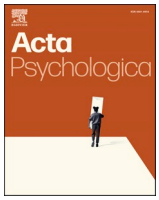
Acta Psychologica 256 (2025) 104993

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



Cross-cultural measurement invariance of the BFI-15p in university students from Argentina, Spain, and Peru

César Merino-Soto [a,b], Sergio Dominguez-Lara [b], Lucas Marcelo Rodriguez [c], Guillermo M. Chans [a,d,*], Manuel Marti-Vilar [e]

[a] Institute for the Future of Education, Tecnologico de Monterrey, Monterrey 64700, Mexico
[b] Instituto de Investigación FCCTP, Universidad de San Martín de Porres, Lima 34, Peru
[c] Instituto de Filosofía, Universidad Austral – CONICET, Facultad de Ciencias Biomédicas, Universidad Austral, Argentina
[d] School of Engineering and Sciences, Tecnologico de Monterrey, Mexico City 01389, Mexico
[e] Departamento de Psicología Básica, Universitat de València, 46010 Valencia, Spain

ARTICLE INFO

Keywords:
Exploratory structural equation modeling
Personality measurement
Big Five
Cross-cultural
Validity
Higher education
Innovational education

ABSTRACT

The Big Five Inventory–15p (BFI-15p) is a relatively new 15-item version developed for Hispanic populations; however, evidence of its cross-cultural equivalence is lacking. This study examined the clarity of item phrasing (Study 1) and measurement invariance (Study 2) in university students from Argentina, Spain, and Peru. Two studies were conducted: one to assess perceived item clarity (n = 70) and another to evaluate the internal structure (n = 737), including dimensionality, reliability, and measurement invariance. In Study 1, consistent item clarity was observed across the three groups, though one item required minor revision. In Study 2, the five-factor structure showed a good fit in the confirmatory factor analysis (CFA), while exploratory structural equation modeling (ESEM) provided slightly better results. However, differences between CFA and ESEM were not substantial, as cross-loadings were close to zero. Omega reliability varied across country groups but was predominantly acceptable for three-item subscales (approximately 0.60). Differential item functioning (DIF) by sex within each country was predominantly negligible. Measurement invariance was supported at the configural, metric, and scalar levels. The findings highlight the implications of measuring the Big Five model in university students across diverse cultural contexts.

1. Introduction

The study of individual differences has given rise to various personality theories, which can broadly be categorized into genetic and biological theories, psychodynamic models, humanistic theories, and trait-based personality approaches (Larsen & Buss, 2002; Schultz & Schultz, 2016; Weiner et al., 2013). Some theories have more substantial empirical support (Larsen & Buss, 2002). A key example is Cattell's work, which identified five core factors through the orthogonal rotation of 35 selected variables. This milestone laid the groundwork for the Big Five Factors (BFF) model (Goldberg, 1990).

Numerous studies have replicated the validity of the BFF model over the years (Digman & Inouye, 1986; Digman & Takemoto-Chock, 1981; McCrae & Costa, 1985, 1987), leading to the conclusion that the BFF model—also known by the acronym OCEAN (openness: OPEN, conscientiousness: CONS, extraversion: EXTR, agreeableness: AGRE, and neuroticism: NEUR)—explains key aspects of human development and functioning.

OCEAN describes the following factors: extraversion, which includes sociability, leadership, joy, energy, and fun-seeking; agreeableness, referring to prosocial traits such as generosity and cooperation; conscientiousness, describing willpower, organization, and purposeful behavior; neuroticism, encompassing tendencies to experience emotions like fear, resentment, and guilt, as well as patterns of behavior such as difficulty resisting impulses or coping with stress; and finally, openness, which relates to a need for variety, aesthetic sensitivity, and an open-minded attitude (McCrae et al., 2013; McCrae & Costa, 1985, 1987).

1.1. Measurement of the Big Five Factors: Big Five Inventory (BFI)

Historically, several widely recognized instruments have been developed to assess personality within the BFF framework. Among the

* Corresponding author at: Institute for the Future of Education, Tecnologico de Monterrey, Monterrey 64700, Mexico.
E-mail addresses: ca.ms@tec.mx (C. Merino-Soto), guillermo.chans@tec.mx (G.M. Chans), manuel.marti-vilar@uv.es (M. Marti-Vilar).

https://doi.org/10.1016/j.actpsy.2025.104993
Received 20 January 2025; Received in revised form 26 March 2025; Accepted 3 April 2025
Available online 15 April 2025

most notable are the Big Five Inventory (BFI-44) (John et al., 1991), consisting of 44 items, and the NEO Personality Inventory (NEO-PI), which has been developed into several subsequent versions (Costa Jr. & McCrae, 1992; Costa Sr. & McCrae, 1985; McCrae & Costa Jr, 1989). Comparatively, the BFI offers several advantages over the NEO-PI.

First, the BFI is not commercially restricted, making it freely accessible, unlike the NEO-PI, which requires a commercial license. While this may not be a significant issue in North America or Europe, it can hinder accessibility in regions with economic and institutional limitations. Additionally, the BFI features a standardized Spanish version devoid of regionalisms, ensuring better comprehension of its items (Benet-Martínez & John, 1998). Although the NEO-PI can be linguistically adapted, this process is complicated by its legal restrictions.

The BFI's shorter length—44 items—also makes it particularly suitable for contexts where time constraints are critical, such as organizational settings, compared to the NEO-PI's 200+ items. Furthermore, the absence of legal restrictions allows for the extraction of shorter versions of the BFI, enabling adaptations to specific contexts and facilitating the removal of less representative items for instrumental studies.

Evidence remains inconsistent regarding psychometric studies of the BFI in Peru, Argentina, and Spain (the focus of the present study). In Spain, no robust psychometric studies exist for the BFI. Although a Spanish translation aimed at adolescents was developed (Castro Solano & Casullo, 2001) in Argentina, its psychometric results were unsatisfactory, showing low reliability and factorial complexity (items influenced by more than one factor). Other Argentine studies (Castro-Solano, 2005; Genise et al., 2020) employed the Little Jiffy procedure (principal component analysis, eigenvalues greater than one, and varimax rotation), a method now discouraged for psychometric studies (Lloret-Segura et al., 2014).

In Peru, a psychometric study of the BFI (Dominguez-Lara et al., 2018) found a factorial structure similar to the original version, albeit with a reduction of 7 items. Despite these inconsistencies, the BFI continues to be used in Argentina (Casari et al., 2019; Castro Solano et al., 2022; Eidman et al., 2019; Fernández et al., 2022; Micillo et al., 2022; Stover et al., 2019), Spain (Herrera Torres & Perandones González, 2021; Molero-Jurado et al., 2020; Suárez & Marrero, 2020; Varchetta et al., 2023), and Peru (Dominguez-Lara & Merino-Soto, 2019; García-Flores et al., 2022; Manrique-Millones & Castillo-Blanco, 2021; Prada-Chapoñan et al., 2020).

However, some studies failed to corroborate the BFI's factorial structure. The initial (Castro Solano & Casullo, 2001) and brief versions (Rammstedt & John, 2007) in Argentina were used without validation in the target samples (Fernández et al., 2022; Micillo et al., 2022). In Spain, several studies used a brief German version (Rammstedt & John, 2007), and only one used the Peruvian short version (Varchetta et al., 2023), but none verified its internal structure. In Peru, both the extended (Dominguez-Lara et al., 2019; Prada-Chapoñan et al., 2020) and short versions (García-Flores et al., 2022; Manrique-Millones & Castillo-Blanco, 2021) were applied, relying on previous evidence of validity (Dominguez-Lara et al., 2019; Dominguez-Lara & Merino-Soto, 2018, 2019) but without verifying it in their datasets.

In summary, there appears to be an inappropriate validity induction in the use of the BFI, characterized by citing evidence from distinct groups while neglecting verification of its internal structure within the specific study samples (Merino-Soto & Angulo-Ramos, 2022; Merino-Soto & Calderón-De la Cruz, 2018).

1.2. Brief versions of the BFI

Shortened measures of the BFI have been developed, though some criticisms of these reduced versions are based on their ability to assess broad and complex constructs (such as the OCEAN factors) using only a few items. Nevertheless, they remain valuable in contexts with limited time availability or research projects involving multiple self-reports, leading to respondent fatigue (Soto & John, 2017) and insufficient effort or careless responses (IE/C) (Gibson & Bowling, 2019; Meade & Craig, 2012).

Two well-known brief versions of the BFI exist: the 10-item version (BFI-10) (Rammstedt & John, 2007) and the 15-item version (BFI-15). However, as outlined Dominguez-Lara and Merino-Soto (2018), these versions have certain limitations, which can be summarized as follows: a) the potential method bias associated with reverse-worded items (Fernández et al., 2023; Lance et al., 2010); b) the absence of a report on corrected correlations (due to shared elements) between the brief and full versions (Smith et al., 2000); c) the lack of reliability reporting; and d) using the outdated Little Jiffy approach (Lloret-Segura et al., 2014) in developing the BFI-15 (Gerlitz & Schupp, 2005).

International studies have revealed variability in the internal structure of brief scales like the BFI-10, which may stem from sampling, translation differences (Ludeke & Larsen, 2017), or cultural characteristics. For instance, the BFI-10 and BFI-15 versions exhibit acceptable psychometric evidence when the samples come from contexts similar to the original study (Courtois et al., 2020; Fernández et al., 2023; Guido et al., 2015; Hahn et al., 2012; Palsson et al., 2023; Rammstedt, 2007) or different samples from the same country over time (Rammstedt et al., 2024). However, when samples are drawn from cultural contexts different from the original, the internal structure often fails to demonstrate invariance (Balgiu, 2018; Dominguez-Lara & Merino-Soto, 2018; Kim et al., 2010; Kunnel John et al., 2019; Lovik et al., 2017; Pejić et al., 2014; Steyn & Ndofirepi, 2022; Stone et al., 2022). This pattern suggests that measures of this type are unlikely to maintain invariance across diverse cultural settings.

To address the methodological limitations identified in the initial studies of the BFI-10 and BFI-15, two alternative versions of 10 and 15 items were developed, specifically tailored for Spanish-speaking Peruvian adults (BFI-10p and BFI-15p) (Dominguez-Lara & Merino-Soto, 2018). These versions demonstrated adequate psychometric properties in independent studies conducted with other Spanish-speaking samples, including Peru (Dominguez-Lara & Merino-Soto, 2019), Mexico (Dominguez-Lara et al., 2022; Dominguez-Lara & Campos-Uscanga, 2024), and Chile (Dominguez-Lara et al., 2024). The reduced versions, particularly the BFI-15p, offer the advantage of reducing the likelihood of correlated uniquenesses among items (i.e., "methodological noise"; (Dominguez-Lara et al., 2022; Dominguez-Lara & Merino-Soto, 2019), due to a) their limited number of items, b) the elimination of method effects stemming from the absence of reverse-scored items, and c) the avoidance of factors derived from item parcels (Chiorri et al., 2016; Meade & Kroustalis, 2006). Together, these features help mitigate specific methodological challenges in defining latent dimensions (Fernández et al., 2023; Marsh, Nagengast, & Morin, 2013; Meade & Kroustalis, 2006).

However, the consistency of the psychometric properties of the BFI-15p remains at an emerging stage and requires evaluation of its internal structure's invariance or the absence of differential item functioning (DIF). In contexts where validated brief versions are unavailable for specific populations, researchers often rely on versions validated in other contexts, potentially inducing validity inappropriately (Merino-Soto & Angulo-Ramos, 2022; Merino-Soto & Calderón-De la Cruz, 2018). This issue is evident in studies conducted in Argentina and Spain (Fernández et al., 2023; Herrera Torres & Perandones González, 2021; Micillo et al., 2022; Molero-Jurado et al., 2020; Suárez & Marrero, 2020; Varchetta et al., 2023), which utilized the version validated in Germany (Rammstedt, 2007) without evidence of its internal structure in their samples, extrapolating psychometric properties from culturally distinct groups. Relying on these induced properties poses a risk of invalid conclusions. When applying measures in a new context, it is essential to evaluate their validity evidence (American Educational Research Association [AERA] et al., 2014). This is particularly critical given reports suggesting that some measures lack invariance, exhibiting variability in item-construct relationships or measurement intercepts across cultures (e.g., BFI-10) (Steyn & Ndofirepi, 2022).

1.3. The present study

In summary, several literature gaps support the present study's objectives. First, the internal structure of short versions of the BFI has often been analyzed using outdated or non-recommended methods (e.g., Little Jiffy). Second, these short versions have not been satisfactorily replicated across diverse cultural contexts. Third, existing evidence is primarily based on studies of the overall measurement model, without examining DIF within key groups—for example, in cross-national studies, equivalence between subgroups within each country is typically not assessed. Finally, research on the BFI-15p has been limited to samples from a single country, leaving its cross-cultural validity largely unexplored.

A key consequence of this lack of evidence on measurement invariance is the risk of inappropriate instrument use—specifically, applying versions of the BFI that lack psychometric validation for the target population (Fernández et al., 2023; Herrera Torres & Perandones González, 2021; Micillo et al., 2022; Molero-Jurado et al., 2020; Suárez & Marrero, 2020; Varchetta et al., 2023).

The BFF model has evidence supporting its etic nature (Allik, 2005; McCrae & Terracciano, 2005; Schmitt et al., 2007; Yamagata et al., 2006) in Western societies characterized by formal education systems, industrialized settings, high economic standards, and established democracies (Henrich et al., 2010). Its validity has also been primarily observed in groups with high levels of education (Rammstedt et al., 2013) and adult populations (Soto et al., 2008). However, the BFF model's replicability is inconsistent in contexts lacking these sociodemographic attributes (Fetvadjiev & Van der Vijver, 2015; Gurven et al., 2013). This inconsistency in the internal structure raises legitimate concerns regarding the potential lack of measurement invariance in its parameters (Fischer & Rudnev, 2024).

Furthermore, studies reporting intercultural similarities in constructs often fail to use methods that confirm measurement equivalence across groups (e.g., measurement invariance) (Church, 2008; Eigenhuis et al., 2015). Consequently, findings that reveal group differences (Galinha et al., 2016; Lippold et al., 2020; Weiss & Roberts, 2018) are typically obtained without evidence of measurement invariance, which presents significant limitations for rigorously studying intercultural differences (Church, 2008; Eigenhuis et al., 2015).

In this sense, measurement invariance is essential for determining the extent to which two constructs are understood and interpreted similarly across cultures. Without first establishing invariance, cross-cultural comparisons may yield misleading conclusions about supposed cultural differences (Chen, 2008; Fischer & Rudnev, 2024; Matsumoto, 2003), as such differences could stem from variations in psychometric calibration rather than substantive differences in the underlying construct. This issue is particularly relevant in developing and using psychological instruments in general and personality assessments in particular, where measurement invariance plays a fundamental role at both empirical and theoretical levels (Fischer & Rudnev, 2024). In this context, the objective of the present study is to evaluate the measurement invariance of the BFI-15p in samples from Argentina, Spain, and Peru. This cross-cultural study is essential for substantive group comparisons (Matsumoto, 2003). It represents an advance in the emic approach to abbreviated OCEAN factor measurement using the BFI-15p, a tool developed in a Latin American country, evaluated in other countries within the region, and, in this case, extended to a Spanish-speaking European country.

While current validation evidence for the BFI-15p suggests that it is a robust tool for samples from various Latin American countries in Peru (Dominguez-Lara & Merino-Soto, 2019), Mexico (Dominguez-Lara et al., 2022; Dominguez-Lara & Campos-Uscanga, 2024), and Chile (Dominguez-Lara et al., 2024), each analysis has been conducted independently. No direct cross-country comparisons based on primary data analysis with a post-Little Jiffy psychometric approach have been performed to date.

This study is justified from two perspectives. The first is the emic perspective (Fetvadjiev & Van der Vijver, 2015), as the study used a Spanish-language version of the instrument free of regionalisms, allowing for responses tied to the construct within specific cultural contexts. The second is the etic perspective (Fetvadjiev & Van der Vijver, 2015) since the BFI, while representing an instrument and construct external to the target groups, may potentially replicate the BFF model in the studied samples (Argentina, Spain, and Peru). The geographical and cultural proximity among the groups could explain this potential similarity in the responses forming the factors (Allik & McCrae, 2004). In both etic (universal) and emic (culture-specific) perspectives, favorable evidence of the validity and reliability of the BFI-15p is expected in these contexts.

Two studies were conducted. Within a pre-testing framework, the first evaluated an aspect of content validity evidence (American Educational Research Association [AERA] et al., 2014): the perceived clarity of each item (Ruel et al., 2016; Vogt et al., 2004). The second focused on assessing the internal structure, including dimensionality, group equivalence, and reliability (American Educational Research Association [AERA] et al., 2014). Both studies were carried out logically and sequentially with independent groups of university students from three Spanish-speaking countries: Argentina, Spain, and Peru.

2. Study 1

2.1. Objective

This study aimed to explore the clarity of the BFI-15p items among university students from three countries (Argentina, Spain, and Peru). This was conducted within a pre-testing framework to evaluate one of the content-related properties of each item (Ruel et al., 2016), specifically a preliminary assessment focused on how understandable the item wording was for the target participants. This quality was determined based on each participant's perception, aligning with content validity studies in which judges from the target population assess the content (Vogt et al., 2004). Instead of the authors rationally evaluating the readability and clarity of the content, as was done in another validity study using the BFI-2 (Soto & John, 2017), this was assessed empirically by exploring it with participant judges.

2.2. Method

2.2.1. Participants

The sample consisted of 70 students enrolled in private universities in Argentina ($n = 23$) and Peru ($n = 23$) and a public university in Spain ($n = 24$). The sampling was non-probabilistic but intentional, aiming to achieve a proportional distribution across countries, sexes, academic programs, and ages while covering various semesters. This approach ensured that the sample represented the natural heterogeneity of students in terms of academic performance and progress.

Table 1
Description of the sample for content clarity evaluation of the BFI-15p (Study 1).

	Argentina ($n = 23$)	Spain ($n = 24$)	Peru ($n = 23$)	Between-group comparison
Sex				$\chi^2 = 4.88$
Men	5	2	8	(df = 2)
Women	18	22	15	$p = .08$
Major				$\chi^2 = 2.93$
Psychology	13	18	12	(df = 2)
Non-psychology	10	6	11	$p = .23$
	M (SD)	*M* (SD)	*M* (SD)	KW-H = 5.85
Age	22.48 (2.67)	23.17 (3.52)	21.39 (3.32)	(df = 5) $p = .053$

Note. KW-H: Kruskal–Wallis H test. M: mean. SD: standard deviation. df: degrees of freedom.

Table 1 presents the demographic characteristics of the sample. The distribution was proportional across the sampled countries and did not differ from a uniform distribution ($\chi^2 = 0.029$, df = 5, $p = .95$). Additionally, there were no statistically significant differences in the distribution of sex, academic programs, or age across the countries (see Table 1). These differences were predominantly small in size for sex (Cramer-$V = 0.264$), educational programs (Cramer-$V = 0.205$), and age ($\eta^2 = 0.02$).

2.2.2. Instrument

2.2.2.1. Sociodemographic questionnaire. This questionnaire was used to gather sociodemographic information, including country, city, university, age, sex, level of education, year of study, and more.

2.2.2.2. Content Clarity Form for the BFI-15p. This questionnaire assessed participants' perceptions of the clarity of each item in the instrument. Responses were scaled from 1 (*unclear at all*) to 6 (*completely clear*). The instructions read as follows:

> "*You have been selected to provide your opinion on the clarity of the items in the questionnaires detailed below, which will later be applied to the population of our country. Clarity is understood as something understandable, clear, and comprehensible to the individuals who will respond to the instrument. Please note that the participants for this study will later include young people, both university students and those without such education. You must rate each item on a scale from 1 (not clear at all) to 6 (completely clear) to indicate the degree of clarity for you.*"

The Spanish version of the BFI-15p used in this form was derived from two different studies (Dominguez-Lara & Merino-Soto, 2018, 2019), which employed the Spanish-language version developed by Benet-Martínez and John (1998).

2.2.3. Procedure

2.2.3.1. Ethical considerations. This study is part of a larger project authorized by the University of Valencia Ethics Committee (code: H14820253925). Voluntary participation was obtained through signed informed consent, which outlined the study's objectives and the activities involved (e.g., answering various questionnaires), as well as the conditions of participation: anonymity of responses, absence of incentives for participation, freedom to withdraw at any time, non-harmful content, and the requirement to answer all questions. No electronic identifiers (e.g., IP addresses) were recorded.

2.2.3.2. Data collection. The study researchers and/or their assistants contacted university participants through direct requests during in-person classes or via their contact networks. Participants were directed to use a link to access an electronic form created in LimeSurvey. The sequence of materials presented to participants was informed consent, sociodemographic questionnaire, and clarity evaluation questionnaires.

2.2.4. Analysis

Aiken's *V* coefficient (Aiken, 1980) was used to quantify the perceived clarity of each item. This coefficient converts the average rating of responses into a metric ranging from 0.0 (minimum or complete absence of validity) to 1.0 (maximum validity). Statistical significance concerning a null hypothesis ($V = 0.60$) was evaluated using asymmetric confidence intervals (Penfield & Giacobbi, 2004). An average *V* was estimated as an overall indicator for each dimension. Two preliminary calculations were performed:

1. Homogeneity or agreement of ratings was assessed using the concentration of responses through the average deviation of the mean (Burke et al., 1999). This approach was a validity measure for calculating *V* due to the potential impact of outlier ratings. The cutoff point to distinguish appropriate concentration from excessive dispersion ($adm < adm_{cutoff}$) was set at $adm_{cutoff} = N_{opt}/6$ (Dunlap et al., 2003), where N_{opt} is the number of response options. For the analyzed data, adm_{cutoff} was 0.83.
2. A priori variability in ratings attributable to group membership (i.e., countries) was estimated using the intraclass correlation coefficient (ICC_1) (Bliese, 2000). Since this variability could be linearly associated with *V*, large ICC_1 values indicate heterogeneity in *V* coefficients among the compared country groups. Therefore, *V* coefficients should be estimated independently for each group. The software *ICAiken* (Merino-Soto & Livia-Segovia, 2009) and the *multilevel* package (Bliese, 2016) were used for the analyses.

2.3. Results

According to the results (Table 2), *adm* showed high values ($> adm_{cutoff}$) in a few items across each country. This was associated with the distribution of item ratings (as observed in the database) but not due to highly discrepant raters (outliers). Therefore, no participant raters were removed. Subsequently, the highest variability in ratings associated with countries occurred in Agreeableness items ($ICC_1 > 0.14$). At the same time, the trend was lower in Openness items ($ICC_1 < 0.09$). Finally, the items with the greatest heterogeneity in ICC_1 were from the Extraversion and Conscientiousness scales. These results generally imply that the *V* coefficients were estimated for each country group instead of treating them as a single sample and were ultimately compared.

Regarding comparative clarity across countries, the *V* coefficients and their confidence intervals revealed that the lowest perceived clarity occurred in the Peruvian sample compared to the Argentine and Spanish samples, which were similar to each other. However, in absolute terms, the items demonstrated sufficient clarity for all three groups, as the lower limit of *V* (V_{inf}) for each BFI-15p scale was predominantly above 0.60. The exceptions were items 1 (in the Spanish and Argentine samples) and 2 and 10 (in the Peruvian sample). Item 10 in Peru was slightly lower ($V_{inf} < 0.60$) but could be considered trivial. Finally, compared to the other items, item 1 had one of the lowest *V* coefficients in Argentina and Spain.

2.4. Discussion

This study aimed to compare the perceived clarity of the BFI-15p items among university students from Argentina, Spain, and Peru. Except for item 2, the remaining items were perceived as clear and understandable. This finding suggests that response variability due to item wording can be considered incidental and that the Spanish version of the BFI-15p is suitable for use in a cross-cultural Spanish-speaking context. Furthermore, given that the participants are university students, who are expected to engage in various activities emphasizing inductive and analytical reasoning through numerous verbal stimuli (de-la-Peña & Luque-Rojas, 2021; Yáñez Botello, 2016), understanding the content of the BFI-15p is likely not a significant challenge.

A noteworthy finding is that clarity ratings were comparatively lower in the Peruvian sample than in the Argentine and Spanish samples. This pattern suggests that the reading comprehension skills of the participants interacted with the wording of the BFI-15p items. This is a novel finding in the context of personality measurement in Spanish-speaking countries, leaving no definitive conjecture regarding its causes. However, it could be argued that the differential perception of item clarity is associated with the high proportion of students in Peru who exhibit a literal level of reading comprehension (de-la-Peña & Luque-Rojas, 2021). A deeper discussion of this issue requires further attention and falls beyond the scope of the present study.

Despite the variability in perceived clarity for two items, the wording of the BFI-15p items was evaluated as at least appropriate and highly

Table 2
Results of variability (ICC_1), concentration (ADm), and content clarity (V) for the BFI-15p (Study 1).

	ICC_1	Argentina					Spain					Peru				
		M	ADm	*V*	V_{Inf}	V_{Sup}	*M*	ADm	*V*	V_{Inf}	V_{Sup}	*M*	ADm	*V*	V_{Inf}	V_{Sup}
Extraversion																
bfi1	−0.00	3.39	1.06	0.598	0.512	0.678	3.63	0.93	0.658	0.574	0.732	3.87	0.86	0.718	0.635	0.788
bfi4	0.03	4.52	0.70	0.880	0.813	0.925	4.71	0.48	0.928	0.871	0.960	4.26	0.70	0.815	0.740	0.872
bfi15	0.18	4.74	0.43	0.935	0.879	0.966	4.67	0.72	0.918	0.859	0.953	4.00	0.78	0.750	0.669	0.816
M				0.80					0.83					0.76		
Neuroticism																
bfi2	0.12	4.26	0.83	0.815	0.740	0.872	4.37	0.93	0.843	0.772	0.894	3.30	1.50	0.575	0.489	0.656
bfi6	0.09	4.48	0.68	0.870	0.802	0.917	4.50	0.79	0.875	0.809	0.920	3.78	1.09	0.695	0.611	0.767
bfi10	0.09	4.52	0.70	0.880	0.813	0.925	4.25	0.93	0.813	0.739	0.869	3.61	1.22	0.653	0.568	0.729
M				0.86					0.77					0.64		
Agreeableness																
bfi3	0.15	4.70	0.47	0.925	0.867	0.959	4.75	0.39	0.938	0.884	0.967	4.13	0.75	0.783	0.704	0.845
bfi12	0.16	4.65	0.51	0.913	0.852	0.950	4.79	0.32	0.948	0.896	0.974	4.17	0.64	0.793	0.715	0.853
bfi14	0.16	4.78	0.35	0.945	0.892	0.973	4.79	0.34	0.948	0.896	0.974	4.26	0.70	0.815	0.740	0.872
M				0.93					0.89					0.80		
Conscientiousness																
bfi5	0.02	4.48	0.72	0.870	0.802	0.917	4.42	0.72	0.855	0.786	0.904	4.00	0.86	0.750	0.669	0.816
bfi8	0.11	4.78	0.35	0.945	0.892	0.973	4.50	0.66	0.875	0.809	0.920	4.13	0.75	0.783	0.704	0.845
bfi11	0.07	4.57	0.60	0.893	0.828	0.935	4.75	0.39	0.938	0.884	0.967	4.26	0.70	0.815	0.740	0.872
M				0.90					0.83					0.78		
Openness																
bfi7	0.08	4.65	0.51	0.913	0.852	0.950	4.63	0.56	0.908	0.847	0.946	4.13	0.83	0.783	0.704	0.845
bfi9	0.04	4.39	0.68	0.847	0.776	0.899	4.25	0.93	0.813	0.739	0.869	3.78	0.93	0.695	0.611	0.767
bfi13	0.06	4.43	0.73	0.857	0.787	0.907	3.96	0.89	0.740	0.661	0.806	3.78	0.78	0.695	0.611	0.767
M				0.87					0.75					0.72		

Note. ICC_1: intraclass correlation coefficient 1. *M*: mean. ADm: coefficient mean deviation from the mean. *V*, V_{inf}, V_{sup}: *V* coefficients to estimate perceived clarity, lower, and upper 90 % interval.

comprehensible. Nonetheless, item 1 may require rewording to minimize potential sources of variability related to item wording.

3. Study 2

3.1. Objective

The internal structure of the BFI-15p was assessed using samples from Argentina, Spain, and Peru. As briefly discussed in the introduction, previous literature on the BFI has generally relied on CFA. In contrast, the BFI-15p was developed using the ESEM approach (Asparouhov & Muthén, 2009), offering several advantages over EFA and CFA. These advantages include greater independence in factor correlations and identifying misspecifications caused by cross-loadings that are inappropriately constrained to zero in CFA modeling (Asparouhov & Muthén, 2009; Marsh et al., 2014).

Using ESEM to develop the BFI-15p aligns with the trend of achieving better statistical fit in personality measures than CFA (Booth & Hughes, 2014). Additionally, there is growing consistency in its application for modeling the BFF (Marsh et al., 2010; Marsh, Lüdtke, et al., 2013). Unlike some studies where principal component analysis persists, e.g., BFI-2 (Reyes Zamorano et al., 2014; Soto & John, 2017), this study follows a post-Little Jiffy trend (Lloret-Segura et al., 2014). Therefore, the present research modeled the internal structure of the instrument in a new context, assessing potential variations across three groups of university students (Argentina, Spain, and Peru) using the ESEM approach.

3.2. Method

3.2.1. Participants

The target population consisted of university students actively engaged in academic activities and enrolled in regular study programs. The total sample included 737 university students from Argentina (207, 28.1 %), Spain (310, 42.1 %), and Peru (220, 29.9 %). The demographic characteristics were distributed as follows: sex: male = 169 (22.9 %), female = 568 (77.1 %); age: M = 21.4 (SD = 4.12, Min = 16, Max = 53); semesters of study: 1 (138, 18.7 %), 2 (105, 14.2 %), 3 (188, 25.5 %), 4 (208, 28.2 %), 5 (98, 13.3 %); major: psychology, 623 (84.5 %), other fields, 114 (15.5 %). The response rate was 100 %. The predominance of women in psychology programs reflects a widely documented trend in higher education across countries (Chavez Irigoyen, 2023; Southwell, 2024; Willyard, 2011). The participating institutions are generally characterized by student populations from middle socioeconomic backgrounds.

The specific sample distribution for each country is detailed in Table 3. Statistically significant differences ($p < .01$) were observed for each demographic variable analyzed; however, effect sizes were small ($d < 0.20$), except for the distribution by major. Therefore, the baseline characteristics of the participants can be considered equivalent across the three groups.

3.2.2. Instrument

3.2.2.1. Sociodemographic questionnaire. A questionnaire was designed to collect sociodemographic information, including country, city, university, age, sex, level of studies, semester of study, and other relevant data.

3.2.2.2. Big Five Inventory - 15p (BFI-15p). The BFI-15p (Dominguez-Lara & Merino-Soto, 2018, 2019) is a brief measure derived from the Spanish adaptation of the Big Five Inventory (Benet-Martínez & John, 1998). It comprises 15 items distributed across five factors, with three

Table 3
Demographic distribution of participants from the three countries (Study 2).

	Argentina (n = 207)	Spain (n = 310)	Peru (n = 220)	Statistics tests and effect size
Sex				χ^2 (df: 2) = 27.40 **
Male	31	61	77	Cramer's V = 0.19 (95 % CI = 0.10, 0.26)
Female	176	249	143	
Study year				KW χ^2 (df: 2) = 111.84 **
1	51	69	18	η^2 = 0.15 (95 % CI = 0.11, 0.20)
2	37	54	14	
3	64	90	34	
4	33	90	85	
5	22	7	69	
Age				KW χ^2 (df: 2) = 10.57 **
M	20.6	21.6	21.8	η^2 = 0.01 (95 % CI = 0.00, 0.03)
SD	2.88	4.35	4.65	
Min	17	18	16	
Max	36	53	50	
Major				χ^2 (df: 2) = 130.18 **
Psychology	178	307	138	Cramer's V = 0.42 (95 % CI = 0.34, 0.49)
Other	29	3	82	

Note. KW χ^2: Kruskal-Wallis χ^2. df: degrees of freedom. η^2: effect size for KW χ^2. **$p < .01$. CI: Confidence interval. *M*: mean. SD: standard deviation.

items per dimension: extraversion, agreeableness, conscientiousness, neuroticism, and openness. All items use a response scale ranging from "strongly disagree" (1) to "strongly agree" (5). None of the items require recoding of the response scores. The same content from the instrument used in Study 1 was employed, except for item 1, which was modified based on the results of Study 1.

3.2.3. Procedure

3.2.3.1. Item modification. Following a conceptual analysis and consensus among the authors regarding potential unintended interpretations by respondents, it was decided to modify one of the items based on the results of Study 1. Specifically, in item 1, the original wording, which included the term "…*hablador*" (talkative/verbose), was replaced with "…*conversador*" (conversationalist). Since item 10 contained two compatible meanings, one of which clarified the other ("… temperamental, moody"), the original wording was retained.

3.2.3.2. Data collection. As in the previous study, researchers and/or their assistants contacted university participants through direct requests or existing contact networks. The selection process aimed to maximize heterogeneity in the participants' semesters of study. Participants were instructed to use a link to access an electronic form hosted on Google Forms. The materials were presented in the following order: informed consent, sociodemographic questionnaire, and construct-related questionnaires. No time limits were imposed for completing any part of the form.

3.2.4. Analysis

The analysis was conducted in three stages: a) Response bias detection: This stage aimed to identify potential biases associated with careless or insufficiently deliberate responses (IE/C) (Meade & Craig, 2012); b) Item comparative analysis: Items were analyzed within a comparative framework across the three groups; and, c) Internal structure analysis: Dimensionality, group equivalence, and reliability were examined (American Educational Research Association [AERA] et al., 2014). A profile analysis approach was used to observe potential differences or similarities in comparative results (Verhelst, 2012). This technique allows for assessing and quantifying the extent to which parameter patterns differ and/or are similar across groups.

3.2.4.1. Potential response bias. Response bias, expressed as the longest sequence of consecutively identical responses, was evaluated using the *longstring* index (Ls) (Johnson, 2005). This method, recommended for multidimensional measures (DeSimone et al., 2015), was selected because the BFI-15p items are dispersed (i.e., items evaluating the same dimension are not consecutive). Any long pattern of consecutively identical responses was flagged as suspicious. The detection criterion Ls was set at seven or more consecutive identical responses, spanning at least two dimensions of the BFI-15p. This criterion aligns with recommendations for datasets with a similar number of variables (Curran, 2016).

3.2.4.2. Item analysis. Descriptive statistics were calculated for the items. Additionally, item responses were compared using the non-parametric Cucconi test for k samples (MC_k) (Marozzi, 2014), an omnibus test designed to detect simultaneous between-group changes in the location and scale of item distributions. A statistically significant MC_k result indicates differences in location, scale, or both. Consequently, follow-up tests were conducted to identify specific differences in location (Kruskal-Wallis test) and scale (Fligner-Killeen test), including effect size estimators for both (ε^2) (Tomczak & Tomczak-Łukaszewska, 2014). MC_k was evaluated with a nominal $\alpha = 0.05$ and Bonferroni adjustment for each country (0.05/15 items = 0.003). Based on Field (2013), ε^2 was interpreted as very small (<0.02), small (<0.13), moderate (<0.26), and large (≥0.26). The *R* packages *coin* (Hothorn et al., 2008), *ksamples*, *MultiSampleCucconiTest* (Marozzi, 2014), and *psych* (Revelle, 2023) were used for these analyses.

3.2.5. Internal structure

3.2.5.1. Dimensionality. The item-construct relationship was modeled using structural equation modeling (SEM). The tested model was a five-factor correlated model based on the item distribution from previous studies (Dominguez-Lara & Merino-Soto, 2018, 2019). Two approaches were implemented: CFA and ESEM with target rotation (Asparouhov & Muthén, 2009). The key difference between these approaches is the estimation of cross-loadings, which occurs in ESEM (Asparouhov & Muthén, 2009). The CFA was modeled to obtain a baseline model, while the ESEM tested whether cross-loadings should not be ignored in modeling the BFI-15p. Both procedures are generally recommended when evaluating the internal structure of a multidimensional measure (Booth & Hughes, 2014).

The global fit of the models was examined using the chi-square goodness-of-fit test (χ^2) and approximate fit indices (AFI) (Kline, 2013). Specifically, the following criteria were considered: CFI > 0.90 (McDonald & Ho, 2002); RMSEA between 0.05 and 0.08 (Browne & Cudeck, 1993) and WRMR <1.0 (DiStefano et al., 2018). The WRMR was selected over the SRMR due to its greater suitability for models involving categorical data and small to moderate sample sizes (n < 1000). In contrast, the SRMR is typically recommended for continuous data and larger samples (Singer et al., 2013).

Local fit focused on the factor loadings of the items on their hypothesized factors (very low: <0.40, low: >0.40, medium: >0.60, high: >0.80) (Kline, 2013; Shi et al., 2019; Shi & Maydeu-Olivares, 2019). For cross-loadings, it was expected that they would be low (<0.20) or around zero.

In a profile analysis framework (Chiorri et al., 2016; Verhelst, 2012), the comparison between both models (CFA and ESEM) focused on the differences in a) factor loadings and b) the factor configuration. For factor loadings, we assessed whether their strength status changed according to the levels of local fit. For factors, this was evaluated by a) the correlation between the factor scores for each factor derived from the CFA and ESEM and b) comparing the interfactor correlation matrix

(Booth & Hughes, 2014; Chiorri et al., 2016; Marsh, Lüdtke, et al., 2013). Both strategies evaluated whether including cross-loadings modified their contribution to the definition of the latent construct (Booth & Hughes, 2014; Tóth-Király et al., 2017).

3.2.5.2. Equivalence. Equivalence between the countries was initially evaluated using the intra-group equivalence approach (within each country) and then the usual between-group equivalence approach (across countries).

3.2.5.3. Intra-group equivalence. Before investigating the internal structure, a DIF study was conducted within each country (intra-country) to assess potential intra-group heterogeneity based on the participants' sex. The sex grouping was chosen because it is commonly included in equivalence studies with OCEAN model measures. However, this analysis presented significant challenges, such as the small size of each subscale (3 items), the inequality of the compared groups (female-to-male ratio in Argentina, Spain, and Peru: 6:1, 4:1, and 2:1, respectively), and the relatively small sample size in each country and sex group, <300 (Scott et al., 2009). Given these conditions, and because the focus of this analysis was on probable differences in the average item responses, two non-parametric methods that are often used together to test for the existence and describe the size of DIF were chosen (Kim et al., 2007; Zwick et al., 1993). The statistical test M^2 (Mantel, 1963) was used, with a Bonferroni adjustment to the p-value = .01 (0.01/3 = 0.003). The standardized mean difference (SMD) (Dorans & Schmitt, 1991; Zwick et al., 1993) was used as a descriptive measure and effect size. The SMD can have negative values (higher scores in the focal group) or positive values (higher scores in the reference group). In estimating the SMD, instead of using the observed standard deviation for each item (Zwick et al., 1993) to standardize the comparison between items (Dorans & Schmitt, 1993; Wells, 2021), the maximum possible category value for each item (category 5, "strongly agree") was used. DIF levels, using SMD, were classified as (Wells, 2021): trivial or small (SMD < 0.05), moderate (SMD < 0.10), large (SMD ≥ 0.10). The *R MeasInv* program (Wells, 2023) was used.

3.2.5.4. Between-group equivalence. After analyzing the scale at the intra-group level, the next step was to assess whether the construct was represented in the same way among participants from Argentina, Spain, and Peru. To do this, a multiple-group confirmatory factor analysis (MGCFA) was implemented, involving the sequential introduction of several constraints (Pendergast et al., 2017): first, specifying internal structure equivalence (configural invariance); second, factor loading equivalence (weak invariance); and third, threshold equivalence (strong invariance). The recommendations by Chen (2007) were followed to evaluate measurement invariance, with a sample size of around 300 in each group (comparative fit index, CFI ≤ | −0.010|, root mean square error of approximation, RMSEA ≤0.015).

3.2.5.5. Reliability. The ω coefficient was estimated, which is appropriate for congeneric models (variability in the magnitude of factor loadings). As a complementary measure, the average proportional distance, APD (Sturman et al., 2009), was estimated, which is a response consistency measure independent of the number of items. For both coefficients, 95 % confidence intervals were calculated using 1000 *bootstrap* samples and the percentile method (Padilla & Divers, 2013).

3.2.5.6. Ethical aspects. This study, like the one described in Study 1, is part of a larger project authorized by the Ethics Committee of the Universitat de València (code: H14820253925). In all cases, before data collection, each participant was asked to sign an informed consent form detailing the scope of the study.

3.3. Results

3.3.1. Possible response biases

A total of 38 participants (5.1 %) exhibited between 8 (10 participants, 1.3 %) and 15 (2 participants, 0.2 %) consecutive identical responses. These participants were distributed as follows: 27 in Peru (71.0 %), 7 in Argentina (18.4 %), and 4 in Spain (10.5 %). These cases were excluded, resulting in a final effective sample size of 699.

3.3.2. Item analysis

The global MC_k test identified two statistically significant contrasts (Table 4) in items 4 and 9 (EXTR and OPEN, respectively). Further examination of these items revealed the following:

For item 4, the difference in location among the three groups was statistically significant (Kruskal-Wallis $\chi^2 = 15.73$, $p < .001$) but of small effect size ($\eta^2 = 0.02$, 95 % CI = 0.01, 0.05). The difference in dispersion was also statistically significant (Fligner-Killeen $\chi^2 = 7.25$, df = 2, p = .02) but similarly small ($\varepsilon^2 = 0.01$).

For item 9, the difference in location among the three groups was statistically significant (Kruskal-Wallis $\chi^2 = 25.91$, $p < .001$) with a small effect size ($\eta^2 = 0.04$, 95 % CI = 0.02, 0.07). The difference in dispersion was statistically significant (Fligner-Killeen $\chi^2 = 22.84$, df = 2, $p < .001$) but small ($\varepsilon^2 = 0.03$). Given these findings, no group-wise post-hoc analyses were conducted.

Subsequently, each country's vector of item means and standard deviations (SD) was correlated. The item mean correlations across the 15 items were statistically significant ($p < .001$) and high between all country pairs (Argentina-Spain: $r = 0.94$, 95 % CI = 0.83, 0.98; Argentina-Peru: $r = 0.92$, 95 % CI = 0.78, 0.97; Spain-Peru: $r = 0.95$, 95 % CI = 0.89, 0.98). A moderately similar pattern was observed for item variability (SD) correlations: Argentina-Spain ($r = 0.85$, 95 % CI = 0.59, 0.95), Argentina-Peru ($r = 0.90$, 95 % CI = 0.73, 0.97), and Spain-Peru ($r = 0.75$, 95 % CI = 0.38, 0.91).

Table 4
Descriptive for items analysis within-country (Study 2).

	Argentina		Spain		Peru		Multi-sample comparison	
	M	SD	*M*	SD	*M*	SD	MC_K	*p*
Extraversion								
bfi1	3.64	1.07	3.49	1.28	3.64	1.01	4.384	0.007
bfi4	3.52	0.92	3.82	0.86	3.80	0.87	6.500	0.002*
bfi15	3.68	1.17	3.92	1.01	3.72	1.12	3.918	0.028
Neuroticism								
bfi2	2.33	1.17	2.23	1.05	2.31	1.08	2.935	0.141
bfi6	3.42	1.10	3.15	1.15	3.13	1.08	3.307	0.035
bfi10	3.12	1.27	2.94	1.23	2.91	1.14	2.402	0.100
Agreeableness								
bfi3	4.03	0.66	4.03	0.64	4.08	0.78	18.399	0.064
bfi12	4.03	0.86	4.16	0.78	3.95	0.86	8.398	0.021
bfi14	4.29	0.76	4.14	0.86	4.17	0.77	7.41	0.320
Conscientiousness								
bfi5	4.01	0.80	3.98	0.87	4.09	0.80	5.281	0.490
bfi8	4.13	0.91	3.96	0.92	4.17	0.81	6.191	0.024
bfi11	3.50	1.00	3.24	1.02	3.54	0.94	5.046	0.007
Openness								
bfi7	3.77	1.03	3.79	0.92	4.02	0.95	5.492	0.011
bfi9	3.33	0.98	3.49	0.99	3.82	0.77	9.044	0.000*
bfi13	3.38	1.07	3.48	1.05	3.61	1.06	2.322	0.357

Note. *M*: mean. SD: standard deviation. MC_K: multi-sample Cucconi test. *$p <$ .003 (0.05 with Bonferroni adjustment).

3.3.3. Internal Structure

3.3.3.1. Dimensionality. The global model fit (WLSMV χ^2) differed substantially between the CFA and ESEM results (Table 5), both in the total sample and within each country. The fit indices (AFIs) consistently indicated a better fit for the ESEM model compared to CFA. The magnitude of improvement in AFIs varied (see "AFI ratio" in Table 5): the difference was more pronounced for WRMR, moderate for RMSEA, and small for CFI. Nonetheless, all indices converged in showing that the ESEM model provided superior fit.

3.3.3.2. CFA results. The CFA modeling showed that item loadings on their intended constructs demonstrated moderately high validity (factor loadings $M_{items} > 0.55$; see Table 6, "Intra-construct" header). At the country level, a few items had lower loadings ($M_{sample} < 0.60$), but most clustered around that value. Items 1 and 11 consistently showed the lowest loadings across the three samples. Overall, the CFA results suggest acceptable item–construct relations.

3.3.3.3. ESEM results. The ESEM modeling (Table 7) generally revealed a consistent pattern of loadings on the expected factors ("target loadings") at both the intra-construct and between-sample levels, aligning with the five OCEAN factors. In other words, the target loadings were substantially higher than cross-loadings ($M_{items} > M_{cross\_i}$, $M_{samples} > M_{cross\_s}$). However, as in the CFA, items 1 and 11 had the lowest average target loadings across all three countries. Additionally, item 11 exhibited a higher number of cross-loadings greater than |0.15|.

3.3.3.4. CFA vs. ESEM comparison. Following prior factorial profile analyses (Chiorri et al., 2016; Verhelst, 2012), the factorial distinctiveness of the CFA and ESEM models was compared using congruence coefficients (at both the factor and item levels) and Pearson correlations between factor scores (Table 8). Congruence at the factor level was high (> 0.94), and remained similarly high at the item level (≥ 0.94), although a few items—specifically items 2, 9, and 11—fell slightly below 0.95.

Regarding differences in inter-factor correlations (Table 9), the Argentine sample showed small differences between CFA and ESEM (median = 0.06; min = –0.06, max = 0.18), with the largest differences (between 0.10 and 0.20) occurring in OPEN and CONS. In the Spanish sample, the central difference was also small (median = 0.09), ranging from –0.04 to 0.39 in AGRE, with CONS also showing a moderate difference (greater than 0.20 but less than 0.25). Finally, in the Peruvian sample, differences were comparatively larger (min = –0.16, max = 0.85), with a median difference of 0.11.

3.3.4. Equivalence

The evaluation of equivalence was conducted within each country and subsequently between countries.

3.3.4.1. Intra-group equivalence. Before assessing cross-country (between-group) measurement invariance, equivalence was evaluated within the sex groups of each country (Table 10). Using the M^2 test, item 2 (NEUR) in Spain was the only one with a *p*-value close to the Bonferroni-adjusted criterion ($\alpha = 0.003$), but it did not meet the threshold. No DIF was detected in the remaining items using the M^2 test; however, some items consistently exhibited moderate SMD in NEUR across all three countries. The DIF of item 2 can be considered moderate based on its SMD.

3.3.4.2. Between-group equivalence. Regarding the measurement invariance analysis, the BFI-15p demonstrated an acceptable level of measurement invariance across countries based on the variability of fit indices (Table 11). Although the CFI exceeded 0.001 in the case of weak invariance, this excess is not practically significant and does not affect the decision.

3.3.5. Reliability

Due to the slight differences in the factor analysis parameters between the CFA and ESEM solutions, reliability and consistency estimates were derived from the CFA solution. The point estimates were predominantly above 0.60, except for three subscales: EXTR in the Spanish sample, NEUR in the Argentine sample, and OPEN in the Peruvian sample (Table 12). However, population-level values at the 95 % confidence interval were consistently above 0.65. Additionally, APD was consistently good for the AGRE score. At the same time, performance was moderately variable for the remaining samples and scales.

Nevertheless, discrepancies were observed between the estimators. For instance, while the point omega values for EXTR (Spanish sample) and OPEN (Peruvian sample) were below 0.60, their APD values fell within an excellent or acceptable range.

4. General discussion

This study represents the first cross-cultural examination of the BFI-15p (Dominguez-Lara & Merino-Soto, 2018, 2019), a new abbreviated version of the Big Five Inventory (Benet-Martínez & John, 1998)

Table 5
Results of ESEM and CFA fit for BFI-15 in 3 countries (Study 2).

	CFA						
	Total	Argentina	Spain	Peru			
χ^2 (df = 80)	309.543**	144.441**	180.882**	173.259**			
CFI	0.953	0.934	0.954	0.971			
RMSEA	0.064	0.064	0.064	0.078			
(90 % CI)	(0.057, 0.072)	(0.047, 0.080)	(0.052, 0.077)	(0.062, 0.094)			
WRMR	1.37	0.941	1.05	1.028			
	ESEM				AFI ratio		
	Total	Argentina	Spain	Peru	CFI	RMSEA	WRMR
χ^2 (df = 40)	44.612	71.340**	72.007**	94.669			
CFI	0.999	0.949	0.977	0.966	0.984		1005
RMSEA	0.013	0.064	0.052	0.087	1.000		0.896
	(0.0. 0.013)	(0.039, 0.087)	(0.032, 0.070)	(0.065, 0.110)			
WRMR	0.520	0.445	0.405	0.484	2.115		2.124

Note: ESEM: Exploratory Structural Equation Modeling. CFA: Confirmatory Factor Analysis. AFI Ratio: approximate fit index (AFI) ratio from two models CFA and ESEM, or AFI_{CFA}/AFI_{ESEM}. CFI: comparative fit index. RMSEA: root mean square error of approximation. WRMR: weighted root mean square residual. $^{**}p < .01$. df: degrees of freedom.

Table 6
Confirmatory factor analysis (CFA) for BFI-15p, in three countries (Study 2).

	Extraversion			Neuroticism			Agreeableness			Conscientiousness			Openness			
	Arg	Spa	Peru	Arg	Spa	Peru	Arg	Spa	Peru	Arg	Spa	Peru	Arg	Spa	Peru	Msample
bfi1	.455	.349	.649													.484
bfi4	.756	.673	.860													.763
bfi15	.690	.871	.857													.806
bfi2				.824	.704	.803										.777
bfi6				.439	.580	.500										.506
bfi10				.520	.663	.609										.597
bfi3							.597	.549	.623							.590
bfi12							.772	.715	.678							.722
bfi14							.734	.849	.996							.860
bfi5										.636	.765	.787				.729
bfi8										.805	.837	.676				.773
bfi11										.421	.478	.487				.462
bfi7													.643	.768	.374	.595
bfi9													.346	.462	.772	.527
bfi13													.712	.779	.593	.695
Intra-construct																
Mitems	.634	.631	.789	.594	.649	.637	.701	.704	.766	.621	.693	.650	.567	.670	.580	

Note. Arg: Argentina. Spa: Spain. Msample: factor loadings average across samples. Mitems: factor loadings average across items.

Table 7
Exploratory structural equations modeling (ESEM) for BFI-15p, in the three countries (Study 2).

	Extraversion			Neuroticism			Agreeableness			Conscientiousness			Openness				
	Arg	Spa	Peru	Arg	Spa	Peru	Arg	Spa	Peru	Arg	Spa	Peru	Arg	Spa	Peru	Msamples	Mcross_s
bfi1	**.432**	**.453**	**.926**	.085	.264	.141	.029	−.032	−.118	.096	.105	−.050	.057	−.097	−.102	.604	.032
bfi4	**.485**	**.711**	**.708**	−.145	−.135	−.163	.112	−.008	−.002	.212	.033	.121	.149	.180	.093	.635	.037
bfi15	**.996**	**.692**	**.731**	−.038	−.055	−.015	.015	.308	.221	−.140	−.099	−.144	−.057	−.060	.067	.806	.000
bfi2	−.183	−.248	.020	**.533**	**.651**	**.839**	−.107	.090	.058	−.132	−.113	−.006	.273	.124	−.032	.674	−.021
bfi6	−.041	.072	−.092	**.738**	**.693**	**.483**	.138	.006	.155	.200	.142	−.023	−.136	−.027	−.041	.638	.029
bfi10	.102	.157	.041	**.611**	**.649**	**.571**	−.017	−.098	−.173	−.084	−.093	.013	−.068	−.009	.215	.610	−.001
bfi3	−.003	.142	.199	−.045	.072	.085	**.637**	**.519**	**.488**	−.053	−.017	.214	−.011	−.021	−.122	.548	.037
bfi12	.038	.041	−.058	−.026	−.035	−.049	**.596**	**.742**	**.748**	.148	.006	−.042	.130	.035	.055	.695	.020
bfi14	.070	.110	.065	.086	−.029	.000	**.871**	**.705**	**.850**	−.083	.130	.097	−.003	.021	.127	.809	.049
bfi5	−.140	−.043	.128	−.076	−.024	−.139	.214	.043	.291	**.525**	**.738**	**.438**	−.086	.034	−.001	.567	.017
bfi8	−.023	.016	−.064	−.029	−.082	.048	−.059	−.009	.061	**.868**	**.844**	**.965**	.045	.018	.043	.892	−.003
bfi11	.221	−.027	.116	.154	.112	−.168	−.127	.084	.049	**.492**	**.495**	**.163**	.000	−.025	.206	.383	.050
bfi7	.030	−.096	−.053	.108	.068	.171	.030	.075	−.001	−.044	−.016	.148	**.667**	**.847**	**.495**	.670	.035
bfi9	.064	.279	.140	−.087	.025	−.075	−.076	−.227	−.069	.134	.062	.429	**.332**	**.469**	**.369**	.390	.050
bfi13	.000	−.018	.006	−.063	−.019	.020	.106	.081	.092	−.030	.013	−.203	**.721**	**.701**	**.962**	.795	−.001
Intra-construct																	
Mitems	.638	.619	.788	.627	.664	.631	.701	.655	.695	.628	.692	.522	.573	.672	.609		
Mcross_i	.011	.032	.037	−.006	.014	−.012	.022	.026	.047	.019	.013	.046	.024	.014	.042		

Note. Arg: Argentina. Spa: Spain. Mitems: mean of factor loadings within the hypothesized construct. Mcross_i: mean of cross-loadings from other non-hypothesized construct items. Msamples: mean of specific items factor loadings on hypothesized construct across the three countries. Mcross_s: mean of specific items factors cross-loadings across the three countries.

tailored for Spanish-speaking adults. The groups compared (university students from Argentina, Spain, and Peru) exhibit cultural differences but also share similarities in their roles and academic tasks (academic demands, academic stress, etc.). In this context, validity evidence was obtained from two studies focusing on content validity and internal structure.

Table 8
Factor distinctiveness: comparison of ESEM and CFA results.

	ESEM-CFA Congruence	Factor scores correlation (95 % CI)	
		ESEM-CFA	ESEM-RAW
Extraversion	0.95	0.98** (0.98, 0.98)	0.94** (0.93, 0.95)
Neuroticism	0.95	0.96** (0.96, 0.97)	0.97** (0.96, 0.97)
Agreeableness	0.97	0.99** (0.99, 0.99)	0.96** (0.96, 0.97)
Conscientiousness	0.95	0.96** (0.95, 0.96)	0.95** (0.94, 0.96)
Openness	0.96	0.98** (0.97, 0.98)	0.97** (0.97, 0.98)

Items	Item-level congruence
Extraversion	
bfi1	0.94
bfi4	0.94
bfi15	0.97
Neuroticism	
bfi2	0.92
bfi6	0.96
bfi10	0.97
Agreeableness	
bfi3	0.99
bfi12	1.00
bfi14	0.99
Conscientiousness	
bfi5	0.97
bfi8	1.00
bfi11	0.93
Openness	
bfi7	0.99
bfi9	0.81
bfi13	0.98

Note. ESEM: Exploratory Structural Equation Modeling. CFA: Confirmatory Factor Analysis. ESEM-CFA: correlation between ESEM and CFA factor scores. ESEM-RAW: correlation between ESEM factor scores and observed scores.

In Study 1, a quantitative assessment of item comprehension was conducted, leveraging the advantages of experiential judges drawn from the target population of the BFI-15p (Vogt et al., 2004). Study 1 explored a single aspect of the BFI-15p content: the perceived clarity of items (Ruel et al., 2016). Overall, the findings indicate that university students from all three countries perceive the items as clear and understandable. Interestingly, the Peruvian group consistently perceived slightly lower clarity in the BFI-15p items. However, their clarity levels were still above the chosen threshold ($V > 0.60$), suggesting that these participants generally judged the item clarity as appropriate.

A notable result was that item 1 (conversation behavior) was consistently rated as having moderate perceived clarity. Upon further review, the authors concluded that this item might carry a socially negative interpretation: the word *hablador* (talkative/verbose) differs from *conversador* (conversationalist), as the former can imply excessive verbalizing, potentially with insignificant or uninteresting content. Consequently, this item was revised and incorporated into Study 2 with the new term (*conversador*). The conclusion of Study 1 indicates that the items are sufficiently understandable for students across the three countries.

These results corroborate the pan-Hispanic nature of the items, as initially proposed in the standard Spanish study (Benet-Martínez & John, 1998). Furthermore, evaluating differences in perceived clarity highlights the sensitivity of item wording to language use in a comparative, cross-cultural context. This aspect should be further investigated in future cross-cultural studies with the BFI-15p.

In a comparative framework, Study 2 assessed the statistical characteristics of the items and the internal structure. Using a global test of distributional similarity that jointly focused on location and scale, a high degree of similarity was found among students from the three countries in their response distributions for the items. This finding is intriguing and remarkable, given the cultural differences among the three countries, as the response trends for the items were practically identical. This suggests two things: at the item level, the behaviors measured may represent a common trend among the groups, with minor differences that may not be significant in everyday life. At the scale score level, slight differences in scores across the three countries might be observed.

Regarding evaluating the internal structure (Study 2), the statistical differences between CFA and ESEM modeling provided valuable insights about the BFI-15p. Initially, CFA can serve as a standard starting point for internal structure analysis (Booth & Hughes, 2014), given its parsimonious representation of constructs and its restrictive characteristics, particularly regarding the absence of cross-loading modeling. For example, CFA results indicated that items showed adequate item-construct relationships, but ESEM provided a more precise distinction of the items. The target loadings (i.e., the loadings of each item on its expected factor) were substantially higher than non-target loadings (or cross-loadings). This finding also provided evidence of discriminant validity among the constructs comprising the BFI-15p, as cross-loadings were low and minimal shared variance with other constructs. The factor loadings obtained through ESEM may approximate their population values (Booth & Hughes, 2014; Tóth-Király et al., 2017) because spurious variability from other factors was removed from the expected factor.

An interesting aspect was the presence of heterogeneous factor loadings within the same factor. However, factor loadings within a factor can vary in magnitude (Logan et al., 2022), particularly when the constructs measured have broad content coverage, as these differences realistically reflect the variability of behaviors within the construct. Lastly, inter-factor correlations were smaller in the ESEM results, which is expected because the absence of cross-loading modeling in CFA artificially inflates correlations in that approach (Booth & Hughes, 2014; Tóth-Király et al., 2017). The magnitude of these correlations, particularly their positive or negative direction, aligns with the literature on the associations among the OCEAN factors (Balgiu, 2018; Benet-Martínez & John, 1998; Steyn & Ndofirepi, 2022).

The statistical fit of the ESEM model was superior to that of the CFA model, indicating that cross-loadings are relevant and retain variance from factors not intended to be measured (Alamer & Marsh, 2022). This finding helps better understand the constructs measured by the BFI-15p, and, given the context of measuring personality attributes, this result can be expected. Thus, although ESEM demonstrates a superior fit compared to CFA, the profile analysis between the two was not significantly different. The congruence indices (at both the factor and item levels) and the inter-factor correlations between the two models (CFA and ESEM) were high ($r > 0.90$), as were their correlations with the observed scores. This observation suggests that, in practical terms, the impact of cross-loadings is insignificant in this study. However, it remains necessary to corroborate this assumption through structural analysis using both types of modeling, as cross-loadings may be significant, for example, >0.30 (Dominguez-Lara & Merino-Soto, 2019). In conclusion, regarding the comparison between CFA and ESEM, the OCEAN factors exhibit evidence of discriminant validity, both empirically and theoretically, when measured using the BFI-15p. However, ESEM provided more consistent evidence of discriminant validity, albeit not markedly different from CFA.

On the other hand, factor scores and observed scores can be used to represent the measurement of a construct in subsequent analyses. The

Table 9
Factor latent correlation: CFA and ESEM models (Study 2).

	CFA					ESEM					Difference[a]				
	EXTR	NEUR	AGRE	CONS	OPEN	EXTR	NEUR	AGRE	CONS	OPEN	EXTR	NEUR	AGRE	CONS	OPEN
Argentina															
EXTR	1.000					1					–				
NEUR	−0.402	1.000				−0.348	1				0.063	–			
AGRE	0.415	−0.167	1.000			0.340	−0.153	1			0.088	0.014	–		
CONS	0.264	−0.272	0.325	1.000		0.154	−0.320	0.155	1		0.115	0.053	0.181	–	
OPEN	0.223	−0.013	0.277	0.043	1.000	0.064	−0.005	0.211	0.054	1	0.163	−0.008	0.070	−0.011	–
Spain															
EXTR	1.000					1					–				
NEUR	−0.170	1.000				−0.127	1				0.044	–			
AGRE	0.641	0.163	1.000			0.577	−0.227	1			0.102	0.395	–		
CONS	0.260	0.157	0.331	1.000		0.178	−0.094	0.183	1		0.086	0.253	0.159	–	
OPEN	0.125	0.089	0.195	0.241	1.000	0.030	−0.051	0.212	0.170	1	0.096	0.140	0.018	0.074	–
Peru															
EXTR	1.000					1					–				
NEUR	−0.325	1.000				−0.320	1				0.006	–			
AGRE	0.647	−0.208	1.000			0.605	−0.243	1			0.069	0.037	–		
CONS	0.564	−0.299	0.678	1.000		0.313	−0.138	0.429	1		0.315	0.170	0.367	–	
OPEN	0.467	0.062	0.578	0.696	1.000	0.287	0.202	0.464	0.007	1	0.211	0.143	0.157	0.852	–

Note: CFA: Confirmatory factor analysis. ESEM: Exploratory structural equations modeling. EXTR: Extraversion. NEUR: Neuroticism. AGRE: Agreeableness. CONS: Conscientiousness. OPEN: Openness.
[a] Absolute difference based in Fisher transformation.

Table 10
DIF analysis within-country (sex groups) in Study 2 (Peru, Argentina, Spain).

	Peru					Argentina					Spain				
	M^2	*p*	SMD	ES_{SMD}	Level	M^2	*p*	SMD	ES_{SMD}	Level	M^2	*p*	SMD	ES_{SMD}	Level
EXTR															
bfi1	0.492	0.483	0.078	0.020	Neg	3.809	0.051	−0.252	−0.063	Mod	0.000	1.000	−0.001	0.000	Neg
bfi4	0.696	0.404	0.054	0.014	Neg	2.216	0.137	0.195	0.049	Neg	0.842	0.359	0.090	0.030	Neg
bfi15	2.180	0.140	−0.132	−0.033	Neg	0.337	0.562	0.057	0.014	Neg	0.920	0.337	−0.089	−0.022	Neg
NEUR															
bfi2	0.017	0.896	0.001	0.000	Neg	0.104	0.747	−0.047	−0.012	Neg	8.081	0.004	0.292	0.073	Mod
bfi6	5.906	0.015	−0.262	−0.066	Mod	3.249	0.071	−0.292	−0.073	Mod	5.219	0.022	−0.241	−0.060	Mod
bfi10	5.338	0.021	0.261	0.065	Mod	3.929	0.047	0.339	0.085	Mod	0.316	0.574	−0.050	−0.013	Neg
AGRE															
bfi3	0.439	0.508	−0.040	−0.010	Neg	0.077	0.781	−0.010	−0.003	Neg	2.077	0.150	0.127	0.032	Neg
bfi12	0.006	0.938	0.028	0.007	Neg	0.638	0.424	0.116	0.039	Neg	0.663	0.416	−0.061	−0.015	Neg
bfi14	0.561	0.454	0.012	0.003	Neg	1.415	0.234	−0.106	−0.026	Neg	0.374	0.541	−0.066	−0.016	Neg
CONS															
bfi5	3.255	0.071	−0.128	−0.032	Neg	1.054	0.305	−0.159	−0.040	Neg	0.867	0.352	−0.121	−0.030	Neg
bfi8	0.705	0.401	−0.065	−0.016	Neg	0.044	0.834	0.028	0.007	Neg	0.342	0.559	0.047	0.012	Neg
bfi11	4.429	0.035	0.193	0.048	Neg	0.574	0.449	0.132	0.033	Neg	0.107	0.744	0.074	0.018	Neg
OPEN															
bfi7	0.699	0.403	−0.108	−0.027	Neg	0.059	0.808	−0.030	−0.007	Neg	1.852	0.174	−0.115	−0.029	Neg
bfi9	0.277	0.599	0.063	0.021	Neg	0.387	0.534	−0.107	−0.027	Neg	4.855	0.028	0.235	0.059	Mod
bfi13	0.128	0.721	0.044	0.011	Neg	0.737	0.391	0.138	0.034	Neg	1.680	0.195	−0.120	−0.030	Neg

Note. DIF analysis: differential item functioning analysis. M^2: Mantel test. *p*: *p*-value for Mantel test (significance level, after Bonferroni adjustment: 0.003). SMD: standardized mean difference. ES_{SMD}: SMD/Maximum category value. Level: effect size category for ES_{SMD} (Neg: negligible; Mod: moderated). EXTR: Extraversion. NEUR: Neuroticism. AGRE: Agreeableness. CONS: Conscientiousness. OPEN: Openness.

results indicate that cross-loadings cannot be ignored in construct modeling but can seemingly be disregarded when calculating scores. This finding suggests that other factors minimally influence the BFI-15p items, and such influence can be ignored for applied purposes.

The study of equivalence between groups consisted of two stages. In the first, group homogeneity within each country was not assumed, and participants' sex was selected as a potential source of heterogeneity. It was found that the DIF of item 2 ("…is depressed, melancholic") was moderate and of greater magnitude in the male group, differing from findings in other studies with Spanish-speaking populations (Dominguez-Lara et al., 2024; Dominguez-Lara & Campos-Uscanga, 2024). There is no clear explanation for the DIF, where males responded differently from females, so it was ultimately disregarded.

In the final step, evidence supported measurement invariance of the OCEAN factors across country groups, a critical aspect of internal structure that allows for valid comparisons of latent parameters such as

Table 11
Measurement invariance of BFI-15p in the three countries (Study 2).

	WLSMV – χ^2 (df)	CFI	RMSEA (90 % CI)	WRMR
Argentina	71.340 (40) $p = .0017$	0.949	0.064 (0.039, 0.087)	0.445
Spain	72.007 (40) $p = .0014$	0.977	0.052 (0.032, 0.070)	0.405
Peru	94.669 (40) $p < .0001$	0.966	0.087 (0.065, 0.110)	0.484

Invariance levels	WLSMV – χ^2 (df)	CFI	RMSEA (90 % CI)	WRMR	Δ CFI	Δ RMSEA
Configurational	235.89 ** (120)	0.968	0.066 (0.053, 0.078)	0.767		
Weak	377.515 ** (220)	0.957	0.056 (0.047, 0.066)	1.30	−0.011	−0.01
Strong	486.065 ** (300)	0.949	0.053 (0.044, 0.061)	1.50	−0.008	−0.003

Note. WLSMV: weighted least square mean and variance adjusted. df: degrees of freedom. CFI: Comparative Fit Index. RMSEA: Root mean square error of approximation. CI: Confidence interval. WRMR: Weighted Root Mean Square Residual. Δ: difference between approximate fit index. **$p < .001$.

Table 12
Reliability and internal consistency estimates (Study 2).

	Extraversion (95 % CI)	Neuroticism (95 % CI)	Agreeableness (95 % CI)	Conscientiousness (95 % CI)	Openness (95 % CI)
Omega coefficient[a]					
Argentina	0.614	0.578	0.684	0.604	0.568
Spain	0.575	0.642	0.697	0.681	0.684
Peru	0.786	0.641	0.768	0.632	0.514
APD					
Argentina	0.205 (0.183, 0.228)	0.302 (0.278, 0.326)	0.144 (0.128, 0.161)	0.212 (0.190, 0.230)	0.229 (0.207, 0.252)
Spain	0.215 (0.195, 0.235)	0.270 (0.252, 0.288)	0.133 (0.120, 0.147)	0.206 (0.190, 0.224)	0.204 (0.188, 0.220)
Peru	0.151 (0.132, 0.172)	0.244 (0.220, 0.268)	0.126 (0.110, 0.12)	0.186 (0.165, 0.206)	0.198 (0.177, 0.219)

Note. CI: Confidence interval. APD: average proportional distance (very good/excellent: 0 to 0.20; acceptable: 0.20 to 0.25; >0.25: questionable).
[a] Coefficients obtained from the CFA solution.

means, covariances, and variances (Church, 2008; Eigenhuis et al., 2015). This finding can be partly attributed to the linguistic similarities among the countries analyzed—Argentina, Spain, and Peru. The observed equivalence also offers partial support for the Spanish version of the BFI developed by Benet-Martínez and John (1998). Moreover, the results suggest that the comparability of the constructs remains stable despite potential contextual differences among university systems—an important consideration given prior findings showing that academic settings can influence the measurement of constructs like academic burnout (Campos & Maroco, 2012).

The evidence of factorial invariance presented in this study is particularly relevant, as it confirms the uniform assessment of personality traits under the BFF model among Spanish-speaking populations from distinct cultural backgrounds. This finding contributes to the theoretical development of personality by supporting both its etic and emic dimensions (Fischer & Rudnev, 2024). Consequently, future cross-group comparisons using this instrument are less likely to yield misleading conclusions regarding cultural differences or similarities (Chen, 2008).

In the reliability assessment, although none of the point estimates for the ω coefficient exceeded 0.70, the confidence intervals were close to this value. The range of these intervals depends on the sample size, and a larger sample is required for precise estimates. For instance, a sample size of 1000 subjects would result in narrower intervals around the population value. Given each scale's broad theoretical content coverage and that each scale was measured with only three items, higher values (e.g., >0.70) could not be expected. These results appear aligned with the original study (Dominguez-Lara & Merino-Soto, 2018) since the confidence intervals generated in the present study tend to include their reported values. Finally, the magnitude of the point ω coefficients can be considered acceptable for short and ultra-short measures (Ponterotto & Charter, 2009), as their magnitude is not significantly affected by the number of items.

Using the APD estimator, a relatively recent measure of response consistency (Sturman et al., 2009), there was general agreement with ω (in terms of appropriate values), and the results suggest an acceptable degree of response consistency. However, although ω and APD are different and complementary measures, they do not need to be strictly concordant, as the similarity of APD was studied with the α coefficient, not with ω (Sturman et al., 2009). An implication of the reliability study is that, given the magnitudes obtained, the BFI-15p would be more precise for group descriptions and basic research purposes, as higher coefficients (>0.90) are required for individual-level interpretations.

A final contribution of Study 2 lies in its inclusion of an analysis of

potential insufficient effort or careless (IE/C) responses—an approach not commonly reported in the prior literature. This step enabled the exclusion of cases likely to reflect inattentive responding, which could compromise the internal structure of the BFI-15p (Magarotto Machado et al., 2024). Although this method does not guarantee a perfect true-positive detection rate when compared to more sophisticated techniques, even an approximate identification is preferable to omitting such checks altogether.

It is also reasonable to consider possible sources of variability related to academic programs or other forms of response bias. However, since the BFI-15p assesses stable intraindividual traits, it is unlikely that variability introduced by specific college courses would account for significant variance in the data. Moreover, several aspects support the robustness of the findings. First, the BFI-15p is less susceptible to method effects due to the absence of negatively worded items. Second, the sample consists predominantly of young adults with higher education levels, contrasting with populations characterized by lower education or older age—contexts where response variability has been more frequently associated with social desirability and acquiescence (Morales-Vives et al., 2017). Third, the measure used to detect possible IE/C responses likely captures at least some of the potential sources of self-report bias. Taken together, these considerations suggest that the observed results can serve as a reliable proxy for construct variance.

This first direct comparison of the BFI-15p structure across countries provides an optimistic perspective regarding its theoretical and practical implications. Among the theoretical implications, the BFF model, as approximated by the BFI-15p, maintains an etic position previously established in its initial model (i.e., the BFI). This means that, in the cross-cultural measurement of the BFF model, the OCEAN factors measured by the BFI-15p represent a parsimonious, equivalent, and similarly calibrated approach among students from different countries. Therefore, any differences in scores can be attributed to variations in construct levels rather than calibration differences between groups (Huang & Wong, 2014). These results suggest that the BFI-15p is an equivalent assessment tool for Argentina, Spain, and Peru university students. While cross-cultural validity does not inherently reveal differences between cultures, the instrument's validity is preserved across all cultures, one of the four types of possible cross-cultural studies (Matsumoto, 2003). This contribution to validity across the three Spanish-speaking countries establishes a baseline for comparative studies between these nations using a culturally valid measure. However, further evidence is required to confirm the equivalence of the BFI-15p's internal structure in Hispanic immigrants or Spanish-speaking individuals from other countries.

Among the practical implications, besides the aforementioned potential for academic and research studies, its application in clinical and psychotherapeutic settings could provide value for therapists working with patients across different Spanish-speaking regions. This idea is supported by the increase in online psychotherapy services since the COVID-19 pandemic (Castro et al., 2022; Open Access Government, 2022), as personality assessment often plays a key role in evaluation and treatment planning.

Despite its contributions, this study presents several significant limitations that warrant critical reflection. First, the use of non-probabilistic sampling methods, based on the researchers' access to participants, limits the sample's representativeness and restricts the generalizability of the findings. Although convergence was observed with previous studies on the BFI-15p in Spanish-speaking populations (Dominguez-Lara et al., 2022, 2024; Dominguez-Lara & Merino-Soto, 2019), the absence of random sampling increases the risk that the findings reflect institutional or contextual idiosyncrasies rather than general patterns across national or cultural groups.

Second, the exclusive use of psychology students may introduce discipline-specific biases, as students in this field are arguably more familiar with psychological terminology, constructs, and self-reflection practices. This familiarity could influence response styles (e.g., reduced social desirability bias or greater cognitive sophistication), which may not generalize to students in other academic domains or non-student populations. Thus, future studies should aim to replicate these findings across more diverse academic disciplines and populations.

Third, while the study examined potential sex differences within countries, other sources of sample heterogeneity—such as academic performance, socio-economic background, or urban versus rural origin—were not measured or controlled for. These unmeasured variables could act as confounding factors, potentially influencing how individuals interpret and respond to personality items. Given that previous literature has shown that variability in self-reports can be influenced by social desirability and acquiescence, particularly in populations with lower education levels or older age groups (Morales-Vives et al., 2017), the relative homogeneity of the current sample may mask important subgroup differences.

A further consideration concerns the use of the BFI-15p itself. As an ultra-brief instrument (with only three items per dimension), its clinical utility for individual-level assessment must be approached with caution. Establishing cut-off scores requires careful evaluation. However, based on the levels of reliability observed in this study, the BFI-15p may be a suitable screening tool for identifying groups with predominant personality patterns.

Finally, while the present study provides the first direct comparison of the BFI-15p's internal structure across Argentina, Spain, and Peru, the linguistic and cultural similarities among these countries may have contributed to the measurement invariance observed. Future research should extend this analysis to Spanish-speaking populations in more culturally diverse regions (e.g., Central America, the Caribbean, or Hispanic communities in the United States) to more comprehensively assess the instrument's cross-cultural validity.

To strengthen the scale's validity, future research should incorporate additional sources of evidence, such as correlations with theoretically relevant variables (Dominguez-Lara et al., 2024), and consider alternative methodological approaches like Item Response Theory. Moreover, given the longstanding critique that personality assessments based on the OCEAN model are largely validated in university populations (Rammstedt et al., 2013), it is advisable to gather psychometric evidence from more diverse, general-population samples.

Building on this study's contributions, future cross-cultural research among Argentina, Spain, and Peru can include cross-sectional and longitudinal studies examining personality and its associations with other constructs (Costa et al., 2019). The availability of an instrument with robust psychometric properties and demonstrated measurement invariance enables such investigations. Additionally, future studies could explore interaction models involving relevant variables within Spanish-speaking contexts and extend invariance testing to other countries in the region to broaden the evidence base.

5. Conclusion

The BFI-15p content, a brief Spanish-language measure of the Big Five Factors model, was perceived as clear and understandable across three undergraduate student samples from Spanish-speaking countries (Argentina, Spain, and Peru). The distributional characteristics of the items were highly similar among the three groups, as was the internal structure of the BFI-15p. The number of dimensions (five), factor loadings, intercepts, and residuals demonstrated invariance across the three groups. The latent correlations between the dimensions showed moderate similarity among specific scales. For research purposes involving group comparisons in Spanish-speaking higher education samples, the BFI-15p can be an efficient and valuable tool for exploring the Big Five Factors model.

CRediT authorship contribution statement

César Merino-Soto: Writing – review & editing, Writing – original

draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Sergio Dominguez-Lara:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Lucas Marcelo Rodriguez:** Writing – review & editing, Writing – original draft, Conceptualization. **Guillermo M. Chans:** Writing – review & editing, Visualization, Resources, Project administration, Funding acquisition. **Manuel Marti-Vilar:** Writing – review & editing, Validation, Supervision, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors extend their gratitude to the participants and institutions for their active involvement in completing the survey and their willingness to support the data collection process. We also acknowledge the financial support provided by the Writing Lab, Institute for the Future of Education, Tecnologico de Monterrey, Mexico, which was instrumental in the development of this work.

Data availability

Data will be made available on request.

References

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40*(4), 955–959. https://doi.org/10.1177/001316448004000419

Alamer, A., & Marsh, H. (2022). Exploratory structural equation modeling in second language research: An applied example using the dualistic model of passion. *Studies in Second Language Acquisition, 44*(5), 1477–1500. https://doi.org/10.1017/S0272263121000863

Allik, J. (2005). Personality dimensions across cultures. *Journal of Personality Disorders, 19*(3), 212–232. https://doi.org/10.1521/pedi.2005.19.3.212

Allik, J., & McCrae, R. R. (2004). Toward a geography of personality traits: Patterns of profiles across 36 cultures. *Journal of Cross-Cultural Psychology, 35*(1), 13–28. https://doi.org/10.1177/0022022103260382

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2014). Standards for educational and psychological testing. In *National Council on measurement in education*. American Educational Research Association.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397–438. https://doi.org/10.1080/10705510903008204

Balgiu, B. A. (2018). The psychometric properties of the Big Five inventory-10 (BFI-10) including correlations with subjective and psychological well-being. *Global Journal of Psychology Research: New Trends and Issues, 8*(2), 61–69. https://doi.org/10.18844/gjpr.v8i2.3434

Benet-Martínez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology, 75*(3), 729–750. https://doi.org/10.1037/0022-3514.75.3.729

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein, & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). Jossey-Bass/Wiley.

Bliese, P. D. (2016). Multilevel: Multilevel Functions. *R package version, 2*, 6. https://CRAN.R-project.org/package=multilevel.

Booth, T., & Hughes, D. J. (2014). Exploratory structural equation modeling of personality data. *Assessment, 21*(3), 260–271. https://doi.org/10.1177/1073191114528029

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–192). Sage.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods, 2*(1), 49–68. https://doi.org/10.1177/109442819921004

Campos, J. A., & Maroco, J. (2012). Maslach burnout inventory - Student survey: Portugal-Brazil cross-cultural adaptation. *Revista de Saúde Pública, 46*(5), 816–824. https://doi.org/10.1590/s0034-89102012000500008

Casari, L. M., Ison, M. S., & Gómez, B. M. M. (2019). Personal style of the therapist and personality dimensions in a sample of Argentinian therapists. *Research in Psychotherapy: Psychopathology, Process and Outcome, 22*(2), 292–307. https://doi.org/10.4081/ripppo.2019.362

Castro Solano, A., & Casullo, M. M. (2001). Rasgos de personalidad, bienestar psicológico y rendimiento académico en adolescentes argentinos. [Personality traits, psychological well-being, and academic achievement of Argentine adolescents.]. *Interdisciplinaria Revista de Psicología y Ciencias Afines, 18*(1), 65–85. https://psycnet.apa.org/record/2001-11537-002.

Castro Solano, A., Lupano Perugini, M. L., & Caporiccio Trillo, M. A. (2022). Motivos de Uso de Instagram. Análisis Automático del Texto de los Usuarios y Asociación con los Rasgos de Personalidad Normal, Psicopatológica y Positiva. *Perspectivas en Psicología, 19*(2), 171–193. http://perspectivas.mdp.edu.ar/revista/index.php/pep/article/view/629.

Castro, T. G. D., Guerin, K. S. D. A., & Pimentel, L. I. R. (2022). Psicoterapia on-line no Brasil: revisão integrativa de publicações nacionais pré-pandemia por COVID-19. *Contextos Clínicos, 15*(2), 611–637. https://doi.org/10.4013/ctc.2022.152.13

Castro-Solano, A. (2005). *Técnicas de Evaluación psicológica en los ámbitos militares*. Paidós.

Chavez Irigoyen, C. (2023). *Género y Educación*. SUNEDU: Una mirada al sistema universitario desde las trayectorias estudiantiles.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005–1018. https://doi.org/10.1037/a0013193

Chiorri, C., Marsh, H. W., Ubbiali, A., & Donati, D. (2016). Testing the factor structure and measurement invariance across gender of the big five inventory through exploratory structural equation modeling. *Journal of Personality Assessment, 98*(1), 88–99. https://doi.org/10.1080/00223891.2015.1035381

Church, A. T. (2008). Current controversies in the study of personality across cultures. *Social and Personality Psychology Compass, 2*(5), 1930–1951. https://doi.org/10.1111/j.1751-9004.2008.00132.x

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual* (Psychological Assessment Resources).

Costa, P. T., McCrae, R. R., & Löckenhoff, C. E. (2019). Personality across the life span. *Annual Review of Psychology, 70*, 423–448. https://doi.org/10.1146/annurev-psych-010418-103244

Costa, P. X., Sr., & McCrae, R. R. (1985). *NEO Personality Inventory (NEO-PI)*. https://doi.org/10.1037/t07564-000

Courtois, R., Petot, J. M., Plaisant, O., Allibe, B., Lignier, B., Réveillère, C., … John, O. (2020). Validation française du Big Five Inventory à 10 items (BFI-10). *L'Encéphale, 46*(6), 455–462. https://doi.org/10.1016/j.encep.2020.02.006

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior, 36*(2), 171–181. https://doi.org/10.1002/job.1962

Digman, J. M., & Inouye, J. (1986). Further specification of the five robust factors of personality. *Journal of Personality and Social Psychology, 50*(1), 116–123. https://doi.org/10.1037/0022-3514.50.1.116

Digman, J. M., & Takemoto-Chock, N. K. (1981). Factors in the natural language of personality: Re-analysis, comparison, and interpretation of six major studies. *Multivariate Behavioral Research, 16*(2), 149–170. https://doi.org/10.1207/s15327906mbr1602_2

DiStefano, C., Jin, L., Ning, J., & Shi, D. (2018). Examination of the weighted root mean square residual: Evidence for trustworthiness? *Structural Equation Modeling: A Multidisciplinary Journal, 25*(3), 453–466. https://doi.org/10.1080/10705511.2017.1390394

Dominguez-Lara, S., & Campos-Uscanga, Y. (2024). Structural analysis and measurement invariance regarding the sex of three brief versions of the Big Five Inventory in Mexican general population. *Testing, Psychometrics, Methodology in Applied Psychology, 31*(2), 207–226. https://doi.org/10.4473/TPM31.2.5

Dominguez-Lara, S., Campos-Uscanga, Y., & Valente, S. N. (2022). Análisis psicométrico de versiones cortas del Big Five Inventory en universitarios mexicanos. *Avaliação Psicológica, 21*, 140–149. http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712022000200003&nrm=iso.

Dominguez-Lara, S., Carmona-Halty, M., Schulmeyer, M. K., Valente, S. N., & Lourenço, A. A. (2024). Análisis psicométrico de dos versiones del Big Five Inventory-15 en estudiantes universitarios chilenos: estructura interna, invarianza de medición y asociación con bienestar subjetivo. *Ciencias Psicológicas, 18*(1), Article e-3369. https://doi.org/10.22235/cp.v18i1.3369

Dominguez-Lara, S., & Merino-Soto, C. (2018). Dos versiones breves del Big Five Inventory en universitarios peruanos: BFI-15p y BFI-10p. *Liberabit, 24*(1), 81–96. https://doi.org/10.24265/liberabit.2018.v24n1.06

Dominguez-Lara, S., & Merino-Soto, C. (2019). Estructura interna del BFI-10P y BFI-15P: un estudio complementario con enfoque CFA y ESEM. *Revista Argentina de Ciencias del Comportamiento, 10*(3), 22–34. https://doi.org/10.32348/1852.4206.v10.n3.21037

Dominguez-Lara, S., Merino-Soto, C., Zamudio, B., & Guevara-Cordero, C. (2018). Big Five Inventory en Universitarios Peruanos: Resultados Preliminares de su Validación. *Psykhe (Santiago), 27*(2), 1–12. https://doi.org/10.7764/psykhe.27.2.105

Dominguez-Lara, S. A., Prada-Chapoñan, R., & Moreta-Herrera, R. (2019). Diferencias de género en la influencia de la personalidad sobre la procrastinación académica en

estudiantes universitarios peruanos. *Acta Colombiana de Psicología, 22*(2), 125–136. https://doi.org/10.14718/ACP.2019.22.2.7

Dorans, N. J., & Schmitt, A. P. (1991). Constructed response and differential item functioning: A pragmatic approach. *ETS Research Report Series, 1991*(2), i–49. https://doi.org/10.1002/j.2333-8504.1991.tb01414.x

Dorans, N. J., & Schmitt, A. P. (1993). Constructed response and differential item functioning: A pragmatic approach. In R. E. Bennett, & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 135–165). Erlbaum.

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for r(WG) and average deviation interrater agreement indexes. *Journal of Applied Psychology, 88*(2), 356–362. https://doi.org/10.1037/0021-9010.88.2.356

Eidman, L., Díaz, J. P., & Landressina, A. (2019). Desarrollo, Bienestar Psicológico, Rasgos de Personalidad y Factores Socio-Demográficos en Población Clínica y General del Nordeste Argentino. *Calidad de Vida y Salud, 12*(1). http://revistacdvs.uflo.edu.ar/index.php/CdVUFLO/article/view/177.

Eigenhuis, A., Kamphuis, J. H., & Noordhof, A. (2015). Personality differences between the United States and the Netherlands: The influence of violations of measurement invariance. *Journal of Cross-Cultural Psychology, 46*(4), 549–564. https://doi.org/10.1177/0022022115570671

Fernández, I., Torres, Z., Martínez-Gregorio, S., Oliver, A., & Tomás, J. M. (2023). Method effects associated to item valence: Evidence from the 10-item Big-Five Inventory in older adults. *Research on Aging, 45*(7–8), 517–525. https://doi.org/10.1177/01640275221132196

Fernández, R. S., Crivelli, L., Guimet, N. M., Allegri, R. F., Picco, S., & Pedreira, M. E. (2022). Psychological distress and mental health trajectories during the COVID-19 pandemic in Argentina: A longitudinal study. *Scientific Reports, 12*(1), 5632. https://doi.org/10.1038/s41598-022-09663-2

Fetvadjiev, V., & Van der Vijver, F. J. R. (2015). Universality of the five-factor model of personality. In J. D. Wright (Ed.) (2nd ed.,, *Vol. 9. International encyclopedia of social and behavioral sciences* (pp. 249–253). Elsevier.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage.

Fischer, R., & Rudnev, M. (2024). From MIsgivings to MIse-en-scène: The role of invariance in personality science. *European Journal of Personality, 0*(0). https://doi.org/10.1177/08902070241283081

Galinha, I. C., Garcia-Martin, M.Á., Oishi, S., Wirtz, D., & Esteves, F. (2016). Cross-cultural comparison of personality traits, attachment security, and satisfaction with relationships as predictors of subjective well-being in India, Sweden, and the United States. *Journal of Cross-Cultural Psychology, 47*(8), 1033–1052. https://doi.org/10.1177/0022022116658262

García-Flores, L. A., Campos-Ugaz, O., Hernández, R. M., Saavedra-López, M. A., Aguinaga Doig, S. G., Bravo Larrea, Y. M., … Campos-Ugaz, W. A. (2022). Personality and attitudes towards machismo as predictors of sexual harassment in university students in northern Peru. *Journal of Educational and Social Research, 12*(5), 54–62. https://doi.org/10.36941/jesr-2022-0122

Genise, G., Ungaretti, J., & Etchezahar, E. (2020). El Inventario de los Cinco Grandes Factores de Personalidad en el contexto argentino: puesta a prueba de los factores de orden superior. *Diversitas, 16*(2), 325–340. https://doi.org/10.15332/22563067.6298

Gerlitz, J. Y., & Schupp, J. (2005). Zur Erhebung der Big-Five-basierten ersoenlichkeitsmerkmale im SOEP. https://www.diw.de/documents/publikationen/73/diw_01.c.43490.de/rn4.pdf.

Gibson, A. M., & Bowling, N. A. (2019). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment, 36*(2), 410–420. https://doi.org/10.1027/1015-5759/a000526

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Guido, G., Peluso, A. M., Capestro, M., & Miglietta, M. (2015). An Italian version of the 10-item Big Five Inventory: An application to hedonic and utilitarian shopping values. *Personality and Individual Differences, 76*, 135–140. https://doi.org/10.1016/j.paid.2014.11.053

Gurven, M., von Rueden, C., Massenkoff, M., Kaplan, H., & Lero Vie, M. (2013). How universal is the Big Five? Testing the five-factor model of personality variation among forager–farmers in the Bolivian Amazon. *Journal of Personality and Social Psychology, 104*(2), 354–370. https://doi.org/10.1037/a0030841

Hahn, E., Gottschling, J., & Spinath, F. M. (2012). Short measurements of personality – Validity and reliability of the GSOEP Big Five Inventory (BFI-S). *Journal of Research in Personality, 46*(3), 355–359. https://doi.org/10.1016/j.jrp.2012.03.008

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. discussion 83–135. https://doi.org/10.1017/s0140525x0999152x

Herrera Torres, L., & Perandones González, T. M. (2021). Fortalezas y virtudes personales, felicidad subjetiva, sentido del humor y personalidad. Análisis según el país y género de los docentes. Revista INFAD de Psicología. *International Journal of Developmental and Educational Psychology, 1*(2), 99–108. https://doi.org/10.17060/ijodaep.2021.n2.v1.2161

Hothorn, T., Hornik, K., van de Wiel, M. A., & Zeileis, A. (2008). Implementing a class of permutation tests: The coin package. *Journal of Statistical Software, 28*(8), 1–23. https://doi.org/10.18637/jss.v028.i08

Huang, W. Y., & Wong, S. H. (2014). Cross-cultural validation. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 1369–1371). Netherlands: Springer. https://doi.org/10.1007/978-94-007-0753-5_630.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). Big five inventory. *Journal of Personality and Social Psychology.*. https://doi.org/10.1037/t07550-000

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*(1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Kim, S.-H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*(2), 93–116. https://doi.org/10.1111/j.1745-3984.2007.00029.x

Kim, S. Y., Kim, J. M., Yoo, J. A., K.Y., B, Kim, S. W., Yang, S. J., … Yoon, J. S. (2010). Standardization and validation of Big Five Inventory – Korean version (BFI-K) in elders. *Korean Journal of Biological Psychiatry, 17*(1), 15–25.

Kline, R. B. (2013). Exploratory and confirmatory factor analysis. In *Applied quantitative analysis in the social sciences* (pp. 171–207). Routledge. https://doi.org/10.4324/9780203108550.

Kunnel John, R., Xavier, B., Waldmeier, A., Meyer, A., & Gaab, J. (2019). Psychometric evaluation of the BFI-10 and the NEO-FFI-3 in Indian adolescents. *Frontiers in Psychology, 10*, 1057. https://doi.org/10.3389/fpsyg.2019.01057

Lance, C. E., Dawson, B., Birkelbach, D., & Hoffman, B. J. (2010). Method effects, measurement error, and substantive conclusions. *Organizational Research Methods, 13*(3), 435–455. https://doi.org/10.1177/1094428109352528

de-la-Peña, C., & Luque-Rojas, M. J. (2021). Levels of reading comprehension in higher education: Systematic review and meta-analysis. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.712901

Larsen, R. J., & Buss, D. M. (2002). *Personality psychology: Domains of knowledge about human nature*. McGraw-Hill.

Lippold, J. V., Laske, J. I., Hogeterp, S. A., Duke, É., Grünhage, T., & Reuter, M. (2020). The role of personality, political attitudes and socio-demographic characteristics in explaining individual differences in fear of coronavirus: A comparison over time and across countries. *Frontiers in Psychology, 11*, Article 552305. https://doi.org/10.3389/fpsyg.2020.552305

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). El Análisis Factorial Exploratorio de los Ítems: una guía práctica, revisada y actualizada. *Anales de Psicología, 30*(3), 1151–1169. https://doi.org/10.6018/analesps.30.3.199361

Logan, J. A. R., Jiang, H., Helsabeck, N., & Yeomans-Maldonado, G. (2022). Should I allow my confirmatory factors to correlate during factor score extraction? Implications for the applied researcher. *Quality & Quantity, 56*(4), 2107–2131. https://doi.org/10.1007/s11135-021-01202-x

Lovik, A., Verbeke, G., & Molenberghs, G. (2017). Evaluation of a very short test to measure the Big Five personality factors on a Flemish sample. *Journal of Psychological and Educational Research, 25*(2), 7–17.

Ludeke, S. G., & Larsen, E. G. (2017). Problems with the Big Five assessment in the world values survey. *Personality and Individual Differences, 112*, 103–105. https://doi.org/10.1016/j.paid.2017.02.042

Magarotto Machado, G., Hauck-Filho, N., Pallini, A. C., Dias-Viana, J. L., Santana, L. H. B. C., da Silva, C. A. N. M., & Valentini, F. (2024). Investigating the acquiescent responding impact in empathy measures. *International Journal of Testing, 24*(4), 277–301. https://doi.org/10.1080/15305058.2024.2364170

Manrique-Millones, D., & Castillo-Blanco, R. (2021). Big Five personality inventory: Sex differences in a Peruvian sample. *Mankind Quarterly, 61*(3), 542–544. https://doi.org/10.46469/mq.2021.61.3.9

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*(303), 690–700. https://doi.org/10.2307/2282717

Marozzi, M. (2014). The multisample Cucconi test. *Statistical Methods & Applications, 23*(2), 209–227. https://doi.org/10.1007/s10260-014-0255-x

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22*(3), 471–491. https://doi.org/10.1037/a0019227

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—Camouflaging misspecification with item parcels in CFA models. *Psychological Methods, 18*(3), 257–284. https://doi.org/10.1037/a0032773

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*(10), 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2013). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology, 49*(6), 1194–1218. https://doi.org/10.1037/a0026913

Matsumoto, D. (2003). Cross-cultural research. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology* (pp. 189–208). Blackwell Publishing Ltd.. https://doi.org/10.1002/9780470756973.ch9

McCrae, R. R., & Costa, P. T., Jr. (1989). More reasons to adopt the five-factor model. *American Psychologist, 44*(2), 451–452. https://doi.org/10.1037/0003-066X.44.2.451

McCrae, R. R., & Costa, P. T. (1985). Comparison of EPI and psychoticism scales with measures of the five-factor model of personality. *Personality and Individual Differences, 6*(5), 587–597. https://doi.org/10.1016/0191-8869(85)90008-X

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*(1), 81–90. https://doi.org/10.1037/0022-3514.52.1.81

McCrae, R. R., Gaines, J. F., & Wellington, M. A. (2013). The five-factor model in fact and fiction. In I. B. Weiner, H. A. Tennen, & J. M. Suls (Eds.), Handbook of psychology: Personality and social psychology (2nd ed., Vol. 5, pp. 6–91). John Wiley & Sons, Inc.

McCrae, R. R., & Terracciano, A. (2005). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology, 89*(3), 407–425. https://doi.org/10.1037/0022-3514.89.3.407

McDonald, R. P., & Ho, M. H. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64–82. https://doi.org/10.1037/1082-989x.7.1.64

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. https://doi.org/10.1037/a0028085

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods, 9*(3), 369–403. https://doi.org/10.1177/1094428105283384

Merino-Soto, C., & Angulo-Ramos, M. (2022). Estudios métricos del Compliance Questionnaire on Rheumatology (CQR): ¿Un caso de inducción de la validez? *Reumatología Clínica, 18*(8), 497–498. https://doi.org/10.1016/j.reuma.2021.03.004

Merino-Soto, C., & Calderón-De la Cruz, G. A. (2018). Validez de estudios peruanos sobre estrés y burnout. *Revista Peruana de Medicina Experimental y Salud Publica, 35*(2), 353–354. https://doi.org/10.17843/rpmesp.2018.352.3521

Merino-Soto, C., & Livia-Segovia, J. (2009). Intervalos de confianza asimétricos para el índice la validez de contenido: Un programa Visual Basic para la V de Aiken. *Anales de Psicología, 25*(1), 169–171. https://revistas.um.es/analesps/article/view/71631.

Micillo, L., Rioux, P.-A., Mendoza, E., Kübel, S. L., Cellini, N., Van Wassenhove, V., … Mioni, G. (2022). Time perspective predicts levels of anxiety and depression during the COVID-19 outbreak: A cross-cultural study. *PLoS One, 17*(9), Article e0269396. https://doi.org/10.1371/journal.pone.0269396

Molero-Jurado, M.d. M., Pérez-Fuentes, M.d. C., Soriano, J. G., Tortosa, B. M., Oropesa, N. F., Simón-Márquez, M.d. M., & Gázquez-Linares, J. J. (2020). Personality and job creativity in relation to engagement in nursing. *Anales de Psicología, 36*(3), 533–541. https://doi.org/10.6018/analesps.36.3.405391

Morales-Vives, F., Lorenzo-Seva, U., & Vigil-Colet, A. (2017). How response biases affect the factor structure of Big Five personality questionnaires. *Anales de Psicología, 33*, 589–596. http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-97282017000300016&nrm=iso.

Open Access Government. (2022, October 17). 21.5% more people accessing talking therapies through the NHS. https://www.openaccessgovernment.org/21-5-more-people-accessing-talking-therapies-through-the-nhs/145977/.

Padilla, M. A., & Divers, J. (2013). Bootstrap interval estimation of reliability via coefficient omega. *Journal of Modern Applied Statistical Methods, 12*(1), 78–89. https://doi.org/10.22237/jmasm/1367381520

Palsson, T. S., Skou, S. T., Pape, M. H., Hirata, R. P., Rafn, T., Bellosta-López, P., & Christensen, S. W. M. (2023). Cross-cultural translation and adaptation of the Danish version of the brief version of the 10-item Big Five Inventory. *Physiotherapy Research International, 28*(3), Article e2004. https://doi.org/10.1002/pri.2004

Pejić, M., Tenjović, L., & Knežević, G. (2014). Validation of the BFI-10 questionnaire - Short version of the Big Five Inventory. *Primenjena psihologija, 7*(1), 45–92. https://doi.org/10.19090/pp.2014.1.45-92

Pendergast, L. L., von der Embse, N., Kilgus, S. P., & Eklund, K. R. (2017). Measurement equivalence: A non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. *Journal of School Psychology, 60*, 65–82. https://doi.org/10.1016/j.jsp.2016.11.002

Penfield, R. D., & Giacobbi, J. P. R. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science, 8*(4), 213–225. https://doi.org/10.1207/s15327841mpee0804_3

Ponterotto, J. G., & Charter, R. A. (2009). Statistical extensions of Ponterotto and Ruckdeschel's (2007) reliability matrix for estimating the adequacy of internal consistency coefficients. *Perceptual and Motor Skills, 108*(3), 878–886. https://doi.org/10.2466/pms.108.3.878-886

Prada-Chapoñan, R., Navarro-Loli, J. S., & Dominguez-Lara, S. (2020). Personalidad y agotamiento emocional académico en estudiantes universitarios peruanos: un estudio predictivo. *Revista Digital de Investigación en Docencia Universitaria, 14*(2), Article e1227. https://doi.org/10.19083/ridu.2020.1227

Rammstedt, B. (2007). The 10-item Big Five Inventory: Norm values and investigation of sociodemographic effects based on a German population representative sample. *European Journal of Psychological Assessment, 23*(3), 193–201. https://doi.org/10.1027/1015-5759.23.3.193

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Rammstedt, B., Kemper, C. J., & Borg, I. (2013). Correcting Big Five personality measurements for acquiescence: An 18–country cross–cultural study. *European Journal of Personality, 27*(1), 71–81. https://doi.org/10.1002/per.1894

Rammstedt, B., Roemer, L., & Lechner, C. M. (2024). Consistency of the structural properties of the BFI-10 across 16 samples from eight large-scale surveys in Germany. *European Journal of Psychological Assessment, 40*(3), 204–215. https://doi.org/10.1027/1015-5759/a000765

Revelle, W. (2023). psych: Procedures for psychological, psychometric, and personality research. *R package version, 2*(2), 9. https://CRAN.R-project.org/package=psych.pac.

Reyes Zamorano, E., Álvarez Carrillo, C., Peredo Silva, A., Miranda Sandoval, A., & Rebolledo Pastrana, I. M. (2014). Psychometric properties of the big five inventory in a Mexican sample. *Salud Mental, 37*(6), 491–497. https://doi.org/10.17711/sm.0185-3325.2014.059

Ruel, E., Wagner, W. E., III, & Gillespie, B. J. (2016). Pretesting and pilot testing. In *The practice of survey research: Theory and applications* (pp. 101–119). SAGE Publications, Inc.. https://doi.org/10.4135/9781483391700

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology, 38*(2), 173–212. https://doi.org/10.1177/0022022106297299

Schultz, D. P., & Schultz, S. E. (2016). *Theories of personality* (10th ed.). Cengage Learning.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., … Sprangers, M. A. G. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology, 62*(3), 288–295. https://doi.org/10.1016/j.jclinepi.2008.06.003

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and Psychological Measurement, 79*(2), 310–334. https://doi.org/10.1177/0013164418783530

Shi, D., & Maydeu-Olivares, A. (2019). The effect of estimation methods on SEM fit indices. *Educational and Psychological Measurement, 80*(3), 421–445. https://doi.org/10.1177/0013164419885164

Singer, S., Engelberg, P. M., Weißflog, G., Kuhnt, S., & Ernst, J. (2013). Construct validity of the EORTC quality of life questionnaire information module. *Quality of Life Research, 22*(1), 123–129. https://doi.org/10.1007/s11136-012-0114-x

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12*(1), 102–111. https://doi.org/10.1037//1040-3590.12.1.102

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of big five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology, 94*(4), 718–737. https://doi.org/10.1037/0022-3514.94.4.718

Southwell, M. L. (2024). Trayectorias académicas y desigualdades de género en la educación superior universitaria en Argentina: Una mirada al campo de las ingenierías. *Políticas Educativas – PolEd, 17*(1). https://seer.ufrgs.br/index.php/Poled/article/view/137895.

Steyn, R., & Ndofirepi, T. M. (2022). Structural validity and measurement invariance of the short version of the Big Five Inventory (BFI-10) in selected countries. *Cogent Psychology, 9*(1), Article 2095035. https://doi.org/10.1080/23311908.2022.2095035

Stone, B. M., Bartholomay, E. M., & Chamberlain, A. B. (2022). Validation of the BFI-10-R: A new BFI scale with strong structural and construct validity. *Journal of Psychopathology and Behavioral Assessment, 44*(4), 1064–1076. https://doi.org/10.1007/s10862-022-09978-4

Stover, J. B., Solano, A. C., & Liporace, M. F. (2019). Dysfunctional personality traits: Relationship with Five Factor Model, adaptation and symptomatology in a community sample from Buenos Aires. *Research in Psychotherapy: Psychopathology, Process and Outcome, 22*(2), 343. https://doi.org/10.4081/ripppo.2019.343

Sturman, E. D., Cribbie, R. A., & Flett, G. L. (2009). The average distance between item values: A novel approach for estimating internal consistency. *Journal of Psychoeducational Assessment, 27*(5), 409–420. https://doi.org/10.1177/0734282908330937

Suárez, C., & Marrero, R. J. (2020). Intervención positiva multicomponente a través del correo electrónico para promover el bienestar en estudiantes universitarios. *Behavioral Psychology, 28*(3), 375–392. https://www.behavioralpsycho.com/wp-content/uploads/2020/12/01.Suarez_28-3.pdf.

Tomczak, M., & Tomczak-Łukaszewska, E. (2014). The need to report effect size estimates revisited. *An overview of some recommended measures of effect size., 1*(21), 19–25.

Tóth-Király, I., Bõthe, B., Rigó, A., & Orosz, G. (2017). An illustration of the Exploratory Structural Equation Modeling (ESEM) framework on the passion scale. *Frontiers in Psychology, 8*, 1968. https://doi.org/10.3389/fpsyg.2017.01968

Varchetta, M., González-Sala, F., Mari, E., Quaglieri, A., Fraschetti, A., Cricenti, C., … Martí-Vilar, M. (2023). Psychosocial risk factors of technological addictions in a sample of Spanish University students: The influence of Emotional (Dys)Regulation, personality traits and Fear of Missing Out on internet addiction. *Psychiatry Research, 329*, Article 115518. https://doi.org/10.1016/j.psychres.2023.115518

Verhelst, N. D. (2012). Profile analysis: A closer look at the PISA 2000 reading data. *Scandinavian Journal of Educational Research, 56*(3), 315–332. https://doi.org/10.1080/00313831.2011.583937

Vogt, D. S., King, D. W., & King, L. A. (2004). Focus groups in psychological assessment: Enhancing content validity by consulting members of the target population. *Psychological Assessment, 16*(3), 231–243. https://doi.org/10.1037/1040-3590.16.3.231

Weiner, I. B., Tennen, H. A., & Suls, J. M. (2013). *Handbook of psychology, personality and social psychology* (2nd ed.) (2nd ed., Vol. 5). Wiley https://elibro.net/es/ereader/sibuca/182626?page=89.

Weiss, S., & Roberts, R. D. (2018). Using anchoring vignettes to adjust self-reported personality: A comparison between countries. *Frontiers in Psychology, 9*, 325. https://doi.org/10.3389/fpsyg.2018.00325

Wells, C. S. (2021). Assessing measurement invariance for applied research. *Cambridge University Press.*. https://doi.org/10.1017/9781108750561

Wells, C. S. (2023). MeasInv: Collection of methods to detect dichotomous and polytomous Differential Item Functioning (DIF). R package version 0.1.*0.* https://github.com/cswells1/MeasInv/.

Willyard, C. (2011, January 11). Men: A growing minority. https://www.apa.org/gradpsych/2011/01/cover-men.

Yamagata, S., Suzuki, A., Ando, J., Ono, Y., Kijima, N., Yoshimura, K., … Jang, K. L. (2006). Is the genetic structure of human personality universal? A cross-cultural twin study from North America, Europe, and Asia. *Journal of Personality and Social Psychology, 90*(6), 987–998. https://doi.org/10.1037/0022-3514.90.6.987

Yáñez Botello, C. R. (2016). Caracterización de los procesos cognoscitivos y competencias involucrados en los niveles de comprensión lectora en Estudiantes Universitarios. *Cuadernos Hispanoamericanos de Psicología, 13*(2), 75–90. https://doi.org/10.18270/chps.v13i2.1350

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233–251. https://doi.org/10.1111/j.1745-3984.1993.tb00425.x