, GIU15/24)

2008; Lohman & Lakin, 2009). Although no significant differences have been concluded in terms of g factor (Deary, Irwing, Der, & Bates, 2007; Jensen, 1998), differences in specific cognitive abilities such as verbal, visuospatial and quantitative have been found (for a review, see Halpern et al., 2007). Gender difference studies are usually based on the comparison of observed or latent scores. But from a methodological point of view, prior to analyzing any score differences, it is important to assess the factorial invariance of the compared model, because lack of factorial invariance can increase Type I errors in observed score comparison (Finch & French, 2012). In this regard, Lemos, Abad, Almeida, & Colom (2013) carried out an invariance study on the BPR Form 2 and Form 3, and concluded differences in the general structure affecting numerical reasoning and mechanical reasoning scales.

Method

Adaptation

The process of adapting the BPR to Spanish followed the recommendations of the International Test Commission (Elosua, Mujika, Almeida, & Hermosilla, 2014; Muñiz, Elosua, & Hambleton, 2013). Applying an independent forward translation design and an iterative review process, the battery was adapted to Spanish. Two independent translators translated the Portuguese version to Spanish and a multidisciplinary team composed of two psychometricians, two professional translators and two primary and secondary education teachers reviewed the product. A final version was ultimately adopted by consensus. Some items in the verbal section were modified to maintain semantic equivalence in terms of familiarity and difficulty: five items from Form 1, three items from Form 2 and three items from Form 3.

Participants

The sample comprised 2624 students, 1299 females and 1325 males. The age of the students ranged from 9 to 22 years. The mean age of the participants completing Form 1 was 10.36 years (SD = 1.03), Form 2, 13.44 years (SD = 1.08) and Form 3, 16.69 years (SD = 1.34). The distribution of the sample is shown in table 1.

Table 1. *Sample Composition*

Form	Females	Males	Total
BPR-1	537	620	1157
BPR-2	498	380	878
BPR-3	264	325	589
Total	1299	1325	2624

Instrument

The Reasoning Test Battery consists of three different forms, each organized into different scales and items. Table 2 shows the scales, the tasks involved and the number of items in each scale.

Procedure

The psychometric analyses focused on score reliability and internal structure.

Consistency of the partial scales of each BPR form was assessed estimating ordinal alpha (Elosua & Zumbo, 2008). The internal structure of the battery was evaluated using two different approaches; first, an exploratory item factor analysis on the item thetracoric matrix for each scale was carried out using ULS estimator to assess the presence of a dominant factor; secondly, a confirmatory factor analysis was performed on each of the three BPR forms to test the hierarchical model underlying the battery. Finally, a factorial invariance study across gender was performed on each of the BPR forms. The measurement invariance examined the equality of the factor pattern matrices (configural invariance), the equality of the loading matrices (measurement invariance), and the equality of the intercepts (scalar invariance). The analyses were performed with the 'lavaan' package (Rosseel, 2012). Model fit was assessed using 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, 2014). The cut-off point for CFI is usually fixed at .90 and for SRMR, at .08 (Hu & Bentler, 1999). 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 (Cheung & Rensvold, 2002).

Results

Internal Consistency

Ordinal alpha values ranged from .86 to .94 (see table 3). The highest values were found in the numerical scales in the three forms (α_{Form1} = .93; α_{Form2} = .94; α_{Form3} = .93), and the lowest values were associated with the mechanical reasoning scales (α_{Form2} = .80; α_{Form3} = .86).

Unidimensionality of Partial Scales

Results of the exploratory factor analyses on the thetracoric correlation matrix performed on each of the scales are summarized in table 3. For abstract, verbal, numerical and practical reasoning scales the percentages of

https://doi.org/10.1017/sjp.2015.75 Published online by Cambridge University Press

Table 2. *Structure of the Reasoning Test Battery*

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

Table 3. *Reasoning Tests Battery Scales. Internal Consistency, Explained Variance and Regression Weights*

	Form 1			Form 2			Form 3		
Scale	α_{or}	%Var	λ	α_{or}	%Var	λ	α_{or}	%Var	λ
Abstract R.	.91	.34	.75	.90	.28	.66	.91	.34	.67
Verbal R.	.91	.34	.75 (.03)	.90	.30	.71 (.05)	.91	.34	.77 (.05)
Numerical R.	.93	.47	.71 (.03)	.94	.50	.66 (.05)	.93	.44	.65 (.05)
Practical R.	.92	.48	.76 (.03)						
Spatial R.				.91	.34	.75 (.05)	.89	.30	.78 (.05)
Mechanical R.				.80	.15	.52 (.05)	.86	.20	.40 (.05)

Notes: α_{or}. Ordinal alpha. %Var Explained variance percentage. λ. Factor loading. Measurement errors in parentheses.

variance associated with the one-dimensional factor were greater than .30 in the three BPR forms; the only exception was for the abstract reasoning scale in Form 2, which yielded a variance percentage of 28%. However, the mechanical reasoning scales showed lower values; 15% of the variance was explained by the mechanical dominant factor in Form 2 and the explained variance was 20% in Form 3.

General Factor Structure

The presence of a general underlying factor was assessed by analyzing scale correlation matrices using maximum likelihood estimation. Table 3 shows the regression weights (λ) and standard errors for each partial scale. The loading values were above .65 for the abstract, verbal, practical and spatial reasoning scales. The mechanical reasoning scale weights were .52 in Form 2, and .40 in Form 3. All the loadings were statistically significant. The SRMR and CFI indexes (table 4) of the confirmatory models met the criteria for a good fit, although RMSEA indexes were little higher than the cut-off points ($RMSEA_{Form2}$ = .10; $RMSEA_{Form3}$ = .14). The extracted general factor explained 54% of the variance in Form 1, 43% in Form 2, and 44% in Form 3.

Factorial Invariance

Results of the factorial invariance are shown in table 5. The CFI indexes obtained in evaluating baseline models yielded values over .97 except for the female sample in Form 3; (CFI = .91; RMSEA = .16). Configural invariance examined the same measurement structure in both samples. This basic invariance showed a reasonable fit to the data in the three forms, with CFI values above .96. Metric equivalence added a restriction to the previous models, which was the equality between the regression coefficients in all six samples. The unidimensional model and the loading matrices were held invariant across samples. Metric invariance was held for Form 1 and Form 2 (ΔCFI_{Form1} = –0; ΔCFI_{Form2} = –.01) but for the oldest students the difference between CFI values

Table 4. *Fit Indexes of the Confirmatory Unidimensional Model*

	χ^2	g.l.	*p*	SRMR	CFI	RMSEA	$IC_{90\%}$ RMSEA
Form 1	0.51	2	.77	.00	.99	.00	.00 – .04
Form 2	40.49	5	<.01	.04	.96	.10	.07 – .13
Form 3	63.43	5	<.01	.06	.93	.14	.11 – .18

https://doi.org/10.1017/sjp.2015.75 Published online by Cambridge University Press

Table 5. *Factorial Invariance Models*

Model	Form 1				Form 2				Form 3			
	χ^2	*df*	RMSEA	CFI	χ^2	*df*	CFI	RMSEA	χ^2	*df*	CFI	RMSEA
Females. Base model	1.76	2	.01	.99	16.21	5	.98	.08	36.54	5	.91	.16
Males. Base model	1.12	2	.01	.99	18.93	5	.97	.10	5.94	5	.97	.03
Configural Invariance	2.88	4	.01	.99	35.14	10	.97	.09	42.48	10	.96	.11
Metric Invariance	9.31	7	.03	.99	44.99	14	.97	.08	81.15	14	.92	.13
Partial Metric (mr free)									56.83	13	.95	.11
Scalar Invariance	72.13	10	.10	.96	88.23	18	.93	.10	144.96	17	.85	.16
Partial scalar (nr free)	26.06	9	.06	.99								
Partial Scalar (mr free)					50.85	17	.965	.07	83.01	16	.92	.12
Partial Scalar (mr and nr free)									61.23	15	.944	.10

was higher than the cut-off point ($CFI_{configural}$ - CFI_{metric} = .96–.92 = .04).The partial metric invariance was assessed by estimating the parameters for the mechanical reasoning scale independently in males and females ($\lambda_{females}$ = .23; λ_{males} = .84); the result was a reduction in the CFI difference and in the RMSEA value ($CFI_{configural}$ - CFI_{metric} = .96–.95 = .01; RMSEA = .10).

In order to evaluate scalar invariance, the intercepts (ν) were added to the model. The fit indexes obtained were slightly poorer than those for the rest of the models assessed. In the three forms analyzed the differences in the CFI values were greater than .01 ($\Delta CFI_{Form\ 1}$ = –.04; $\Delta CFI_{Form\ 2}$ = –.04, $\Delta CFI_{Form\ 3}$ = –.05), and in all three cases the RMSEA values were equal to or exceeded .10 ($RMSEA_{Form\ 1}$ = .10; $RMSEA_{Form\ 2}$ = .10; $RMSEA_{Form\ 3}$ = .16). The statistical analysis of the modification indexes suggested the freeing of the intercepts associated with numerical reasoning in Form 1, and mechanical reasoning in Forms 2 and Form 3. After the new models were estimated, the fit indexes for Form 1 and Form 2 met the established criteria ($\chi^2_{Form\ 1}$ = 26.06; *df* = 9; $\Delta CFI_{Form\ 1}$ = –.01, $RMSEA_{Form\ 1}$ = .06, $\chi^2_{Form\ 2}$ = 50.85; *df* = 17; $\Delta CFI_{Form\ 2}$ = –.005, $RMSEA_{Form\ 2}$ = .07). However, the results of the partial invariance model adjustment were not acceptable in Form 3. Only after freeing the parameters for the abstract reasoning scale, were adequate adjustment indexes found ($\chi^2_{Form\ 3}$ = 61.23; *df* = 15; $\Delta CFI_{Form\ 3}$ = –.006; $RMSEA_{Form\ 3}$ = .10).

Although the latent mean and covariance structure analysis (MACS) did not estimate the absolute means, latent mean differences across groups can be estimated by fixing the latent means values to zero for one of the groups. The female group was defined as the base group in the three forms; therefore, the means in the general factor were fixed to zero for this group and were freely estimated in the male groups. Comparisons among groups were based on the statistical significance of the difference evaluated by *t* values (estimated mean divided by the standard error). In none of the three forms were statistically significant differences found in general reasoning ($M_{Form\ 1}$ = .336; Se = .181; *p* = .06; $M_{Form\ 2}$ = –.024; Se = .223; *p* = .914; $M_{Form\ 3}$ = .492; Se = .265; *p* = .064). Table 6 shows the estimated parameters of the final models.

Discussion

According to the latest version of the standards for Educational and Psychological Testing (AERA, APA, &

Table 6. *Final models parameter estimations*

Scale	BPR Form 1			BPR Form 2			BPR Form 3			
	λ	ν	ν_{female}	λ	ν	ν_{female}	λ	ν	λ_{female}	ν_{female}
Abstract R.	1.02 (0.04)	12.33 (0.13)		1.00	14.09 (0.18)		1.00	12.01 (0.21)		9.97 (1.06)
Verbal R.	1.00	14.68 (0.13)		1.10 (0.07)	15.70 (0.18)		1.21 (0.08)	15.23 (0.23)		
Numerical R.	.91 (0.04)	7.71 (0.14)	6.63 (0.16)	1.12 (0.08)	7.37 (0.19)		0.92 (0.07)	8.58 (0.19)		
Practical R.	.80 (0.03)	9.65 (0.11)								
Spatial R.				1.33 (0.08)	10.80 (0.21)		1.23 (0.08)	10.16 (0.22)		
Mechanical R.				0.74 (0.06)	10.70 (0.21)	9.15 (0.18)	0.84 (0.09)	9.84 (0.22)	0.23 (0.08)	7.46 (0.23)

Note: Standard errors in parenthesis.

https://doi.org/10.1017/sjp.2015.75 Published online by Cambridge University Press

NCMEA, 2014), "validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (p. 11). The BPR, in its different forms, was constructed to assess general cognitive ability. Therefore, the score validation process should provide arguments for the conceptual framework that supports the BPR, the hierarchical structure for reasoning.

The aim of this work was to empirically assess this conceptual framework by gathering evidence based on the internal structure of the tests. The three forms of the BPR consist of different scales or tasks (4 or 5) which jointly configure a general reasoning factor. For each partial reasoning scale item homogeneity was assessed through the ordinal reliability coefficient and exploratory item factor analyses on thetracoric correlation matrices. The estimated coefficients ranged from .80, obtained for Mechanical Reasoning Form 2, to the highest reliability coefficients for the numerical reasoning scales in the three BPR test forms (α_{Form1} = .93; α_{Form2} = .94; α_{Form3} = .93). These results agree with the reliability studies carried out with the BPR (Almeida & Lemos, 2006; Baumgarti & Primi, 2006). It is important to note that the high values obtained in the numerical reasoning scales may be explained by the fact that the scales are the only ones that require a constructed response rather than a multiple-choice response (Primi, Rocha da Silva, Rodriguez, Muniz, & Almeida, 2013).

The hypothesis of the presence of a dominant factor for each partial scale was assessed by using exploratory item factor analysis on the tetrachoric correlation matrices. The percentage of variance explained by the unidimensional factors ranged from .15 in the mechanical reasoning scale Form 2 to .48 in the practical reasoning test Form 1. The lowest values were related systematically to the mechanical reasoning scales, with percentages of 15% for Form 2 and 20% for Form 3. The values shown in the mechanical reasoning test are 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 intuition 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 et al., 2013; Primi et al., 2013).

The partial factors defined by each of the reasoning scales (abstract, mechanical, spatial, numerical, verbal, practical) could be considered elements that contribute to the formation of a general reasoning factor. According to the theoretical model on which BPR was built, a common general factor was expected to be found that 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 54%, 43% and 44% for Form 1, Form 2 and Form 3 respectively. As noted by Almeida, Guisande, Primi, & Lemos (2008), the percentage of explained variance decreased slightly with each grade level, and the mechanical reasoning scales were included in the model, which lower coefficients in the general factor.

Finally, the factorial invariance across gender was assessed. The fit of the baseline models was good for all of the forms in the male and female samples, although in Form 3 the goodness-of-fit was worse for the female behavior. The factorial invariance analysis continued with an examination of the equivalence between the unidimensional configurations of the factorial model in both samples. The configural invariance showed a good fit to the data in all of the BPR forms. The next step was to analyze the metric equivalence. This level of invariance added a restriction to the previous model: the equality between the regression coefficients. The fit indexes remained acceptable except for Form 3. After freeing the loadings associated with the mechanical reasoning scale in Form 3, the final indexes showed good fit between the model and the data. The weight of the mechanical scales on general reasoning in the male sample ($\lambda_{females}$ = .84) was higher than the weight estimated in the female sample ($\lambda_{females}$ = .23).

The study then went on to assess scalar invariance, a necessary step for comparing groups with regard to the general reasoning factor. The analyses showed that none of the BPR forms demonstrated scalar noninvariance across gender. The intercept values for men and women were different in the numerical reasoning scales (Form 1), mechanical reasoning scales (Form 2 and Form 3) and abstract reasoning scales (Form 3). In all three cases the estimated parameters were greater in the male group than in the female group. A substantive interpretation of the findings is directly connected to gender differences in cognitive skills, where scores are systematically higher for males in numerical and visuospatial abilities (Hyde, 2005; Spelke, 2005; Voyer, Voyer, & Bryden, 1995) and no gender differences are associated with the g factor (Deary et al., 2007). However, in this study it is worth pointing out that the only differences in the numerical reasoning test were found in the youngest group of students (9–12 years old).

Focusing on BPR Forms 2 and 3 (age range 13–22), it is interesting to note that the results concur in part with earlier research conducted in a Portuguese sample (Lemos et al., 2013). While the authors of the 2013 study found partial invariance for the mechanical and numerical scales, the data reported in this study show that the gender differences are associated with the mechanical and abstract reasoning tests. Differences between Spanish and Portuguese samples affecting these

https://doi.org/10.1017/sjp.2015.75 Published online by Cambridge University Press

results – numerical scale partial invariance and mechanical scale partial invariance – should be further analyzed. Although 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 that would have to be approached through intercultural research.

From a practical and methodological point of view, these results warn against the incorrect practice of directly comparing g scores across gender without evaluating the previous condition of factorial invariance. If the weights and/or intercept values for the factorial model are not invariant, these inequalities might mask any differences in g scores. Any g score comparison must take this different configuration into account.

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.[Contribution of general factor and specific factors into the relation between intelligence and scholar achivement] *European Journal of Education and ychology*, *1*, 5–16.

Almeida L. S., & Lemos G. (2006). *Batería de Provas de Raciocínio: Manual Técnico*. Braga, Portugal: 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 10ª classe. [Reasoning and scholar achievement. Study with mozambican students in grades 8–10]. In *Atas do 1° Congresso "Cognição, Aprendizagem & Rendimento"*. Braga, Portugal: 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, DC: AERA.

Baumgarti V. O., & Primi R. (2006). Evidências de validade da Bateria de Provas de Raciocínio (BPR-5) para seleção de pessoal. [Validity evidences of the Reasoning Test Battery for recruitment] *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 general intelligence: Tribute to Arthur R. Jensen* (pp. 5–21). Amsterdam, The Netherlands: Pergamon.

Cattell R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*. 1–22. http://dx.doi.org/10.1037/h0046743

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.

Deary I. J., Irwing P., Der G., & Bates T. C. (2007). Brother–sister differences in the g factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY1979. *Intelligence*, *35*, 451–456.

Deary I. J., Penke L., & Johnson W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience*, *11*, 201–211. http://dx.doi.org/10.1038/nrn2793

Elosua P., Mujika J., Almeida L., & Hermosilla D. (2014). Procedimientos analítico-racionales en la adaptación de tests. Adaptación al español de la Batería de Pruebas de Razonamiento [Judgmental-analytical procedures for adapting tests: Adaptation to Spanish of the Reasoning Tests Battery]. *Revista Iatinoamericana de Psicología*, *46*, 117–126.

Elosua P., & Zumbo B. (2008). Coeficientes de fiabilidad para escalas de respuesta categórica ordenada. *Psicothema*, *20*, 896–901.

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.

Halpern D. F., Benbow C. P., 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. http://dx.doi.org/10.1111/j.1529-1006.2007.00032.x

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

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.

Hyde J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*, 581–592. http://dx.doi.org/10.1037/0003-066X.60.6.581

Jensen A. R. (1998). *The g Factor: The science of mental ability*. Westport, CT: Praeger.

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. http://dx.doi.org/10.1111/j.1745-6924.2008.00096.x

Kenny D. A., Kaniskan B., & McCoach D. B. (2014). The performance of RMSEA in Models with Small Degrees of Freedom. *Sociological Methods Research*, *44*, 486–507. http://dx.doi.org/10.1177/0049124114543236

Lemos G. C., Abad F. J., Almeida L. S., & Colom R. (2013). Sex differences on g and non-g intellectual performance reveal potential sources of STEM discrepancies. *Intelligence*, *41*, 11–18. http://dx.doi.org/10.1016/j.intell.2012.10.009

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. http://dx.doi.org/10.1348/000709908X354609

Meuris G. (1969). *Tests de raisonnement différentiel*. Brussels, Belgium: Editest.

https://doi.org/10.1017/sjp.2015.75 Published online by Cambridge University Press

Muñiz J., Elosua P., & Hambleton R. K. (2013). Directrices para la traducción y adaptación de los tests: Segunda edición Guidelines for test translation and adaptation: Second edition]. *Psicothema*, *25*, 149–155.

Primi R., & Almeida L. S. (2000). Estudo de Validação da Bateria de Provas de Raciocínio (BPR-5). [Validation of the Reasoning Test Battery]. *Psicologia: Teoria e Pesquisa*, *16*, 165–173.

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.

Spelke E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? A critical review. *American Psychologist*, *60*, 950–958. http://dx.doi.org/10.1037/0003-066X.60.9.950

Vernon E. (1961). *The structure of human abilities*. London, UK: 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. http://dx.doi.org/10.1037/0033-2909.117.2.250

https://doi.org/10.1017/sjp.2015.75 Published online by Cambridge University Press