ISSN online 1688-4221

Ciencias Psicológicas January-June 2025; 19(1), e-4005
DOI: 10.22235/cp.v19i1.4005

Validation of the Interpersonal Reactivity Index (IRI) in a Uruguayan sample

Validación del Índice de Reactividad Interpersonal (IRI) en una muestra uruguaya

Validação do Índice de Reatividade Interpessoal (IRI) em uma amostra uruguaia

Martín Bidegain[1]
Marcela Carballo[1]
Daniel Costa-Ball[1]

[1] Universidad Católica del Uruguay

Received: 04/10/2024
Accepted: 03/27/2025

Correspondence:
Martín Bidegain,
martin.bidegain@gmail.com

How to cite: Bidegain, M., Carballo, M., Costa-Ball, D. (2025). Validation of the Interpersonal Reactivity Index (IRI) in a Uruguayan sample. *Ciencias Psicológicas*, *19*(1), e-4005. https://doi.org/10.22235/cp.v19i1.4005

Data availability: The dataset that supports the findings of this study is not publicly available.

Abstract: Background: The Interpersonal Reactivity Index (IRI), introduced by Davis in 1980, remains a widely used self-report tool designed to assess empathy from a multidimensional perspective, with adaptations for various languages and populations. However, many studies examining its psychometric properties have failed to replicate the original four-factor structure. Method: This instrumental study applied Classical Test Theory to adapt the IRI for use in Uruguay, with a sample of 858 adult participants (640 females, 218 males). Results: The original 28-item scale did not show a good fit with the four-factor model. Removing the reversed items led to an acceptable fit for both the original and a three-factor model. Conclusion: These findings suggest that a shortened version of the scale, excluding the reversed items, would be more suitable for this population. Further research on the impact of reversed items is recommended. The study highlights the importance of ongoing investigation into the psychometric properties of the IRI, as well as the theoretical and practical implications of using a shortened version.
Keywords: empathy; IRI; validation; psychometric properties; assessment

Resumen: Antecedentes: el Índice de Reactividad Interpersonal (IRI), introducido por Davis en 1980, sigue siendo una escala de autoinforme ampliamente utilizada para evaluar la empatía desde una perspectiva multidimensional, con adaptaciones a diferentes idiomas y poblaciones. Sin embargo, muchos estudios sobre propiedades psicométricas no lograron replicar la estructura original de cuatro factores. Método: Se desarrolló un estudio instrumental que aplicó la Teoría Clásica de Test para adaptar el instrumento IRI a Uruguay en una muestra de 858 participantes adultos uruguayos (640 mujeres, 218 hombres). Resultados: La escala original de 28 ítems no mostró un buen ajuste al modelo de cuatro factores. La eliminación de los ítems invertidos resultó en un ajuste adecuado para el modelo original y un modelo de tres factores. Conclusión: los hallazgos sugieren que en esta población se debe utilizar una versión abreviada de la escala, sin ítems invertidos. Se deben realizar más estudios sobre los problemas relacionados con los ítems invertidos. Este estudio enfatiza la importancia de la investigación continua sobre las propiedades psicométricas del instrumento, los fundamentos teóricos y las implicaciones prácticas del uso de una versión abreviada.
Palabras clave: empatía; IRI; validación; propiedades psicométricas; evaluación

Resumo: Antecedentes: O Índice de Reatividade Interpessoal (IRI), introduzido por Davis em 1980, continua sendo uma escala de autorrelato amplamente utilizada para avaliar a empatia a partir de uma perspectiva multidimensional, com adaptações para diferentes idiomas e populações. No entanto, muitos estudos sobre propriedades psicométricas não conseguiram replicar a estrutura original de quatro fatores. Método: Foi desenvolvido um estudo instrumental que aplicou a Teoria Clássica dos Testes para adaptar o instrumento IRI ao Uruguai em uma amostra de 858 participantes adultos uruguaios (640 mulheres, 218 homens). Resultados: A escala original de 28 itens não apresentou um bom ajuste ao modelo de quatro fatores. A eliminação dos itens invertidos resultou em um ajuste adequado para o modelo original e um modelo de três fatores. Conclusão: Os resultados sugerem que uma versão abreviada da escala, sem itens invertidos, deve ser utilizada nesta população. Devem ser realizados mais estudos sobre os problemas relacionados aos itens invertidos. Este estudo enfatiza a importância da pesquisa contínua sobre as propriedades psicométricas do instrumento, os fundamentos teóricos e as implicações práticas do uso de uma versão abreviada.
Palavras-chave: empatia; IRI; validação; propriedades psicométricas; avaliação

The Interpersonal Reactivity Index (IRI) is a self-report scale designed in 1980 by Mark H. Davis to assess empathy from a multidimensional perspective, conceptualized as a group of related but independent cognitive and affective processes. It is one of the first tools to assess cognitive and affective dimensions separately, and it has been one of the most widely used instruments in the field (Ilgunaite et al., 2017).

The scale consists of 28 items evenly distributed across four subscales: Perspective Taking (PT) and Fantasy (F), corresponding to the cognitive dimension, and Personal Distress (PD) and Empathic Concern (EC), to the affective dimension. The PD subscale contains items that assess negative feelings arising in the individual when perceiving another person's suffering, while EC evaluates feelings of compassion and a desire to alleviate the other person's suffering. PT assesses the ability to understand that others may have thoughts and feelings different from one's own, and F comprises items related to the tendency to put oneself in the position of fictional characters. The psychometric properties analysis reported by Davis (1980, 1983) shows satisfactory internal reliability of the four scales (ranging from .71 to .77) and test-retest reliability ranging from .62 to .71.

The question regarding the multidimensional nature of empathy has been central to both theoretical discussions and the development of measurement techniques.

Currently, different authors have considered that empathy involves at least two types of processes: cognitive and affective (e.g., Chakrabarti & Baron-Cohen, 2006; Decety & Jackson, 2004; Preston & De Waal, 2002). The former is responsible for making inferences about others' mental states (beliefs, ideas, desires, feelings), while the latter is linked to the ability to detect and/or experience others' emotional states. Despite the fact that multidimensional and integrative models of empathy are currently the most accepted, the discussion about the structure, components, and whether they are independent from each other is still unresolved (e.g., Lawrence et al., 2004; Spreng et al., 2009).

The four-factor model presented by Davis (1983) has been tested in various psychometric studies in normative population yielding mixed and inconsistent results. While in some studies it was successfully confirmed, in others, the fit of the original model is only marginally acceptable or requires some changes like item elimination (Table 1).

Furthermore, some have yielded results that do not confirm the four-factor model, and among these, some authors identify different dimensions than those proposed by Davis (Baldner & McGinley, 2014, 2020; Koller & Lamm, 2015; Wang et al., 2020; Yarnold et al., 1996). Cliffordson (2002) and Hawk et al. (2013) propose that the original four factors contribute to a higher-order general factor, while Pulos et al. (2004) found a structure with a higher-order factor encompassing the TP, F, and EC subscales, while PD constitutes an independent factor.

Another prominent model in the literature is the two-factor model: affective and cognitive, each composed of their corresponding subscales. Although this model is widely used in empathy research, no psychometric studies have confirmed its validity (Chrysikou & Thompson, 2016).

Table 1

Factorial structure: studies that have confirmed the original model (Davis, 1983)

Authors, year	N	Software	χ^2	df	χ^2/df	CFI	TLI	RMSEA
Model fit for the complete scale								
Fernández et al. (2011)	435	Mplus	781.74	344	2.34	.813	-	.054
Gilet et al. (2013)	322	AMOS	789	344	2.29	.81	-	.065
Chrysikou & Thompson (2016)	435	Mplus	-	-	-	.955	.950	.110
Lucas-Molina et al. (2017)	2499	Mplus	2075.1	272	7,63	.89	.85	.052
Budagovskaia et al. (2017)	318	PASW	-	-	-	-	-	-
Manarte & Andrade (2018)	275	AMOS	-	-	2.13	.77	-	.064
Ahuatzin-González et al. (2019)	729	LISREL	-	-	4.78	.91	-	.079
Rajput et al. (2020)	100	AMOS	2081.9	-	-	.957	-	.056
Diotaiuti et al. (2021)	300	AMOS	117.99	-	1.29	.975	.968	.031
	300		128.85	90	-	.974	.965	.034
Model fit eliminating items								
Pérez-Albéniz et al. (2003)	1997	LISREL	-	-	9.29	-	-	.063
	692		-	-	6.38	-	-	.083
	515		-	-	2.48	-	-	.053
De Corte et al. (2007)	651	LISREL	-	-	2.67	.9	-	.05
Sampaio et al. (2011)	251	AMOS	-	-	0.89	.99	-	.010
Limpo et al. (2010)	478	-	496.47	243	2.04	.86	-	.07
Muller et al. (2015)	266	EQS	142.95	98	1.45	.93	-	.045
Model fit eliminating negative items								
Palmese & Schmidth (2013)	509	LISREL	503.11	242	2.01	.94	-	.46
Braun et al. (2015)	1244	R	196.01	344	0.57	.92	-	.05
Garcia-Barrera et al. (2016)	548	Mplus	412.77	129	3.20	.9	.88	.063
Murphy et al. (2020)	401	R	950.69	316	3.01	.95	.94	.08
Arenas-Estevez et al. (2021)	413	R	477.3	146	3.27	.962	.956	.074
	1366		1591	146	10.90	.95	.941	.085
Grimaldo et al. (2022)	859	R	714	129	5.54	.924	.910	.073

Note. CFI: comparative fit index; TLI: Tucker-Lewis index; RMSEA: root mean square error of approximation; RMSR: square root of the mean of the squared residuals.

Several psychometric studies have found issues with the reversed items of the IRI (Arenas-Estevez et al., 2021; Braun et al., 2015; Garcia-Barrera et al., 2017; Grimaldo et al., 2022; Murphy et al., 2020; Palmese & Schmidth, 2013). The inclusion of reversed items in self-report scales has been debated as a potential threat to validity (Lundgren et al., 2018). Reversed items can lead to interpretation issues (Haladyna, 2002), resulting in the emergence of a method factor (Sonderen et al., 2013; Zhang et al., 2016). Current guidelines for the development and validation of items for tests or assessment instruments recommend wording items in a direct or positive manner, avoiding negative phrases (Haladyna & Rodriguez, 2013).

In Spanish, the issue with reverse-coded items has shown great significance. Venta et al. (2022) evaluated the psychometric performance of reverse-coded items in a Spanish-speaking population in the United States. The results demonstrated a decline in the psychometric performance of these items. When the reverse-coded items were reverted to their original format, the item correlations with their respective subscale scores improved, along with the overall internal consistency of the scale. This confirms that reverse-coded items present more challenges in Spanish than in other languages, supporting the recommendation to avoid their use in scales designed for this language.

The IRI scale is not exempt from these limitations, as it presents nine items worded in a reversed manner (3, 12, 13, 19), with five of them also written with a negation (4, 7, 14, 15, 18). At least five studies in Romance languages—three in Spanish (Arena-Estevez et al., 2021; García-Barrera et al, 2017; Grimaldo et al., 2022), one in Italian (Palmese & Schmidt, 2013), and one in French (Braun et al., 2015)—have reported issues with the reverse-coded items in the IRI.

Among studies conducted in Spanish-speaking contexts, Grimaldo et al. (2022) examined structural validity, invariance, and reliability among university students in Peru. The results from confirmatory factor analysis (CFA) challenged the original four-dimensional model. However, a suitable model fit was attained by eliminating the reversed items and those with negligible variance. In different

research, the four items with reverse wording and negation (4, 14, 18, and 15) were excluded, resulting in a satisfactory fit for the four-dimensional model (Palmese & Schmidt, 2013). In another study involving university students from Colombia, CFA revealed poor indicators for nine items. Subsequently, they attempted to enhance model fit by removing the reversed items, which led to favorable fit indicators (Arenas-Estevez et al., 2021).

Another study on the factorial structure of the IRI in incarcerated population found that all reversed-worded items loaded onto a single component (method factor) despite belonging to different subscales. Upon removing these items, the four-component structure originally reported by Davis was replicated (Lauterbach & Hosser, 2007).

In 2004, Beven et al. (2004) administered IRI to 88 violent criminals from maximum-security prisons. They found that negative items were clustered in one component, alongside another positive item, albeit with lengthy and complex wording. The authors suggest that linguistic complexity and potential low level of reading comprehension might impact the responses, as observed in other cases.

In addition to the contributions from psychometric studies, the theoretical relevance of some scales like F and PD has been discussed, considering that they do not align with a current conceptualization of empathy (Baldner & McGinley, 2014). Criticism of the F scale suggests that it does not truly measure what it claims to but instead reflects aspects more related to imagination or self-control (Batchelder et al., 2017; Cliffordson, 2002). Regarding the PD scale, Murphy et al. (2020) found poor fit and argued that it is lacking in terms of construct validity. Furthermore, Cliffordson (2002) posits that this may not be a central component of empathy.

Despite these criticisms, the IRI continues to be a widely used instrument for measuring empathy, with numerous adaptations and translations. Validation and adaptation studies have been conducted in different languages, including Portuguese (Sampaio et al., 2011; Shiramizu & Yamamoto, 2018), Dutch (De Corte et al., 2007), Chinese (Siu & Shek, 2005), German (Paulus, 2009), Kannada (Rajput et al., 2020), French (Gilet et al., 2013), Farsi (Yaghoubi Jami & Wind, 2022), Swedish (Cliffordson, 2002), Korean (Kang et al., 2009), Russian (Budagovskaia et al., 2017), and Italian (Diotaiuti et al., 2021; Ingoglia et al., 2016). For Spanish-speaking populations, adaptations have been conducted in Spain (Lucas-Molina et al., 2017; Mestre-Escrivá et al., 2004; Pérez-Albéniz et al., 2003), Argentina (Muller et al., 2015), Colombia (Arenas-Estevez et al., 2021; Garcia-Barrera et al., 2017), and Chile (Fernández et al., 2011).

In the Spanish adaptation used in the present study (Pérez Albéniz et al., 2003), adequate reliability values were obtained using Cronbach's Alpha as the parameter. In a sample of 515 students, for males, values of .73 were obtained for PT, .76 for F, .68 for EC (including item 13), and .70 for PD. For females, values of .75 were obtained for PT and F, .70 for EC (including item 13), and .72 for PD.

In sum, the IRI is characterized by proposing a multidimensional assessment of empathy, contemplating affective and cognitive aspects across four subscales. Thus far, various psychometric studies have failed to yield consistent results regarding this model. Furthermore, several studies have demonstrated issues with the reverse-worded items. This study aims to analyse evidence of content, construct, and convergent validity of the Spanish-language version of the IRI (Davis, 1980; Pérez Albéniz et al., 2003) in a sample of Uruguayan adults. Specifically, we aim to test the fit of the four-factor model in this sample, along with other relevant models proposed in the literature (Chrysikou & Thompson, 2016; Cliffordson, 2002; Hawk et al., 2013; Pulos et al., 2004).

Method

Participants

The sample comprised 858 Uruguayan participants, 640 females and 218 males, ranging from 18 to 90 years old (M = 34.27; SD = 15.43), with medium to high socioeconomic level. Of the total participants, 59.6 % completed secondary education and 40.4 % completed tertiary education.

Instruments

Interpersonal Reactivity Index (IRI) (Davis, 1983; Pérez Albéniz et al., 2003): a self-report scale used for studying empathy that consists of four subscales, with seven items each that independently assess cognitive and affective aspects of the construct. Empathic Concern (EC, items 2, 4, 9, 13, 14, 18, 20, 22) and Personal Distress (PD, items 6, 10, 17, 19, 24, 27) subscales assess the affective dimension, whereas Perspective Taking (PT, items 3, 8, 11, 15, 21, 25, 28) and Fantasy (F, items 1, 5, 7, 12, 16, 23,

26) evaluate the cognitive aspects of empathy. Respondents use a five-point Likert scale (A: *Does not describe me well* to E: *Describes me very well*). The original study (Davis, 1983) reported adequate reliability values for the four subscales (F = .75, PT = .75, EC = .72, and PD = .78). In this study the adapted Spanish version by Pérez-Albéniz et al. (2003) was used, which demonstrates reliability values ranging from .60 to .78, with an adjustment made to item 13 being part of the EC subscale, not in PD as in the original model.

Toronto Empathy Questionnaire (TEQ) (Spreng et al., 2009): a self-report questionnaire designed for assessing empathy that comprises 16 items, eight negatively worded and eight positively worded. Respondents use a five-point Likert scale (0: *never* to 4: *always*) to indicate the frequency with which they feel or act in a particular manner. Psychometric properties of the TEQ demonstrate a good fit to a unidimensional model, encompassing a single-factor structure comprising 16 items, each with a factor loading exceeding .40 and high reliability reported with the α value of Cronbach's .85 (Spreng et al., 2009). In this study we used the version translated into Spanish and validated for the Uruguayan population (Carballo et al., 2023). TEQ-R scale demonstrated the following fit indices: CFI = .932, TLI = .905, $\chi^2(20) = 213.58$, $p < .001$, $\chi^2/df = 10.65$, RMSEA = .106, and SRMR = .045. The reliability index (McDonald's Omega) was .82 for de full scale and .90 for the reduced one (TEQ-R). In this sample reliability was ω= .804 for de full scale and ω = .782 for TEQ-R.

Socioeconomic Level Index (INSE) (Perera, 2018): a questionnaire developed in Uruguay to assess the household's socioeconomic level. The reduced version consisting of six items was used, allowing for sorting households based on their socioeconomic level, inferring consumption capacity from socio-demographic information and possession of both tangible and intangible assets.

Procedure

An instrumental study (Montero & León, 2002) was carried out using Classical Test Theory (Muñiz, 2010) aiming to adapt the IRI instrument to Uruguay.

The content validity study was based on the expert judgment procedure. The total number of items (28) was submitted for consideration of three expert judges in the area of affective processes: a doctoral psychologist specialized in affective regulation, a doctoral psychologist specialized in emotional regulation, and a psychologist with a doctorate in biology and expertise in empathy research. First, the sufficiency criterion was evaluated with the whole scale. Then, each item was assessed in terms of language clarity, theoretical coherence, and relevance, using a four-point Likert scale (Escobar-Pérez & Cuervo-Martínez, 2008).

The sample was collected through non-probabilistic sampling using the snowball technique, by disseminating the study through social networks and inviting university students to participate. After accessing and accepting the informed consent, participants completed the IRI, TEQ, and INSE scales using the Qualtrics platform (Qualtrics, 2021).

The procedure, consents, and protocols have been approved by the Ethics Committee of the Catholic University of Uruguay, complying with the country's regulations on human research as governed by Executive Decree 001-4573/2007 and Law No. 18331 on Data Privacy, regarding the protection of personal data.

Data Analysis

To assess the content validity of the instrument, the Content Validity Coefficient (CVC) proposed by Hernández-Nieto (2002) was employed, applying the least stringent criterion for item retention. CVC values below .70 were considered unsatisfactory, while values above .80 were regarded as highly satisfactory. Subsequently, a statistical analysis of the items was conducted, evaluating their psychometric quality through the calculation of the mean, variance, skewness, and kurtosis.

Regarding the confirmatory factor analysis (CFA), the weighted least squares mean and variance adjusted (WLSMV) estimator was used, as it is recommended for ordered categorical data due to its robustness and the fact that it does not assume a normal distribution of variables (Flora & Curran, 2004; Freiberg et al., 2013; Li, 2016). Several factorial structures were tested, including the original four-factor solution proposed by Davis (1980), a second-order factorial structure, a two-factor model distinguishing between the cognitive (CO) and affective (AF) dimensions, a three-factor model comprising PT, EC, and

PD, a three-dimensional structure including PT, F, and EC, and finally, a two-dimensional model including PT and EC. Finally, the same factorial structures were tested after removing the negative items.

Model fit was assessed using absolute fit indices, such as the chi-square test, the chi-square/degrees of freedom ratio, and the root mean square error of approximation (RMSEA), as well as incremental fit indices, including the comparative fit index (CFI) and the Tucker-Lewis index (TLI). In general, a good fit is indicated when the chi-square value is nonsignificant ($p \geq .05$) or when the chi-square/degrees of freedom ratio is below 2 or 3 (Schreiber et al., 2006). Furthermore, according to Hair et al. (2013), for samples exceeding 250 participants and scales containing between 12 and 29 items, CFI and TLI values equal to or greater than .92, as well as RMSEA values equal to or lower than .07, are recommended as cutoff points for evaluating model fit.

The reliability of the instrument was estimated using McDonald's omega coefficient (McDonald, 1999), with values ranging between .70 and .90 considered indicative of adequate reliability (Campo-Arias & Oviedo, 2008; Dunn et al., 2014).

Regarding concurrent validity evidence, the normality of the scores was assessed using the Kolmogorov-Smirnov (*K-S*) test, considering a distribution to be normal when $p > .05$. The scores obtained in the IRI dimensions were correlated with the full version of the Toronto Empathy Questionnaire (TEQ; Spreng et al., 2009) and the abbreviated version adapted for Uruguay (Carballo et al., 2023). The Spearman's rho coefficient was used for the correlation analysis, with interpretation criteria based on Akoglu (2018), where $r \geq .20$ indicates a low correlation, $r \geq .50$ a moderate correlation, and $r \geq .80$ a strong correlation. Finally, effect size was calculated using G*Power software (Faul et al., 2009), considering values between 0.10 and 0.30 as small effects, between 0.30 and 0.50 as moderate effects, between 0.50 and 0.80 as large effects, and values greater than 0.80 as very large effects (Ferguson, 2009).

All analyses were conducted using Mplus version 8.4 (Muthén & Muthén, 1998-2011) and SPSS v.29.

Results

Content validity and item analysis

Table 2 reports CVC and descriptive statistics with a normality test. Results of the agreement procedure among judges showed very good values in the criterion validity coefficients. All items exceeded the acceptable threshold (CVI > .70), and only four out of the 28 items had CVC values below .80; while the remaining 86 % exhibited excellent CVC indices > .80. Normality test, using the Kolmogorov-Smirnov statistic, rejected the normality hypothesis, indicating that IRI items do not conform to a normal distribution.

Mardia's analysis (Mardia, 1970) was employed to assess multivariate skewness and kurtosis. The results indicated a skewness coefficient of 57.36 ($df = 816$, $p = 1.00$) and a kurtosis coefficient of KI = 338.65, with a p-value < .000, suggesting a lack of multivariate normality in the data. Regarding the IRI items, skewness coefficients ranged between -2 and 2, except for item 18, which exhibited severe skewness (KI > 3) and a kurtosis value exceeding eight (Kline, 2005).

Table 2

Content validity and descriptive statistics of the 28 items of the IRI

Item	CLA	COH	REL	*M*	*SD*	As	Kurtosis	*K-S*
1	.96	.88	.80	3.17	1.27	.01	-1.16	.18**
2	.96	.96	.88	3.66	1.10	-.47	-0.63	.21**
3	.96	.96	.96	1.97	1.04	1.06	0.47	.26**
4	.88	.88	.88	1.75	1.07	1.44	1.21	.33**
5	.96	.88	.88	2.64	1.35	.35	-1.10	.20**
6	.88	.80	.80	2.34	1.19	.62	-0.59	.24**
7	.88	.80	.71	2.08	1.14	.89	-0.18	.25**
8	.96	.96	.96	3.18	1.15	-.05	-0.89	.17**
9	.96	.88	.80	3.91	1.00	-.58	-0.54	.22**
10	.96	.80	.80	2.84	1.28	.15	-1.12	.20**
11	.88	.88	.88	3.48	1.04	-.29	-0.54	.21**
12	.88	.80	.71	1.85	1.09	1.31	0.95	.28**
13	.88	.96	.96	1.66	1.02	1.54	1.47	.36**
14	.88	.80	.80	1.64	0.95	1.66	2.34	.34**
15	.88	.80	.80	2.27	1.20	.77	-0.35	.25**
16	.96	.80	.80	2.06	1.20	.97	-0.09	.24**
17	.88	.80	.80	2.59	1.31	.41	-1.01	.23**
18	.88	.96	.88	1.26	0.70	3.23	10.59	.49**
19	.88	.80	.80	3.20	1.15	-.14	-0.88	.19**
20	.96	.71	.71	3.60	1.09	-.41	-0.69	.22**
21	.96	.88	.80	3.44	1.12	-.23	-0.81	.19**
22	.96	.71	.71	3.85	1.19	-.74	-0.51	.23**
23	.96	.96	.96	3.00	1.26	.01	-1.09	.18**
24	.88	.80	.80	1.62	0.88	1.69	2.95	.32**
25	.80	.96	.96	2.39	1.07	.50	-0.46	.23**
26	.96	.88	.88	2.92	1.32	.13	-1.17	.19**
27	.88	.80	.80	1.46	0.78	1.98	4.21	.39**
28	.96	.96	.96	3.06	1.17	.11	-0.95	.19**

Note. K-S: normality test Kolmogorov-Smirnov; CLA: clarity; COH: Coherence; REL: relevance. ** $p < .005$

Confirmatory Factor Analyses

Six different factorial structures were tested. Table 3 presents the fit indices for each model using the full IRI scale. None of the models exhibited a satisfactory fit. Given the previously reported issues with negatively worded items (Arenas-Estevez et al., 2021; Beven et al., 2004; Grimaldo et al., 2022; Lauterbach & Hosser, 2007; Palmese & Schmidt, 2013), the analyses were repeated after removing these items from the scale.

Table 4 presents the results of the confirmatory factor analysis (CFA) after removing the reversed items. Three models achieved acceptable fit indices: Model 1 (corresponding to the original four-factor model), Model 6 (excluding the PD subscale), and Model 7 (including the scales that assess the cognitive and affective dimensions of empathy).

Since the original model has the strongest theoretical and empirical support in literature, we decided to retain it for further analysis. Table 5 presents the final model, detailing the retained items within each subscale and their respective factor loadings.

Table 6 presents the descriptive statistics for the four subscales of the IRI for the overall group and by gender, based on the four-factor model without reversed items.

Table 3

Goodness fit indices for models of the complete IRI scale

Model	χ^2	df	χ^2/df	RMSEA	CFI	TLI	SRMR
I. Davis (1980)	2192.41	344	6.37	.08	.82	.81	.08
II. Unidimensional	6162.87	350	17.61	.14	.45	.41	.13
III. Second-order CFA	4742.75	350	13.55	.12	.29	.24	.14
IV. 2 Factors (CO/AF)	4891.22	349	14.01	.12	.57	.54	.13
V. 3 Factors (PT, EC, PD)	1461.89	186	7.86	.09	.82	.79	.85
VI. 3 Factors (PT, F, EC)	2935.76	185	15.87	.13	.65	.61	.09
VII. 2 Factors (PT, EC)	539.64	76	7.49	.08	.89	.87	.06

Note. CO: cognitive; AF: affective; χ^2: chi-squared; *df :* degree of freedom.

Table 4

Goodness fit indices for models of the IRI scale without negative items.

Model	χ^2	df	χ^2/df	RMSEA	CFI	TLI	SRMR
I. Davis (1980)	611.41	146	4.19	.061	.94	.94	.05
II. Unidimensional	3860.44	152	25.39	.169	.56	.51	.13
III. Second-order CFA	860.80	148	5.94	.075	.92	.90	.06
IV. 2 Factors (CO/AF)	2796.23	151	18.52	.143	.69	.65	.12
V. 3 Factors (PT, EC, PD)	1697.85	74	22.94	.160	.74	.68	.10
VI. 3 Factors (PT, F, EC)	307.32	74	4.15	.061	.96	.95	.04
VII. 2 Factors (PT, EC)	159.97	26	6.15	.077	.96	.94	.04

Note. CO: cognitive; AF: affective; χ^2: chi-squared; *df*: degree of freedom.

Table 5

Factor load, communality, uniqueness of each item

IRI Items	Factor loading				Com	Uni
	1	2	3	4		
Factor 1: Perspective Taking						
8.	.665				.442	.557
11.	.695				.483	.517
21.	.676				.457	.543
25.	.668				.446	.554
28.	.678				.460	.540
Factor 2: Fantasy						
1.		.443			.196	.803
5.		.733			.537	.462
16.		.719			.517	.483
23.		.853			.727	.272
26.		.762			.580	.419
Factor 3: Empathic Concern						
2.			.623		.388	.611
9.			.585		.342	.657
20.			.710		.504	.496
22.			.707		.500	.500
Factor 4: Personal Distress						
6.				.614	.377	.623
10.				.667	.445	.555
17.				.693	.480	.519
24.				.835	.697	.302
27.				.780	.608	.391

Note. Com: communality; Uni: Uniqueness.

Table 6

Descriptive statistics

	Total		Women		Men	
	M	*SD*	*M*	*SD*	*M*	*SD*
1.TP	15.5	4.03	15.6	4	15.3	4.13
2.F	13.8	4.74	14.2	4.71	12.68	4.65
3.PE	15.0	3.17	15.5	3.05	13.72	3.18
4.PD	10.9	3.9	11.1	3.95	10	3.64

Note. M: mean; *SD*: standard deviation.

Finally, Table 7 presents the convergent validity analysis of the IRI without negatively worded items, examining its correlations with both the full version of the TEQ empathy scale (Spreng et al., 2009) and its short version (Carballo et al., 2023), as well as the intercorrelations among its subscales. Significant positive correlations were found between the PT, F, and EC subscales, while no correlation was observed between PT and PD. The four-factor model for the scale without reversed items demonstrated adequate consistency, as indicated by McDonald's omega coefficients. The reliability of the four subscales ranged from .75 to .84, which is considered acceptable.

Table 7

Convergent validity study with Toronto Empathy Questionnaire (TEQ), reliability and intercorrelation of the subscales

Scale		1	2	3	4	TEQ		TEQ_R	
						r	*d*	*r*	*d*
1.	PT	**.81**	-	-	-	.34**	0.58	.33**	0.57
2.	F	.30**	**.83**	-	-	.20**	0.44	.14**	0.37
3.	EC	.38**	.37**	**.75**	-	.54**	0.73	.55**	0.74
4.	PD	.02	.34**	.21**	**.84**	-.06	-	-.06	-

Note. PT: perspective taking; F: fantasy; EC: emphatic concern; PD: personal distress; TEQ_R: reduced scale. McDonald omega of each IRI subscale is presented on the main diagonal.

***p* < .01

Discussion

The study of empathy has gained significant relevance in psychology, impacting the way it is conceptualized and measured. A multidimensional approach is key for accurately assessing empathy, making the validation of widely used instruments like the Interpersonal Reactivity Index (IRI) crucial. The IRI, developed in 1980, evaluates both affective and cognitive empathy, though its original version treats these dimensions separately. Despite its widespread use, studies attempting to replicate the original factorial structure have yielded inconsistent results, with some showing poor fit indices. Psychometric research suggests that modifications to certain items are often necessary to achieve better model fit, and alternative factorial structures have also been proposed.

This study aimed to test the original four-factor model and some of the alternative models frequently seen in the literature. The results did not show a good fit for any of the models using the complete 28-item scale.

Several studies had previously noticed issues with the inverted items on this scale, achieving better results upon their removal (Arenas-Estevez et al., 2021; Beven et al., 2004; Braun et al., 2015; García-Barrera et al., 2017; Grimaldo et al., 2022; Murphy et al., 2020; Palmese & Schmidt, 2013). The inclusion of inverted items in psychometric scales has been heavily criticized. It is common in scale construction to formulate positively and negatively worded items to avoid response acquiescence or biases, although this strategy introduces more problems than solutions (Muñiz & Fonseca-Pedrero, 2019; Suárez et al., 2018). The effects of inverted items on the factorial structure of the scale have already been studied (Tomás et al., 2012), mostly concluding that these scales end up being affected by method variance (Conway, 2002). Method variance is a form of systematic error, introducing extraneous variables related to the measurement method rather than the trait being measured (Campbell & Fiske, 1959) impacting the psychometric characteristics of the scale (Tomás et al., 2010). Particularly, reliability deteriorates, and the unidimensionality of each evaluated dimension of the test is compromised by secondary sources of variance (Suárez et al., 2018; Woods, 2006), often resulting in the emergence of spurious factors or method factors that are not substantially meaningful concerning their semantic representation (Woods, 2006).

In Spanish, issues with reverse-coded items have been found to be more significant than in other languages, and it has been recommended to exclude them from questionnaires to avoid compromising the validity of the scale. Some explanations of why these items may be more problematic in Spanish point to the grammatical complexity of the language, which makes negative sentences more difficult to interpret. Additionally, reverse-coded items tend to increase cognitive load, which is further heightened in Spanish, potentially leading to misinterpretation. Furthermore, Romance languages generally avoid the use of negative constructions in everyday speech, making such items less intuitive for respondents (Venta et al., 2022). Another complicating factor is the translation of quantity and frequency adverbs such as "very," "usually," and "often." These could alter the perceived intensity of an action within a sentence. For instance, the English adverb "usually" may convey a different intensity than its Spanish equivalent, "normalmente." Such discrepancies can distort self-reported responses—an issue that is critical for validating translated constructs but has received little attention in research (Arenas-Estevez et al., 2021). The authors emphasizes that these translation challenges can significantly impact how

empathy constructs are understood in Spanish-speaking populations, further complicating the psychometric properties of tools like the IRI.

Considering all these antecedents and the poor fit obtained in the present study, the inverted items were removed and the five models with the reduced scale were retested. In this case, a good fit was found for two models: the original four-factor model, and a model that eliminates the PD subscale. Our results confirm previous studies on the IRI (Arena-Estevez et al., 2021; García-Barrera et al., 2017; Grimaldo et al., 2022), highlighting the need to review reverse-coded items in Spanish versions of the scale.

Regarding PD subscale, it has been previously questioned by other researchers (Koller & Lamm, 2015; Murphy et al., 2020), together with its theoretical relevance, which has also been called into question concerning its inclusion as a central aspect of empathy, considering that it might measure an aspect more related to emotional dysregulation or neuroticism. Furthermore, a lack of correlation has been reported between this subscale and other measures of empathy (Carballo et al., 2023).

Taking into account the psychometric indicators, both models —the original and the three dimensional model without PD— achieved good fit. As it is an instrument with four independent subscales, the presence of the PD scale does not impair or limit the others; therefore, its elimination at this point might not be necessary. It is recommended to continue investigating its theoretical relevance, considering the observations made by some of the aforementioned authors.

When comparing mean scores by gender, we found results consistent with previous studies (e.g., Braun et al., 2015; Davis, 1983; Pang et al., 2023). Pang et al. (2023) conducted a comprehensive investigation into sex/gender differences in empathic ability through three distinct studies, employing both large-sample self-report questionnaires and electroencephalography (EEG) measures. While self-reports showed higher empathy scores in women, particularly in PD and EC, neurophysiological data revealed no significant differences in neural responses to others' pain, suggesting a possible influence of social biases on subjective responses. Aligned with prior research, in our sample, women tended to obtain higher self-reported empathy scores.

This study has limitations. The sample is not fully representative, as it is biased toward a higher proportion of women, young individuals, and those from middle to high socioeconomic backgrounds. However, it meets the recommendation of using a sample size greater than 400 for categorical data (Mundfrom et al., 2005). On the other hand, at the psychometric level, some researchers note that there are no clear suggestions regarding the application of fit indices when analysing categorical variables, and conventional cutoff rules for categorical data have not yet been adopted (Xia & Yang, 2019). Garrido et al. (2016) studied the performance of the four commonly used fit indices (CFI, TLI, RMSEA, and SRMR) for estimating the number of factors to retain with categorical data, comparing them with the current gold standard of fit and the parallel analysis procedure by Horn (1965). They found that the CFI and TLI indices provide nearly identical estimates and are accurate fit indices, followed one step lower by the RMSEA. They do not recommend using SRMR as it provides deficient estimates.

Many authors consider it excessive to solely emphasize statistics based on fit indices (Marsh et al., 2005). Although fit indices provide useful information for assessing model fit to data, there are several notable limitations. Simulation studies suggest that implications of cutoff values change when manipulating sample size and load (Stone, 2021). Stone (2021) suggests not to rely exclusively on conventional fit indices that rigidly assess model fit to data. Three procedures should be considered: analysing conventional fit indices, analysing the relative fit procedure by testing fit in different models and selecting the best-fitting one, and lastly, using theory and logic to determine which models better fit to select a theoretically justifiable model.

Considering the obtained results and the study limitations, it is recommended to review the cost-benefit of including inverted items. Future studies could work on directly wording the inverted items in the IRI. Moreover, it is encouraged to continue reviewing the theoretical and empirical relevance of the Personal Distress subscale and the adequacy of the original four-factor model.

Based on the results of this study, a shortened 16-item version of the IRI has been developed and is ready for use with the Uruguayan population. This marks the first short and valid psychometric instrument for assessing empathy from a multidimensional perspective in Uruguayan adults.

References

Ahuatzin-González, A., Martínez-Velázquez, E. S., García Aguilar, G., & Vázquez-Moreno, A. (2019). Propiedades psicométricas del Interpersonal Reactivity Index (IRI) en mexicanos universitarios. *Revista Iberoamericana de Psicología, 12*(1), 111-122.

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91-93. https://doi.org/10.1016/j.tjem.2018.08.001

Arenas-Estevez, L. F. A., Quiñonez, H. S. R., Aguilar, A. C., & García, L. A. P. (2021). Validación en español del Índice de Reactividad Interpersonal–IRI-en estudiantes universitarios colombianos. *Psychology, Society & Education*, *13*(3), 121-135. https://doi.org/10.25115/psye.v13i3.3307

Baldner, C., & McGinley, J. J. (2014). Correlational and exploratory factor analyses (EFA) of commonly used empathy questionnaires: New insights. *Motivation and Emotion*, *38*, 727-744. https://doi.org/10.25115/psye.v13i3.3307

Baldner, C., & McGinley, J. J. (2020). Self-report empathy scales lack consistency: Evidence from exploratory and confirmatory factor analysis. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, *27*(1). https://doi.org/10.4473/TPM27.1.7

Batchelder, L., Brosnan, M., & Ashwin, C. (2017). The development and validation of the empathy components questionnaire (ECQ). *PloS one*, *12*(1), e0169185. https://doi.org/10.1371/journal.pone.0169185

Beven, J. P., O'Brien-Malone, A., & Hall, G. (2004). Using the interpersonal reactivity index to assess empathy in violent offenders. *International Journal of Forensic Psychology*, *1*(2), 33-41.

Braun, S., Rosseel, Y., Kempenaers, C., Loas, G., & Linkowski, P. (2015). Self-report of empathy: A shortened French adaptation of the Interpersonal Reactivity Index (IRI) using two large Belgian samples. *Psychological Reports*, *117*(3), 735-753. https://doi.org/10.2466/08.02.PR0.117c23z6

Budagovskaia, N. A., Dobrovskaia, S. V., & Kariagina, T. D. (2017). Adapting M. Davis's multifactor empathy questionnaire. *Journal of Russian & East European Psychology*, *54*(6), 441-469. https://doi.org/10.1080/10610405.2017.1448179

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81. https://doi.org/10.1037/h0046016

Campo-Arias, A., & Oviedo, H. C. (2008). Propiedades psicométricas de una escala: la consistencia interna. *Revista de Salud Pública, 10*(5), 831-839. https://doi.org/10.1590/s0124-00642008000500015

Carballo, M., Costa-Ball, C. D., Bidegain, M., & Álvarez, J. (2023). Validación del Toronto Empathy Questionnaire (TEQ) en una muestra uruguaya. *Revista Iberoamericana de Diagnóstico y Evaluación-e Avaliação Psicológica*, (69), 101-115. https://doi.org/10.21865/RIDEP69.3.09

Chakrabarti, B., & Baron-Cohen, S. (2006). Empathizing: neurocognitive developmental mechanisms and individual differences. *Progress in Brain Research, 156*, 403-417. https://doi.org/10.1016/S0079-6123(06)56022-4

Chrysikou, E. G., & Thompson, W. J. (2016). Assessing cognitive and affective empathy through the interpersonal reactivity index: an argument against a two-factor model. *Assessment*, *23*(6), 769-777. https://doi.org/10.1177/1073191115607974

Cliffordson, C. (2002). The hierarchical structure of empathy: Dimensional organization and relations to social functioning. *Scandinavian Journal of Psychology*, *43*(1), 49-59. https://doi.org/10.1111/1467-9450.00268

Conway, J. M. (2002). Method variance and method bias in industrial and organizational psychology. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 344-365). Blackwell Publishers Inc.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *Journal of Personality and Social Psychology, 44*(1), 113-126. https://doi.org/10.1037/0022-3514.44.1.113

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*(1), 113. https://doi.org/10.1037/0022-3514.44.1.113

De Corte, K., Buysse, A., Verhofstadt, L. L., Roeyers, H., Ponnet, K., & Davis, M. H. (2007). Measuring empathic tendencies: Reliability and validity of the Dutch version of the Interpersonal Reactivity Index. *Psychologica Belgica*, *47*(4), 235. https://doi.org/10.5334/pb-47-4-235

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews, 3*(2), 71-100. https://doi.org/10.1177/1534582304267187

Diotaiuti, P., Valente, G., Mancone, S., Grambone, A., & Chirico, A. (2021). Metric goodness and measurement invariance of the Italian brief version of interpersonal reactivity index: A study with young adults. *Frontiers in Psychology*, *12*, 773363. https://doi.org/10.3389/fpsyg.2021.773363

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399-412. https://doi.org/10.1111/bjop.12046

Escobar-Pérez, J., & Cuervo-Martínez, Á. (2008). Validez de contenido y juicio de expertos: una aproximación a su utilización. *Avances en medición*, *6*(1), 27-36.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149-1160. https://doi.org/10.3758/brm.41.4.1149

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532-538. https://doi.org/10.1037/a0015808

Fernández, A. M., Dufey, M., & Kramp, U. (2011). Testing the psychometric properties of the Interpersonal Reactivity Index (IRI) in Chile. *European Journal of Psychological Assessment*, *27*(3), 179-185. https://doi.org/10.1027/1015-5759/a000065

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, *9*(4), 466-491. https://doi.org/10.1037/1082-989X.9.4.466

Freiberg, A., Stover, J. B., de la Iglesia, G., & Fernández Liporace, M. (2013). Polychoric and tetrachoric correlations in exploratory and confirmatory factorial studies. *Ciencias Psicológicas*, *7*(2), 151-164. https://doi.org/10.22235/cp.v7i1.1057

Garcia-Barrera, M. A., Karr, J. E., Trujillo-Orrego, N., Trujillo-Orrego, S., & Pineda, D. A. (2017). Evaluating empathy in Colombian ex-combatants: Examination of the internal structure of the Interpersonal Reactivity Index (IRI) in Spanish. *Psychological Assessment*, *29*(1), 116. https://doi.org/10.1037/pas0000331

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods*, *21*(1), 93. https://doi.org/10.1037/met0000064

Gilet, A. L., Mella, N., Studer, J., Grühn, D., & Labouvie-Vief, G. (2013). Assessing dispositional empathy in adults: A French validation of the Interpersonal Reactivity Index (IRI). *Canadian Journal of Behavioural Science / Revue Canadienne des Sciences du Comportement*, *45*(1), 42. https://doi.org/10.1037/a0030425

Grimaldo, M., Correa-Rojas, J., Manzanares Medina, E., & Macavilca Milera, K. (2022). Validez e invarianza factorial del Índice de Reactividad Interpersonal en universitarios peruanos. *Ciencias Psicológicas*, *16*(2). https://doi.org/10.22235/cp.v16i2.2810

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). *Multivariate data analysis* (7th ed.). Pearson Education.

Haladyna, T. M. (2002). Research to improve large-scale testing. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 483-497). Routledge.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. Routledge. https://doi.org/10.4324/9780203850381

Hawk, S. T., Keijsers, L., Branje, S. T. J., van der Graaff, J., de Wied, M., & Meeus, W. H. J. (2013). Examining the Interpersonal Reactivity Index (IRI) among early and late adolescents and their mothers. *Journal of Personality Assessment, 95*(1), 96-106. https://doi.org/10.1080/00223891.2012.696080

Hernández-Nieto, R. A. (2002). *Contributions to statistical analysis*. Universidad de Los Andes.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179-185. https://doi.org/10.1007/BF02289447

Ilgunaite, G., Giromini, L., & Di Girolamo, M. (2017). Measuring empathy: A literature review of available tools. *BPA-Applied Psychology Bulletin (Bollettino di Psicologia Applicata)*, *65*(280). https://doi.org/10.14744/nci.2022.55649

Ingoglia, S., Lo Coco, A., & Albiero, P. (2016). Development of a brief form of the Interpersonal Reactivity Index (B–IRI). *Journal of Personality Assessment*, *98*(5), 461-471. https://doi.org/10.1080/00223891.2016.1149858

Kang, I., Kee, S., Kim, S. E., Jeong, B., Hwang, J. H., Song, J. E., & Kim, J. W. (2009). Reliability and validity of the Korean-version of Interpersonal Reactivity Index. *Journal of Korean Neuropsychiatric Association*, 352-358. https://doi.org/10.1007/s12564-019-09621-0

Kline, P. (2005). *A handbook of test construction: Introduction to psychometric design*. Routledge.

Koller, I., & Lamm, C. (2015). Item response model investigation of the (German) interpersonal reactivity index empathy questionnaire. *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000227

Lauterbach, O., & Hosser, D. (2007). Assessing empathy in prisoners-A shortened version of the Interpersonal Reactivity Index. *Swiss Journal of Psychology*, *66*(2), 91-101. https://doi.org/10.1024/1421-0185.66.2.91

Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S., & David, A. S. (2004). Measuring empathy: Reliability and validity of the Empathy Quotient. *Psychological Medicine, 34*(5), 911-920. https://doi.org/10.1017/S0033291703001624

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*, 936-949. https://doi.org/10.3758/s13428-015-0619-7

Limpo, T., Alves, R. A., & Castro, S. L. (2010). Medir a empatia: Adaptação portuguesa do Índice de Reactividade Interpessoal. *Laboratório de Psicologia*, *8*, 171-184. https://doi.org/10.14417/lp.640

Lucas-Molina, B., Pérez-Albéniz, A., Ortuño-Sierra, J., & Fonseca-Pedrero, E. (2017). Dimensional structure and measurement invariance of the Interpersonal Reactivity Index (IRI) across gender. *Psicothema*, *29*(4), 590-595. https://doi.org/10.7334/psicothema2017.19

Lundgren, O., Garvin, P., Andersson, G., Jonasson, L., & Kristenson, M. (2018). Inverted items and validity: A psychobiological evaluation of two measures of psychological resources and one depression scale. *Health Psychology Open, 5*(1), 2055102918755045. https://doi.org/10.1177/2055102918755045

Manarte, L. F., & Andrade, A. R. (2018). Comparative analysis and validation of the Portuguese version of the Interpersonal Reactivity Index. *Psilogos*, *16*(1), 45-59. https://doi.org/10.25752/psi.14905

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519-530. https://doi.org/10.2307/2334770

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of Fit in Structural Equation Models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275-340). Lawrence Erlbaum Associates.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

Mestre-Escrivá, V., Frías Navarro, M. D., & Samper García, P. (2004). La medida de la empatía: análisis del Interpersonal Reactivity Index. *Psicothema*, 255-260.

Montero, I., & León, O. G. (2002). Clasificación y descripción de las metodologías de investigación en Psicología. *International Journal of Clinical and Health Psychology*, *2*(3), 503-508.

Muller, M. E., Ungaretti, J., & Etchezahar, E. D. (2015). Evaluación multidimensional de la empatía: Adaptación del Interpersonal Reactivity Index (IRI) al contexto argentino. *Revista de Investigación en Psicología Social, 3*(1), 42-52.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(2), 159-168. https://doi.org/10.1207/s15327574ijt0502_4

Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems. *Papeles del Psicólogo*, *31*(1), 57-66.

Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. *Psicothema*, *31*(1), 7. https://doi.org/10.7334/psicothema2018.291

Murphy, B. A., Costello, T. H., Watts, A. L., Cheong, Y. F., Berg, J. M., & Lilienfeld, S. O. (2020). Strengths and weaknesses of two empathy measures: A comparison of the measurement precision, construct validity, and incremental validity of two multidimensional indices. *Assessment*, *27*(2), 246-260. https://doi.org/10.1177/1073191118777636

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus User's Guide* (6th ed.). Muthén & Muthén.

Palmese, F., & Schmidt, S. (2013). Empathy during life span. *TPM. Testing, Psychometrics, Methodology in Applied Psychology, 20*(2), 169-183. https://doi.org/10.4473/TPM20.2.5

Pang, C., Li, W., Zhou, Y., Gao, T., & Han, S. (2023). Are women more empathetic than men? Questionnaire and EEG estimations of sex/gender differences in empathic ability. *Social Cognitive and Affective Neuroscience, 18*(1), nsad008. https://doi.org/10.1093/scan/nsad008

Paulus, C. (2009). Der Saarbrücker Persönlichkeitsfragebogen SPF (IRI) zur Messung von Empathie. *Psychometrische Evaluation der deutschen Version des Interpersonal Reactivity Index*. https://doi.org/10.23668/psycharchives.9249

Perera, M. (2018). Índice de nivel socioeconómico (INSE). *Centro de Investigaciones Económicas, Montevideo, Uruguay*. https://ceismu.org/site/wp-content/uploads/INSE-2018-documento-final.pdf

Pérez-Albéniz, A., De Paúl, J., Etxeberría, J., Montes, M. P., & Torres, E. (2003). Adaptación de interpersonal reactivity index (IRI) al español. *Psicothema*, 267-272.

Preston, S. D., & De Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences, 25*(1), 1-20. https://doi.org/10.1017/s0140525x02000018

Pulos, S., Elison, J., & Lennon, R. (2004). The hierarchical structure of the Interpersonal Reactivity Index. *Social Behavior & Personality: an International Journal*, *32*(4). https://doi.org/10.2224/sbp.2004.32.4.355

Qualtrics. (2021). *Qualtrics XM* [Software de encuesta]. Qualtrics. https://www.qualtrics.com

Rajput, S., Puranik, M. P., Shanbhag, N., & Kumar, A. (2020). Factors affecting empathy among Indian dentists. *Indian Journal of Dental Research*, *31*(1), 14-21. https://doi.org/10.4103/ijdr.ijdr_365_18

Sampaio, L. R., Guimarães, PRB, Camino, CP dos S., Formiga, N. S., & Menezes, I. G. (2011). Estudios sobre la dimensionalidad de la empatía: traducción y adaptación del Índice de Reactividad Interpersonal (IRI). *Psicosis, 42*(1), 67-76.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, 323-337. https://doi.org/10.3200/JOER.99.6.323-338

Shiramizu, V. K. M., & Yamamoto, M. E. (2018). *Validation of the Empathy Index in a Brazilian Sample*. https://doi.org/10.31234/osf.io/zwu26

Siu, A. M., & Shek, D. T. (2005). Validation of the interpersonal reactivity index in a Chinese context. *Research on Social Work Practice*, *15*(2), 118-126. https://doi.org/10.1177/1049731504270384

Sonderen, E. V., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PloS one*, *8*(7), e68967. https://doi.org/10.1371/journal.pone.0068967

Spreng, R. N., McKinnon, M. C., Mar, R. A., & Levine, B. (2009). The Toronto Empathy Questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of Personality Assessment, 91*(1), 62-71. https://doi.org/10.1080/00223890802484381

Stone, B. M. (2021). The ethical use of fit indices in structural equation modeling: Recommendations for psychologists. *Frontiers in Psychology*, *12*, 783226. https://doi.org/10.3389/fpsyg.2021.783226

Suárez, J., Pedrosa, I., Lozano, L. M., García Cueto, E., Cuesta Izquierdo, M., & Muñiz Fernández, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema, 30*. https://doi.org/10.7334/psicothema2018.33

Tomás, J. M., Meléndez, J. C., Oliver, A., Navarro, E., & Zaragoza, G. (2010). Efectos de método en las escalas de Ryff: Un estudio en población de personas mayores. *Psicológica, 31*, 383-400.

Tomás, J. M., Sancho Requena, P., Oliver Germes, A., Galiana Llinares, L., & Meléndez Moral, J. C. (2012). Efectos de método asociados a ítems invertidos vs. ítems en negativo. *Revista Mexicana de Psicología*, *29*(2), 105-115. https://psycnet.apa.org/record/2013-19813-013

Venta, A., Bailey, C. A., Walker, J., Mercado, A., Colunga-Rodriguez, C., Ángel-González, M., & Dávalos-Picazo, G. (2022). Reverse-coded items do not work in Spanish: Data from four samples using established measures. *Frontiers in Psychology*, *13*, 828037. https://doi.org/10.3389/fpsyg.2022.828037

Wang, Y., Li, Y., Xiao, W., Fu, Y., & Jie, J. (2020). Investigation on the rationality of the extant ways of scoring the interpersonal reactivity index based on confirmatory factor analysis. *Frontiers in Psychology*, *11*, 1086. https://doi.org/10.3389/fpsyg.2020.01086

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*, 186-191. https://doi.org/10.1007/s10862-005-9004-7

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior research methods*, *51*, 409-428. https://doi.org/10.3758/s13428-018-1055-2

Yaghoubi Jami, P., & Wind, S. A. (2022). Evaluating the psychometric properties of a proposed Farsi version of the Interpersonal Reactivity Index using item response theory. *Research on Social Work Practice*, *32*(8), 1003-1018. https://doi.org/10.1177/10497315221089322

Yarnold, P. R., Bryant, F. B., Nightingale, S. D., & Martin, G. J. (1996). Assessing physician empathy using the Interpersonal Reactivity Index: A measurement model and cross-sectional analysis. *Psychology, Health & Medicine*, *1*(2), 207-221. https://doi.org/10.1080/13548509608400019

Zhang, X., Noor, R., & Savalei, V. (2016). Examining the effect of reverse worded items on the factor structure of the Need for Cognition Scale. *PloS One, 11*(6), e0157795. https://doi.org/10.1371/journal.pone.0157795

Authors' contribution (CRediT Taxonomy): 1. Conceptualization; 2. Data curation; 3. Formal Analysis; 4. Funding acquisition; 5. Investigation; 6. Methodology; 7. Project administration; 8. Resources; 9. Software; 10. Supervision; 11. Validation; 12. Visualization; 13. Writing: original draft; 14. Writing: review & editing.
M. B. has contributed in l, 2, 3, 6, 9, 11, 13, 14; M. C. in 1, 2, 3, 6, 9, 11, 13, 14; D. C.-B. in 1, 2, 3, 6, 9, 11, 13, 14.

Scientific editor in charge: Dra. Cecilia Cracco.