

### Adapting the BFI-2 Around the World - Coordinated Translation and Validation in Five Languages and Cultural Contexts

Rammstedt, Beatrice; Roemer, Lena; Lechner, Clemens; Soto, Christopher J.

Veröffentlichungsversion / Published Version

Zeitschriftenartikel / journal article

#### Empfohlene Zitierung / Suggested Citation:

Rammstedt, B., Roemer, L., Lechner, C., & Soto, C. J. (2024). Adapting the BFI-2 Around the World - Coordinated Translation and Validation in Five Languages and Cultural Contexts. *European Journal of Psychological Assessment*. <https://doi.org/10.1027/1015-5759/a000844>

#### Nutzungsbedingungen:

Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:

<https://creativecommons.org/licenses/by/4.0/deed.de>

#### Terms of use:

This document is made available under a CC BY Licence (Attribution). For more information see:

<https://creativecommons.org/licenses/by/4.0>



# Adapting the BFI-2 Around the World – Coordinated Translation and Validation in Five Languages and Cultural Contexts

Beatrice Rammstedt<sup>1</sup> , Lena Roemer<sup>1</sup> , Clemens M. Lechner<sup>1</sup>, and Christopher J. Soto<sup>2</sup> 

<sup>1</sup>Survey Design and Methodology, GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

<sup>2</sup>Department of Psychology, Colby College, Waterville, ME, USA

**Abstract:** In the course of the PIAAC international pilot studies conducted in 2016 and 2017, the Big Five Inventory-2 (BFI-2; Soto & John, 2017) was translated into and validated in five languages (French, German, Polish, Spanish, and Japanese). Translation was coordinated and conducted centrally following the same state-of-the-art procedures. The performance and comparability of the resulting BFI-2 versions were investigated in parallel in a comprehensive international online study based on quota samples in each country. In this paper, we present the different language versions of the BFI-2 and our investigation of their psychometric properties (reliability, structural, and criterion-related validity) as well as their measurement invariance. Overall, the results reveal high comparability of the psychometric properties across all six versions of the BFI-2 and pairwise between the five adaptations and the English-language original.

**Keywords:** BFI-2, Big Five, translation, measurement invariance, PIAAC



During the last half a century, consensus has grown among personality researchers that a person's personality can be parsimoniously described in terms of the Big Five traits (Goldberg, 1981; John et al., 2008)<sup>1</sup>, namely, Extraversion, Agreeableness, Conscientiousness, Negative Emotionality (or Neuroticism), and Open-Mindedness (or Openness to Experience).

In addition to these global dimensions, researchers have placed increasing emphasis on also examining the more specific facets of the Big Five (Danner et al., 2021) in recent years. Personality traits at the facet level were shown to add incremental validity for a broad range of criteria. For example, Paunonen and Ashton (2001) showed that taking facets into account – in addition to the global Big Five

domains – can incrementally predict academic achievement. The specific consideration of personality facets is all the more warranted because facets – even within the same trait domain – vary in their developmental trajectories (e.g., Brandt, 2023).

However, for a long time, only very few, and comparatively lengthy, instruments were available to assess these more fine-grained facets of the Big Five domains, for example, the Revised NEO Personality Inventory (NEO PI-R; Costa & McCrae, 1992) or the AB5C-IPIP (Goldberg, 1999). To meet the need for a more efficient Big Five measure that also allows assessing the facet structure of the Big Five domains, the well-established Big Five Inventory (BFI; John et al., 1991) was recently revised (Soto & John, 2017). The revised version, the Big Five Inventory-2 (BFI-2), allows the assessment of both the Big Five domains and the three most prototypical facets of each domain.

Due to increased global collaboration and exchange, there is more cross-cultural comparative psychological research (Byrne et al., 2009). For such purposes, measures

<sup>1</sup> Historically, the term “Big Five” has been associated with psycholexical research examining personality-descriptive terms in natural language, whereas the term “Five-Factor Model” has been associated with research examining the content and structure of traditional personality questionnaires (for a review, see John et al., 2008). For the sake of simplicity and readability, we generally use the term “Big Five” throughout this paper, while acknowledging that the connotations of the “Big Five” and the “Five-Factor Model” differ somewhat.

are needed that have been validated in, and are comparable across, multiple languages. The BFI-2 – originally developed for the US population – has therefore been adapted to several languages and cultural contexts (e.g., to Chinese: Zhang et al., 2022; for a full list see Colby Personality Lab, n.d.). These adaptations were conducted to allow the assessment of the BFI-2 in the corresponding language and cultural context. Therefore, the translations were conducted with a focus of comparability of the adapted version with the original English BFI-2. The comparability across these adaptations, however, was not systematically ensured. Further, for each language, these adaptations did not follow the same procedures and standards of questionnaire translation and adaptation, thereby leading to potential variations in measurement quality and especially in cross-cultural comparability.

In psychology and other disciplines, back translation (i.e., translation from the source language into the target language, followed by independent translation of the translated version back into the source language) is still a widely used procedure (Klotz et al., 2023). However, this approach has been widely criticized over the last decades, mainly for two reasons, that is, its focus on the source language to the detriment of the ‘real’ translation and the chance that it may foster too literal a translation. Evidence has shown that other, more elaborate translation methods that focus on the “real” translation and include two forward translations, a reconciliation and committee approach, and some form of testing result in more readable and accurate translations (Acquadro et al., 2008; Behr & Braun, 2023; DuBay et al., 2022; Epstein et al., 2015; Hagell et al., 2010). Leading institutions such as the International Test Commission (ITC, 2017) have accordingly revised their translation and adaptation guidelines and cautioned researchers against relying on a narrow, forward, and backward translation design. State-of-the-art translation procedures, which are used in many international social surveys and recommended by the ITC, are based on a double-translation and reconciliation procedure that is largely consistent with the team-based TRAPD (translation, review, adjudication, pretesting, documentation) approach proposed by Harkness (2003).

Another limitation of the previous monolingual BFI-2 adaptations is that no study has yet compared the different language adaptations jointly with each other and with the US source version using the same translation and validation procedure. In the present study, we therefore aimed (a) to capitalize on high-quality translated versions of the BFI-2 for five languages, namely, French, German, Japanese, Polish, and Spanish<sup>2</sup>; and (b) to examine their psychometric

properties and cross-cultural comparability in a joint investigation. For these national versions, the OECD-initiated translations of the BFI-2 into the target languages followed a double-translation, reconciliation, and review procedure. To empirically validate the translated BFI-2 versions and to evaluate their comparability, a cross-national study was conducted which comprised comparable, diverse samples from the five countries in question as well as data from the United States using the original version of the BFI-2. For the purpose of criterion validity, a set of external variables central in the context of PIAAC was included.

## Method

### Samples

Data were collected as part of pilot studies for the Organisation for Economic Co-operation and Development (OECD) Programme for the International Assessment of Adult Competencies (PIAAC), with the aim of investigating the predictive power of various non-cognitive skills for the second cycle of PIAAC (<https://www.gesis.org/en/piaac/rdc/data/piaac-pilot-studies-on-non-cognitive-skills>). The studies were conducted using an Amazon Mechanical Turk panel based on samples of adult populations in corresponding countries. As demographic quotas were applied for age and gender, the samples were broadly representative of the general population as measured in the respective national census. The first pilot containing only US data was fielded in 2016, the second pilot including France, Germany, Japan, Poland, and Spain, in 2017. The cleaned dataset (excluding  $n = 1,127$  respondents who provided poor-quality data according to a set of data quality indicators such as having indicated to have visited the ISS) used for the present analyses comprised a total of 6,987 respondents, with sample sizes for the various countries ranging from 979 for Japan to 1,328 for Germany. Overall, participants were  $M = 42.8$  years old ( $SD = 12.8$ ), 54% were female. More detailed demographic information of the samples is provided in Table E1 in Electronic Supplementary Material 1 (ESM 1).

### Instruments

#### Big Five Inventory-2

The BFI-2 (Soto & John, 2017) assesses the Big Five personality domains and three facets per domain. It comprises 60 items that respondents rate on a 5-point scale ranging from 1 = *fully disagree* to 5 = *fully agree*.

<sup>2</sup> For all the languages investigated in the PIAAC Pilots conducted in 2016 and 2017, translated versions of the BFI-2 were developed and some even published in the meantime (Gallardo-Pujol et al., 2022; Lignier et al., 2022; Yoshino et al., 2022).

## Other Measures

Based on relevant outcome measures investigated in the context of PIAAC (see Lechner et al. 2019; OECD, 2017; Rammstedt et al., 2024) and based on previous studies on the associations of the Big Five with life outcomes (e.g., Soto, 2019), the following correlates were used to investigate the criterion validity of the BFI-2:

- (a) Self-rated health based on the single item “How would you describe your health status in general?” rated on a scale from 1 = *excellent* to 5 = *poor*, which we recoded such that higher values represented better health.
- (b) Annual gross income based on six categories ranging from 1 = *less than 10%* to 6 = *90% or more*, where the percentages were presented as numbers adjusted to the country-specific income distributions.<sup>3</sup> To improve comparability, we converted scores on income to percent of maximum possible (POMP) scores (Cohen et al., 1999).
- (c) Job satisfaction measured with the single item “All things considered, how satisfied are you with your current job?” rated on a scale from 1 = *extremely satisfied* to 5 = *extremely dissatisfied*, which we recoded such that higher values represented higher satisfaction.
- (d) General life satisfaction measured with the well-established single item (see Nießen et al., 2020) “Overall, how satisfied are you with your life nowadays?” rated on a scale from 1 = *not satisfied at all* to 10 = *completely satisfied*.

## Translations of the Big Five Inventory-2

Translations of the BFI-2 (and the other instruments) were conducted by the language service provider cApStAn, which is highly experienced in state-of-the-art questionnaire translation for international surveys such as PIAAC or PISA. First-draft translations of the BFI-2 existed for Polish and Spanish. As the quality and process of these translations were unclear, new translations for these languages were produced under acknowledgement of the existing translations. For Germany, an existing BFI-2 translation (Danner et al., 2019) developed parallel to the approach described below was used without changes. In all other cases, professional questionnaire translators at cApStAn created two independent translations of the items per country. These translations were then combined into a single pre-final version (most often resulting in an aggregated third solution per item) by means of a discussion and reconciliation process. Big Five experts from the respective countries were asked to review these prefinal versions and recommend

changes where necessary. The final adaptations of the 60 BFI-2 items in the five languages are provided in Table E2 in ESM 1.

## Analyses

We studied the comparability of the psychometric properties of the Spanish, German, French, Polish, and Japanese translations of the BFI-2 domain and facet scales by largely following the analyses reported by Soto and John (2017; Study 3). Specifically, we investigated scale means, reliability coefficients, and the criterion-related and structural validity of the translations and contrasted them with scores from the US source version.

To investigate the reliability of the domain and facet scales, we used Cronbach’s  $\alpha$  and McDonald’s  $\omega$  as indicators. Both Cronbach’s  $\alpha$  and McDonald’s  $\omega$  are measures of the reliability of a unit-weighted scale score. However, while  $\alpha$  assumes an at least essentially  $\tau$ -equivalent model,  $\omega$  assumes a  $\tau$ -congeneric model and can be used even if there are correlated errors (Widaman & Revelle, 2022; Zinbarg et al., 2006). Therefore, we focused mainly on  $\omega$ .

To investigate the structural validity of the domain and facet scales, we used two approaches. In a first step, to examine the *domain-level structures* of the adapted versions and their comparability to the English-language source version, we computed a random intercept exploratory factor analysis (RI-EFA) with an orthogonal acquiescence factor (Aichholzer, 2014) separately for each country. In line with Soto and John (2017), we within-person-centered the items to account for differences in acquiescence. We analyzed (a) the extent to which the orthogonal target-rotated solution exhibited items loadings on the intended vs. non-intended domains (i.e., primary and secondary loadings) and (b) the pairwise congruence of the solutions (Lorenzo-Seva & ten Berge, 2006) for France, Germany, Poland, Japan, and Spain with the US solution.

In a second step, we tested and compared the multidimensional structure at the *facet level* across the different countries using confirmatory factor analyses (CFA). In line with the procedure described by Soto and John (2017), we analyzed the hypothesized structure for each of the five domains separately and allowed the 12 items per domain to load on three correlated factors representing the three facets per domain. Additionally, also as described by Soto and John (2017), we included an orthogonal factor to represent individual differences in acquiescent response style (see also Billiet & McClendon, 2000). All (non-recoded) items had unit loadings on this acquiescence factor. Item

<sup>3</sup> Note that for the United States, income was assessed as household income with nine categories ranging from 1 = *under \$10,000* to 9 = *more than \$150,000*.

responses were treated as interval-scaled, and models were estimated with robust maximum likelihood estimation and robust standard errors and scaled test statistics. We judged the fit of these models against typical cut-offs for fit indices (CFI > .90, RMSEA < .08, SRMR < .08; e.g., Hopwood & Donnellan, 2010; Hu & Bentler, 1999). Given recent contributions cautioning against binary accept-reject decisions and overgeneralizing cut-offs for latent-variable models (e.g., Groskurth et al., 2023), we took these cut-offs as orientations rather than strict criteria.

Finally, to evaluate the comparability of these facet-level measurement models, we conducted a series of measurement invariance tests (e.g., Meredith, 1993). Measurement invariance is of interest because it informs about the extent to which the relations between latent factors and their indicators (i.e., items) are identical across countries. Finding measurement invariance generally helps minimize potential bias in cross-cultural comparisons as it ensures that group differences in means or correlations can be unambiguously attributed to differences in the latent factor; that said, it should be noted that a lack of measurement invariance does not automatically disqualify an instrument from being used for cross-cultural comparisons (e.g., Funder & Gardiner, 2024; Robitzsch & Lüdtke, 2023). We tested whether the measurement structure (configural invariance), loadings (metric invariance), and intercepts (scalar invariance) could be fixed to equality across countries without substantial deterioration in fit (Chen, 2007). Specifically, deterioration of fit greater than  $\Delta\text{CFI} = -.010$ , combined with  $\Delta\text{RMSEA} = .015$  or  $\Delta\text{SRMR} = .030/\Delta\text{SRMR} = .010$  (for metric and scalar invariance, respectively) indicated that invariance was not supported. We examined the measurement invariance both (a) in a pairwise fashion, comparing the U.S. version to each of the five adaptations, and (b) across all country versions simultaneously. All analyses were conducted in R (version 4.3.2, R Core Team, 2023).

## Results

### Descriptive Statistics

Figure 1 displays the domain and facet scale means as well as their standard deviations per country. Table E3 in ESM 1 also lists these coefficients. In general, these means and variances were quite comparable across the investigated countries, and especially compared pairwise with the USA. More specifically, Table E4 in ESM 1 informs about domain and facet score differences above certain thresholds in terms of Cohen's  $d$ . Across all countries, the strongest deviations from the coefficients for the USA were found for Japan, with all domains and facets showing at least small effect sizes (i.e.,  $d \geq 0.20$ ), and 60%, respectively 40% of these scores showing strong effects (i.e.,

$d \geq 0.80$ ). On average, deviations in Japan amounted to  $d = 0.83$  for the domains and  $d = 0.70$  for the facets. Interestingly, all 20 of these scale score differences pointed in the direction of lower socially desirable scores for Japan compared to the USA (and also to the other countries).

Mean score deviations to the USA for all other countries were markedly lower, with 20–40% of the domains and 27–47% of the facets per country showing small effects, and 0% of the domain or facets showing moderate effects (i.e.,  $d \geq 0.50$ ). Average absolute effect sizes for the domains varied between 0.15 (Poland) and 0.17 (France, Germany, and Spain); for the facets the deviations ranged from 0.14 (Poland) to 0.17 (France and Germany). These deviations are also shown graphically in Figure E1 in ESM 1.

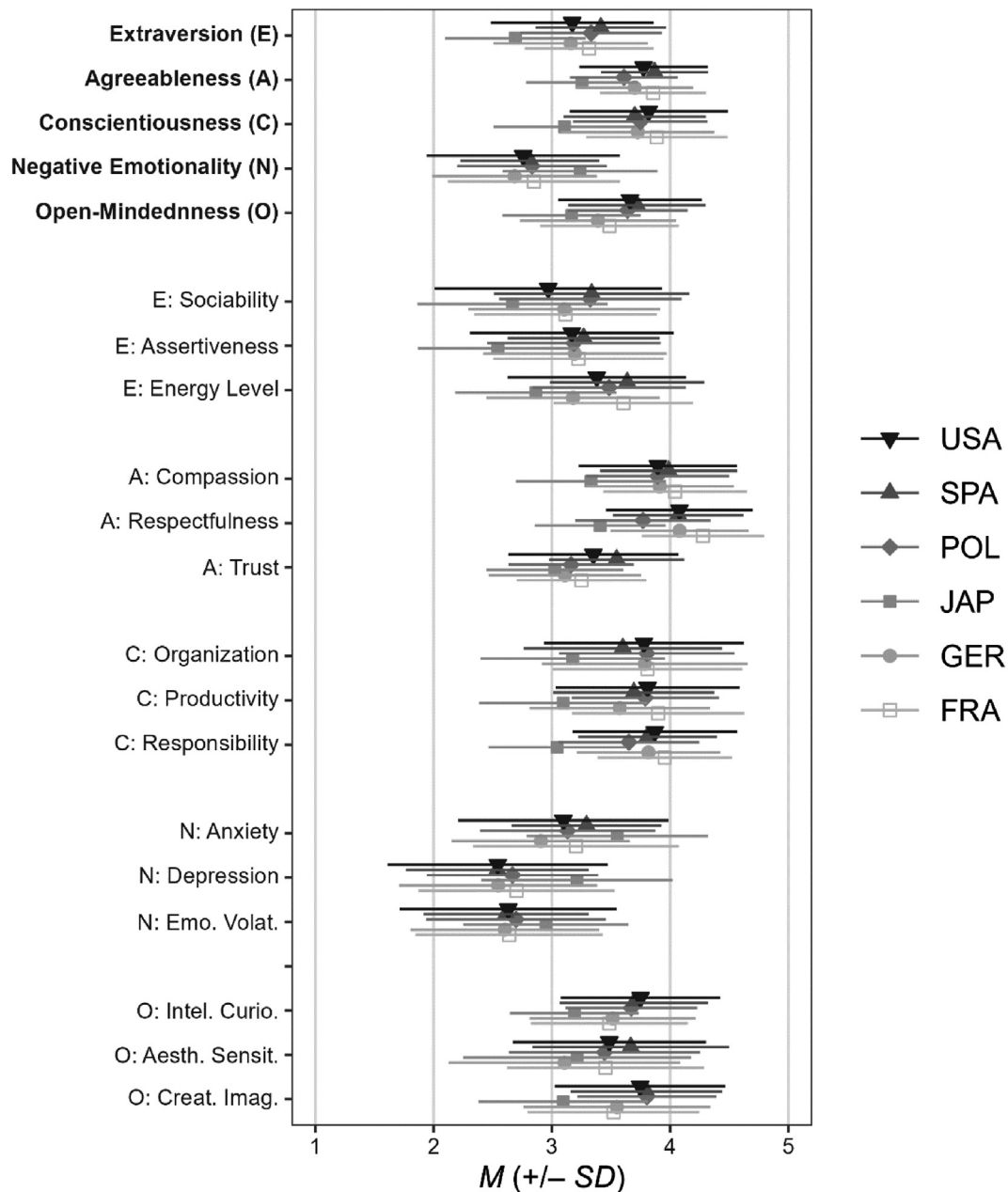
### Reliability Coefficients

The reliability coefficients per country in terms of Cronbach's  $\alpha$  and McDonald's  $\omega$  are displayed in Figure 2 (see also Table E5 in ESM 1). For the domain scales, comprising 12 items each, coefficients reached an average  $\omega$  of .84 ( $SD = 0.05$ ; range: .76–.91). Thus, for none of the countries and none of the domains reliability estimates fell below the minimum standard of .70, according to Nunnally (1978). The resulting estimates were quite homogeneous across countries and mostly similar in size to the US version. That is, average absolute discrepancies from the US estimates were  $\Delta = .03$  for France, Poland, and Spain;  $\Delta = .02$  for Germany and Japan). As evidenced by non-overlapping 95% confidence intervals, significant differences to the US reliability estimates occurred for only one or two domains per country (see Table E5 in ESM 1).

For the 15 facet scales, across all six countries, mean  $\omega$  coefficients ranged between .68 for France and .75 for Germany, with an overall average of .70 ( $SD = 0.10$ ). Surprisingly, for the (original) US and the Spanish version of the BFI-2, nearly half of the facets (47%) did not meet the often-cited minimum standard for reliability estimates of .70 (Nunnally, 1978). In contrast, for Germany, about three-quarters of the scales met the criterion (73%). Compared to the US version, coefficients for Germany also significantly deviated in only a third of the 15 cases (33%,  $\Delta = .06$ ). Slightly more deviations were detected for Poland (40%,  $\Delta = .05$ ), Spain (47%,  $\Delta = .07$ ), and France (53%,  $\Delta = .08$ ). For Japan, only one-third (33%,  $\Delta = .09$ ) of the facet reliability coefficients was comparable in size to the original US version.

### Criterion-Related Validity

To investigate the criterion-related validity of the BFI-2 adaptations, we compared the explained variance ( $R^2$ ) of the Big Five domain and facet scales in the set of criterion



**Figure 1.** Means and standard deviations of the BFI-2 domain and facet scales for the investigated countries. Emo. Volat. = Emotional Volatility; Intel. Curio. = Intellectual Curiosity; Aesth. Sensit. = Aesthetic Sensitivity; Creat. Imag. = Creative Imagination.

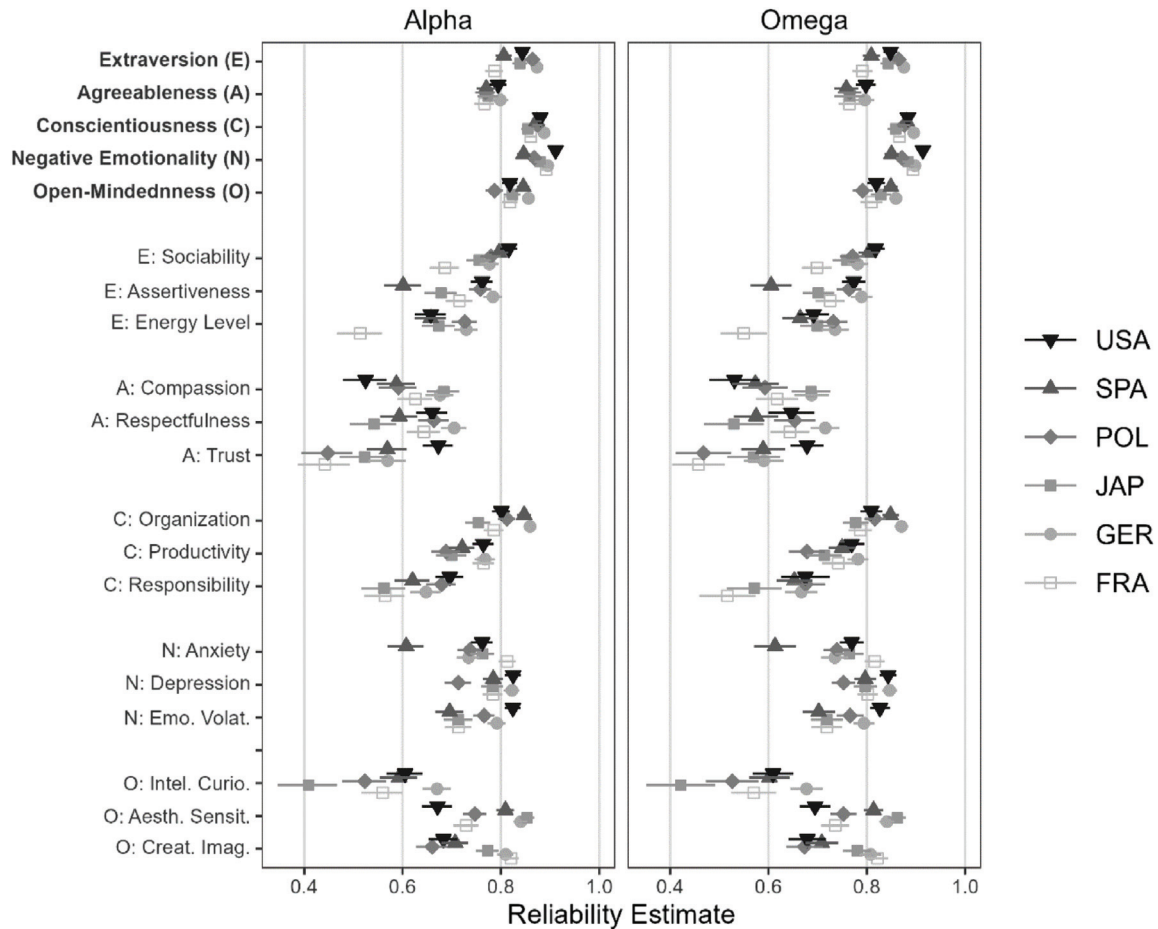
variables described above (Danner et al., 2019; Lechner et al., 2019).

Figure 3 shows the  $R^2$  values for the four criteria as well as averages across these criteria (detailed results are reported in Table E6 in ESM 1, and the full correlational pattern is provided in Table E7 in ESM 1). Across all criteria and all countries, facet scores explained comparatively higher levels of variance (average  $R^2 = .18$ ) in the criteria than the domain scores (average  $R^2 = .13$ ). In line with the literature (Danner et al. 2019; Lechner et al., 2019) and for all countries,  $R^2$ s were highest for life satisfaction

and lowest for income. Comparing the averaged  $R^2$  among the different countries, values for the domain scales were lowest for France and Poland (.10 each) and highest for Germany (.16). For the facet scales, the pattern was similar, with the lowest  $R^2$  value for Poland (.14) the highest for Germany (.23).

## Structural Validity

In the first step, we examined the domain-level structures of the adapted versions and their comparability to the



**Figure 2.** McDonald's  $\omega$  and 95% Confidence intervals of the BFI-2 domain and facet scales for the investigated countries. Emo. Volat. = Emotional Volatility; Intel. Curio. = Intellectual Curiosity; Aesth. Sensit. = Aesthetic Sensitivity; Creat. Imag. = Creative Imagination.

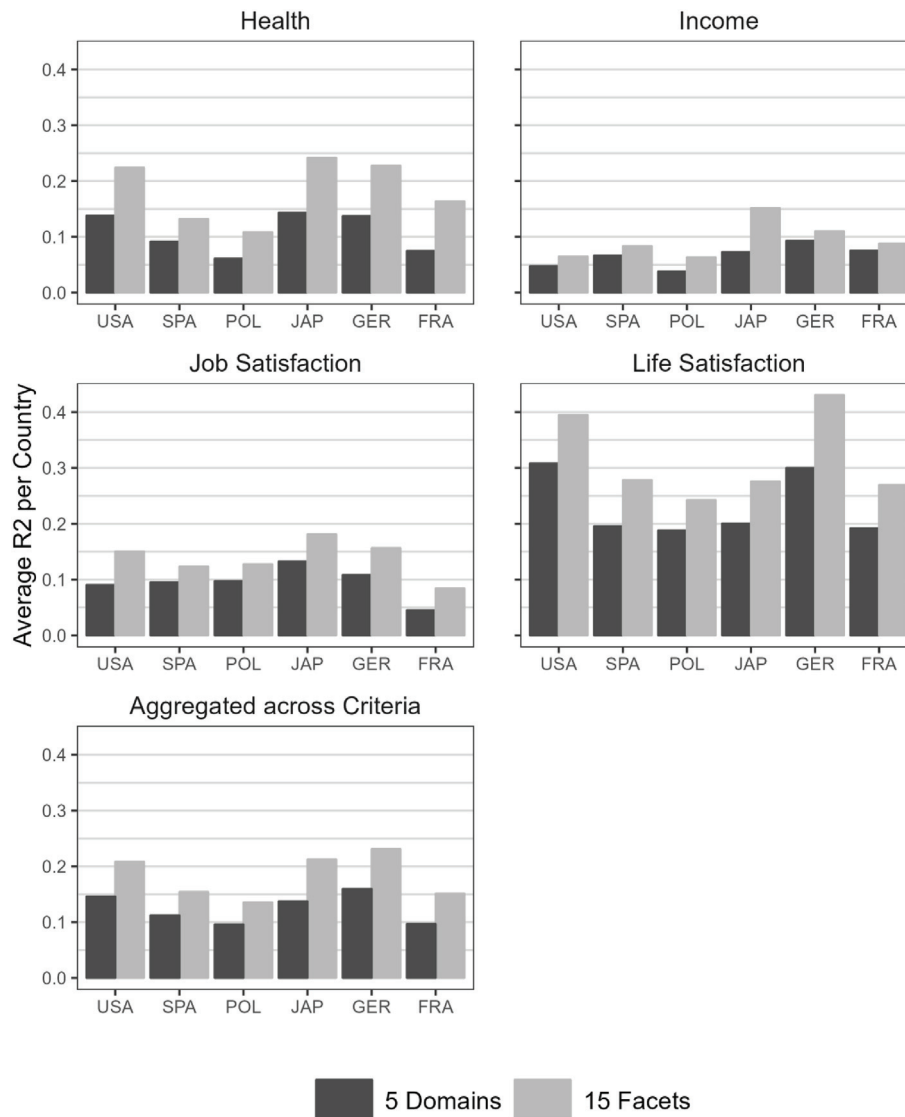
English-language source version. To that aim, we computed RI-EFA (Aichholzer, 2014) with orthogonal target rotation. With very few exceptions (usually 1–5 items, i.e., 1.7–8.3% of the items), the 60 items loaded highest on the corresponding factors in each country (see Table E8 in ESM 1). In Japan, 8 items (i.e., 13.3%) had unexpected highest loadings, and 4 items (i.e., 6.7%) had secondary loadings exceeding .40, which indicates that the factorial structure of the Japanese BFI-2 version was somewhat less robust. Yet, as can be seen from Figure 4, average primary and secondary loadings were quite similar across the investigated countries. Averaged across all domains, primary loadings ranged from .51 for Japan to .56 for Germany. Average secondary loadings varied between .11 for France to .14 for Japan with maximum secondary loadings ranging between .36 in the USA and .59 in Japan. More details on secondary loadings are shown in Table E9 in ESM 1. As a robustness check and to adhere to the standard methodology used in many publications on the BFI-2, we have also examined the domain-level structure using principal component analyses with within-person centered items and Varimax rotation. These solutions, presented in Table E10, yielded

very similar results to those presented in Table E8, showcasing the robustness of our results.

In terms of pairwise comparability with an idealized five-factor solution, in all countries but Japan, congruence coefficients (Tucker's phi) for the loadings on each component (range: .85–.87; for Japan Tucker's phi averaged at .81, see Table E9 in ESM 1) met or exceeded the benchmark for “fairly” similar structures of .85 (Lorenzo-Seva & ten Berge, 2006), indicating that in most countries, the five-factorial solution was well-recovered.

In terms of pairwise comparability with the USA, congruence coefficients for all countries exceeded the benchmark. Averaged across components, congruence ranged between .91 for Japan and .96 for Poland and Spain. Thus, all factorial solutions on the domain-scale level can be regarded as reflecting the solution for the USA.

In a second step, we tested and compared the multidimensional structure at the facet level across the different countries using CFA. Fit indices for the above-described measurement models (i.e., 12 items per trait domain loading on three correlated facet factors and on one acquiescence factor) for the five domains in the six countries are



**Figure 3.** Share of variance explained by the BFI-2 domain and facet scales in the individual criteria and aggregated across all criteria.

displayed in Table 1. The indices can be regarded as generally acceptable. That is, all but three comparative fit index (CFI) values exceeded .90, nearly all root-mean-square error of approximation (RMSEA) values were below .08, and all standardized root-mean-square residual (SRMR) values were below .07. Across all domains, the models fit best for the US source BFI-2 and for Germany and Poland; model fit was often weakest for Japan.

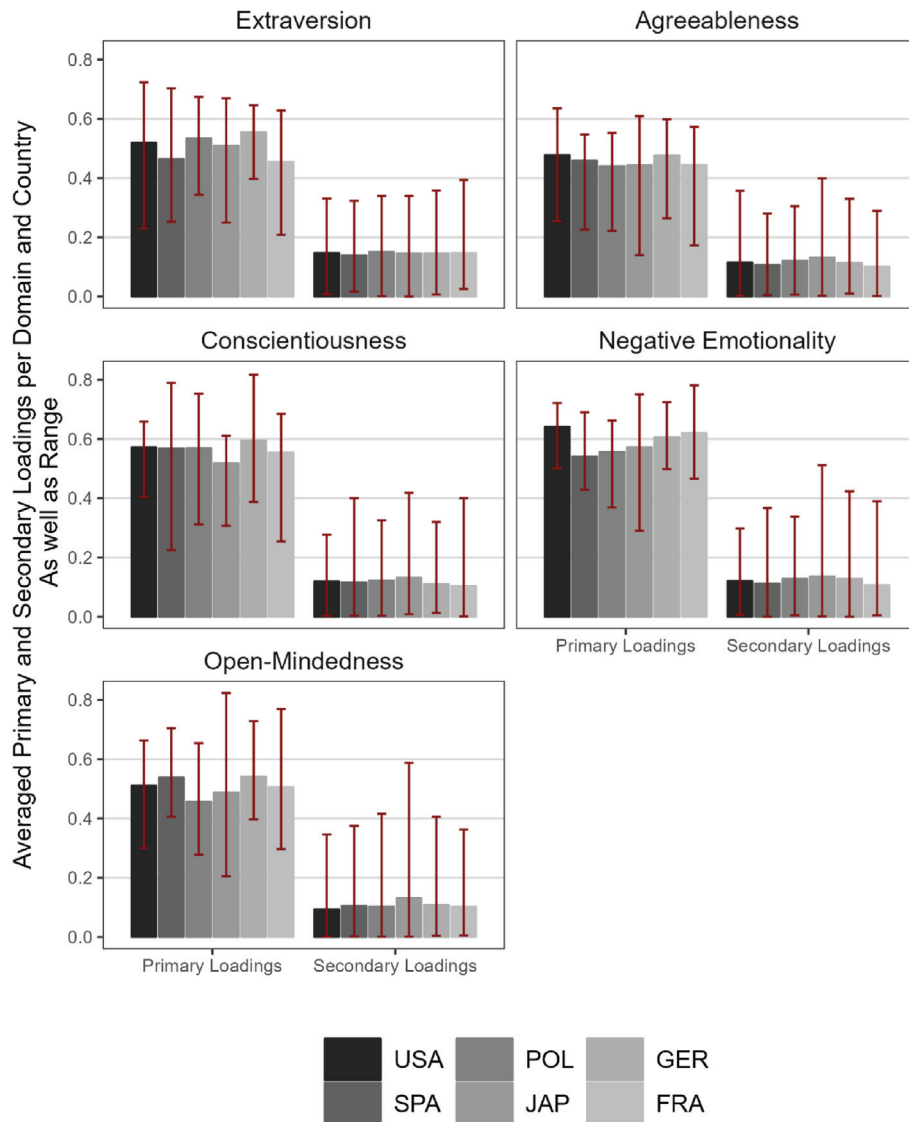
## Measurement Invariance

To formally test the psychometric comparability of the facet-level scores as assessed by the different language versions of the BFI-2, we tested the measurement invariance (e.g., Meredith, 1993) of the above-described measurement

models. In a first step, we tested the measurement invariance in a pairwise fashion. That is, we used the original US version of the BFI-2 as the target against which we compared each of the five adaptations. Results – presented in Table E11 in ESM 1 – indicate that for all domains and in each country (except for Open-Mindedness in Japan), at least metric invariance with the US version can be established.

In the second step, we tested all countries/languages against each other – that is, we repeated the analytical procedures but constrained the structure, loadings, and intercepts to be equal across all six countries simultaneously. As can be seen from the fit indices (Table 2), for all domains except Conscientiousness, the assumptions of metric invariance held across all six countries. However,





**Figure 4.** Averaged primary and secondary loadings per domain and country, as well as their ranges, from the random-intercept exploratory factor analyses with orthogonal target rotation.

although the meaning of the latent facets and the relative strength of relationship between the latent constructs and items of these domains were comparable across all six countries, the rejection of scalar invariance suggests that specific item responses may vary between countries.

## Summary of Results

Table 3 summarizes the main findings of our study. Results for almost all countries reflected the psychometric properties of the original BFI-2 well. Specifically, major deviations were found for the domain and facet scale means in Japan. For Japan, the structural validity in terms of the congruence of the PCA solutions also showed the greatest discrepancies

compared with the US source version. For all other countries, the performance of the BFI-2 adaptations was quite comparable to that of the US version.

## Discussion

Our study aimed to develop and validate parallel adaptations of the BFI-2 in five languages. The translations of the BFI-2 conducted for the present endeavor followed a state-of-the-art double-translation, reconciliation, and review procedure. We investigated the quality of the resulting language versions based on a cross-cultural comparative study using diverse samples in each country that were

**Table 1.** Fit measures for the domain measurement models for BFI-2 separately for the investigated countries

Domain	$\chi^2$	CFI(robust)	RMSEA(robust)	SRMR	BIC
Extraversion					
USA	431.6	.924	.079	.050	39,615
France	421.5	.882	.079	.052	36,274
Germany	330.3	.953	.063	.034	40,070
Japan	447.1	.888	.088	.056	30,034
Poland	292.2	.952	.063	.039	33,491
Spain	423.1	.899	.079	.050	35,858
Agreeableness					
USA	225.4	.954	.052	.036	35,558
France	193.0	.951	.047	.038	31,231
Germany	375.7	.913	.069	.054	37,603
Japan	405.7	.870	.083	.061	28,055
Poland	212.7	.943	.050	.039	32,570
Spain	393.1	.883	.075	.055	31,948
Conscientiousness					
USA	330.3	.952	.066	.041	35,138
France	479.5	.913	.085	.054	32,615
Germany	249.3	.972	.053	.030	37,116
Japan	575.7	.866	.102	.068	29,057
Poland	360.2	.944	.072	.048	30,420
Spain	332.9	.947	.068	.040	32,289
Negative Emotionality					
USA	193.6	.982	.046	.026	37,240
France	396.1	.944	.076	.044	35,661
Germany	452.0	.946	.076	.045	39,054
Japan	359.2	.937	.078	.043	28,878
Poland	245.8	.962	.056	.035	34,822
Spain	221.7	.961	.052	.034	34,962
Open-Mindedness					
USA	192.7	.966	.046	.038	38,102
France	307.3	.944	.065	.036	34,921
Germany	270.2	.964	.056	.038	41,111
Japan	204.8	.964	.053	.045	29,485
Poland	142.9	.974	.037	.034	34,246
Spain	221.2	.963	.052	.038	34,157

Note. CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual. Degrees of freedom in all models = 50.

comparable in their sociodemographic composition. Results indicate that the BFI-2 adaptations perform very similarly in almost all the five languages and that their psychometric quality is comparable to that of the English-language (US) source version. Factorial structures and reliability coefficients were generally comparable across countries and especially to those of the source version. Scale means were also largely similar.

However, our findings also indicate some deviations in terms of psychometric performance from the BFI-2 source version that might either reflect methodological biases in the translations or actual cultural differences. For all countries investigated, criterion validity in particular revealed

differences among the countries in the correlational patterns with the different criteria which might reflect cultural differences in the relevance of different personality domains and facets for these criteria.

While the French, German, Polish, and Spanish versions of the BFI-2 were similarly comparable with the US source version, deviations were often largest for Japan. In Japan, the BFI-2 exhibited the largest scale mean deviations, the lowest reliability coefficients, and the least clear structural validity. As the present results are widely in line with those for a recent Japanese BFI-2 adaptation (Yoshino et al., 2022), we suspect that the observed differences are not caused by the present translation itself. Instead, they may

**Table 2.** Fit measures for the domain measurement models for the BFI-2 across all six countries

Domain	MI type	$\chi^2$	df	CFI (robust)	RMSEA (robust)	SRMR	AIC	BIC2	Interpreted
Extraversion	Configural	2,345.77	300	.923	.075	.046	214,891	215,773	Holds
	Metric	2,881.86	345	.903	.078	.064	215,337	216,053	Holds
	Scalar	6,479.99	385	.763	.116	.092	218,855	219,425	Rejected
	$\Delta$ Metric-configural	536.10	45	-.019	.003	.018	446	281	
	$\Delta$ Scalar-metric	3,598.13	40	-.140	.038	.028	3,518	3,371	
Agreeableness	Configural	1,805.56	300	.920	.064	.047	196,514	197,396	Holds
	Metric	2,134.85	345	.905	.065	.060	196,753	197,470	Holds
	Scalar	3,824.03	385	.812	.086	.078	198,362	198,932	Rejected
	$\Delta$ Metric-configural	329.29	45	-.015	.001	.013	239	74	
	$\Delta$ Scalar-metric	1,689.17	40	-.093	.021	.018	1,609	1,462	
Conscientiousness	Configural	2,327.83	300	.938	.074	.046	196,184	197,065	Holds
	Metric	3,248.25	345	.910	.083	.080	197,014	197,730	Rejected
	Scalar	6,286.58	385	.813	.114	.107	199,972	200,542	Rejected
	$\Delta$ Metric-configural	920.42	45	-.028	.009	.034	830	665	
	$\Delta$ Scalar-metric	3,038.33	40	-.096	.030	.027	2,958	2,811	
Negative Emotionality	Configural	1,868.37	300	.956	.065	.038	210,166	211,048	Holds
	Metric	2,312.31	345	.945	.068	.060	210,520	211,236	Holds
	Scalar	4,219.56	385	.889	.091	.075	212,347	212,917	Rejected
	$\Delta$ Metric-configural	443.94	45	-.012	.003	.023	354	189	
	$\Delta$ Scalar-metric	1,907.25	40	-.055	.023	.014	1,827	1,680	
Open-Mindedness	Configural	1,338.96	300	.962	.052	.038	211,570	212,452	Holds
	Metric	1,947.24	345	.940	.061	.059	212,089	212,805	Holds
	Scalar	4,318.31	385	.847	.092	.080	214,380	214,949	Rejected
	$\Delta$ Metric-configural	608.28	45	-.022	.009	.021	518	353	
	$\Delta$ Scalar-metric	2,371.07	40	-.093	.031	.021	2,291	2,144	

Note. CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual; AIC = Akaike information criterion; BIC2 = Bayesian information criterion 2; DIM = Big Five dimension; MI = measurement invariance. We considered the configural model to hold when CFI > .90, RMSEA < .08, and SRMR < .08.

**Table 3.** Summary and evaluation of the main results

	France		Germany		Japan		Poland		Spain		
	USA Av.	Coeff. $\Delta$ to USA Av.	USA Av.	Coeff. $\Delta$ to USA Av.	USA Av.	Coeff. $\Delta$ to USA Av.	USA Av.	Coeff. $\Delta$ to USA Av.	USA Av.	Coeff. $\Delta$ to USA Av.	
Domains											
Means		.17		.17		.83		.15		.17	
Reliability (McDonald's $\omega$ )	.85	.82	.03	.86	.02	.84	.02	.83	.03	.83	.03
Construct validity (av. $R^2$ with criteria)	.15	.10	.10	.16	.04	.14	.05	.10	.07	.11	.04
Structural validity ( $\varphi$ )	.87	.86	.95	.87	.95	.81	.91	.85	.96	.85	.96
Facets											
Means		.17		.17		.70		.14		.16	
Reliability (McDonald's $\omega$ )	.72	.68	.08	.75	.06	.69	.09	.69	.05	.69	.07
Construct validity (av. $R^2$ with criteria)	.21	.15	.08	.23	.05	.21	.06	.14	.06	.15	.05
Measurement invariance		3 $\times$ metric, 2 $\times$ scalar		2 $\times$ metric, 3 $\times$ scalar		1 $\times$ config., 2 $\times$ metric, 2 $\times$ scalar		2 $\times$ metric, 3 $\times$ scalar		3 $\times$ metric, 2 $\times$ scalar	

be due to a combination of linguistic and cultural differences between Japan (the only Asian country investigated here), on the one hand, and the Central and Western European countries and the USA, on the other hand. Overall, the results of both validation studies examining Japanese BFI-2 adaptations suggest that the BFI-2's intended multidimensional structure can be recovered in Japan, but that there may be greater differences in this cultural context than for some other translations. It is especially noteworthy that, in both studies, the Japanese participants see themselves, on average, as less socially desirable, in terms of their personality traits, compared to the US and European participants. Interestingly, this effect was not found for a recent Chinese BFI-2 adaptation (Zhang et al., 2022), suggesting that the Japanese population in particular may be more modest and less self-enhancing than other cultures around the world (see also Heine & Hamamura, 2007). Future research using the BFI-2 can further investigate this possibility.

In this context, analyses on partial measurement invariance might be particularly fruitful, as they may inform about potential differences in the interpretation of specific items. Similarly, examining measurement invariance of the different versions across additional split criteria such as age, gender, or education might allow for additional insights into the quality of the translation and potential cross-cultural differences. Especially with regard to education, previous research (e.g., Rammstedt et al., 2013; Soto & John, 2017) has suggested that lower-educated respondents have a higher tendency for acquiescence which in turn reduces the response quality in these subpopulations.

For all the languages investigated, other translated versions of the BFI-2 also exist (Colby Personality Lab, n.d.; Danner et al., 2019; Gallardo-Pujol et al., 2022; Lignier et al., 2022; Yoshino et al., 2022). The degree of overlap of these versions with those presented here varies. As described above, only for Germany could we use a translation of the BFI-2 that had already been validated at the time of testing and that followed the same translation approach as the one applied here. For other countries, such as Poland, translation efforts for the BFI-2 had only started in 2017. In these cases, we informed the local teams of our plans and translation results. This might have led to strongly overlapping BFI-2 translations. The major advantage of the BFI-2 translations presented here is that they were translated by a single professional provider following best practices of questionnaire translation and the same principles for each language and that all the translations were geared toward comparability across countries and languages. Comparing these translations in terms of reliability, validity, and especially comparability (measurement invariance) with other translations for the same languages that have since been published might be a fruitful task for future research.

## Conclusion

In sum, our study demonstrates the high psychometric quality and measurement invariance of five adaptations of the BFI-2, which were conducted using a state-of-the-art double-translation, reconciliation, and review procedure applied consistently for all language versions. By making these adapted BFI-2 versions available to the research community, we hope to enhance the quality of cross-cultural personality research. More broadly, our findings highlight the potential utility of coordinating translation and validation procedures across multiple languages and cultural contexts.

## Electronic Supplementary Material

The following electronic supplementary material is available at <https://doi.org/10.1027/1015-5759/a000844>

**ESM 1.** Supplementary information and tables.

## References

- Acquadro, C., Conway, K., Hareendran, A., & Aaronson, N. (2008). Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. *Value in Health, 11*(3), 509–521. <https://doi.org/10.1111/j.1524-4733.2007.00292.x>
- Aichholzer, J. (2014). Random intercept EFA of personality scales. *Journal of Research in Personality, 53*, 1–4. <https://doi.org/10.1016/j.jrp.2014.07.001>
- Behr, D., & Braun, M. (2023). How does back translation fare against team translation? An experimental case study in the language combination English–German. *Journal of Survey Statistics and Methodology, 11*(2), 285–315. <https://doi.org/10.1093/jssam/smac005>
- Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling: A Multidisciplinary Journal, 7*(4), 608–628. [https://doi.org/10.1207/S15328007SEM0704\\_5](https://doi.org/10.1207/S15328007SEM0704_5)
- Brandt, N. D., Drewelies, J., Willis, S. L., Schaie, K. W., Ram, N., Gerstorf, D., & Wagner, J. (2023). Beyond Big Five trait domains: Stability and change in personality facets across midlife and old age. *Journal of Personality, 91*(5), 1171–1188. <https://doi.org/10.1111/jopy.12791>
- Byrne, B. M., Oakland, T., Leong, F. T., van de Vijver, F. J., Hambleton, R. K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3*(2), 94–105. <https://doi.org/10.1037/a0014516>
- 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>
- Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research, 34*(3), 315–346. [https://doi.org/10.1207/S15327906MBR3403\\_2](https://doi.org/10.1207/S15327906MBR3403_2)

- Colby Personality Lab. (n.d.). *The Big Five Inventory-2 (BFI-2)*. <https://www.colby.edu/academics/departments-and-programs/psychology/research-opportunities/personality-lab/the-bfi-2/>
- Costa, P. T., & McCrae, R. R. (1992). *NEO Personality Inventory-Revised (NEO PI-R)*. Psychological Assessment Resources.
- Danner, D., Rammstedt, B., Bluemke, M., Lechner, C., Berres, S., Knopf, T., Soto, C., & John, O. (2019). Das Big Five Inventar 2: Validierung eines Persönlichkeitsinventars zur Erfassung von 5 Persönlichkeitsdomänen und 15 Facetten [The German Big Five Inventory 2: Measuring five personality domains and 15 facets]. *Diagnostica*, 65(3), 121–132. <https://doi.org/10.1026/0012-1924/a000218>
- Danner, D., Lechner, C. M., Soto, C. J., & John, O. P. (2021). Modelling the incremental value of personality facets: The Domains-Facets-Acquiescence-Bifactor (DFAB) model. *European Journal of Personality*, 35(1), 67–87. <https://doi.org/10.1002/per.2268>
- DuBay, M., Sideris, J., & Rouch, E. (2022). Is traditional back translation enough? Comparison of translation methodology for an ASD screening tool. *Autism Research*, 15(10), 1868–1882. <https://doi.org/10.1002/aur.2783>
- Epstein, J., Osborne, R. H., Elsworth, G. R., Beaton, D. E., & Guillemin, F. (2015). Cross-cultural adaptation of the Health Education Impact Questionnaire: Experimental study showed expert committee, not back-translation, added value. *Journal of Clinical Epidemiology*, 68(4), 360–369. <https://doi.org/10.1016/j.jclinepi.2013.07.013>
- Funder, D. C., & Gardiner, G. (2024). Misgivings about measurement invariance. *European Journal of Personality*. Advance online publication. <https://doi.org/10.1177/08902070241228338>
- Gallardo-Pujol, D., Rouco, V., Cortijos-Bernabeu, A., Oceja, L., Soto, C. J., & John, O. P. (2022). Factor structure, gender invariance, measurement properties, and short forms of the Spanish adaptation of the Big Five Inventory-2. *Psychological Test Adaptation and Development*, 3(1), 44–69. <https://doi.org/10.1027/2698-1866/a000020>
- Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 141–165). Sage.
- Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg University Press.
- Groskurth, K., Bluemke, M., & Lechner, C. M. (2023). Why we need to abandon fixed cutoffs for goodness-of-fit indices: An extensive simulation and possible solutions. *Behavior Research Methods*. Advance online publication. <https://doi.org/10.3758/s13428-023-02193-3>
- Hagell, P., Hedin, P.-J., Meads, D. M., Nyberg, L., & McKenna, S. P. (2010). Effects of method of translation of patient-reported health outcome questionnaires: A randomized study of the translation of the Rheumatoid Arthritis Quality of Life (RAQoL) instrument for Sweden. *Value in Health*, 13(4), 424–430. <https://doi.org/10.1111/j.1524-4733.2009.00677.x>
- Harkness, J. (2003). Questionnaire translation. In J. Harkness, F. J. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Wiley.
- Heine, S. J., & Hamamura, T. (2007). In search of East Asian self-enhancement. *Personality and Social Psychology Review*, 11(1), 4–27. <https://doi.org/10.1177/1088868306294587>
- Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, 14(3), 332–346. <https://doi.org/10.1177/1088868310361240>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). ITC. [https://www.intestcom.org/files/guideline\\_test\\_adaptation\\_2ed.pdf](https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf)
- John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory: Versions 4a and 54*. University of California, Berkeley, Institute of Personality and Social Research.
- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). The Guilford Press.
- Klotz, A. C., Swider, B. W., & Kwon, S. H. (2023). Back-translation practices in organizational research: Avoiding loss in translation. *The Journal of Applied Psychology*, 108(5), 699–727. <https://doi.org/10.1037/apl0001050>
- Lechner, C. M., Anger, S., & Rammstedt, B. (2019). Socioemotional skills in education and beyond: Recent evidence and future research avenues. In R. Becker (Ed.), *Research handbook on the sociology of education* (pp. 427–453). Edward Elgar Publishing.
- Lignier, B., Petot, J.-M., Canada, B., De Oliveira, P., Nicolas, M., Courtois, R., John, O. P., Plaisant, O., & Soto, C. (2022). Factor structure, psychometric properties, and validity of the Big Five Inventory-2 facets: Evidence from the French adaptation (BFI-2-Fr). *Current Psychology*, 42, 26099–26114. <https://doi.org/10.1007/s12144-022-03648-0>
- Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2), 57–64. <https://doi.org/10.1027/1614-2241.2.2.57>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543. <https://doi.org/10.1007/BF02294825>
- Nießen, D., Groskurth, K., Rammstedt, B., & Lechner, C. M. (2020). *General Life Satisfaction Short Scale (L-1)*. <https://doi.org/10.6102/zis284>
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.
- Organisation for Economic Co-operation, Development (OECD). (2018a). *Programme for the International Assessment of Adult Competencies (PIAAC), International Pilot Study on Non-Cognitive Skills*. [Open data]. <https://doi.org/10.4232/1.13063>
- Organisation for Economic Co-operation, Development (OECD). (2018b). *Programme for the International Assessment of Adult Competencies (PIAAC), English Pilot Study on Non-Cognitive Skills*. [Open data]. <https://doi.org/10.4232/1.13062>
- Paunonen, S. V., & Ashton, M. C. (2001). Big Five predictors of academic achievement. *Journal of Research in Personality*, 35(1), 78–90. <https://doi.org/10.1006/jrpe.2000.2309>
- Rammstedt, B., Kemper, C. J., & Borg, I. (2013). Correcting Big Five 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., Lechner, C., & Danner, D. (2024). *Beyond Literacy: The incremental value of non-cognitive skills* [OECD Education Working Papers, No. 311]. OECD. <https://doi.org/10.1787/7d4fe121-en>
- R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, 30(6), 859–870. <https://doi.org/10.1080/10705511.2023.2191292>

- Roemer, L. (2024). *Adapting the BFI-2 around the world: Coordinated translation and validation in five languages and cultural contexts*. [Open materials and code]. <https://osf.io/tydb7/>
- Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science*, 30(5), 711–727. <https://doi.org/10.1177/0956797619831612>
- 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>
- Widaman, K. F., & Revelle, W. (2022). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, 55, 788–806. <https://doi.org/10.3758/s13428-022-01849-w>
- Yoshino, S., Shimotsukasa, T., Oshio, A., Hashimoto, Y., Ueno, Y., Mieda, T., Migiwa, I., Sato, T., Kawamoto, S., Soto, C. J., & John, O. P. (2022). A validation of the Japanese adaptation of the Big Five Inventory-2. *Frontiers in Psychology*, 13, Article 924351. <https://doi.org/10.3389/fpsyg.2022.924351>
- Zhang, B., Li, Y. M., Li, J., Luo, J., Ye, Y., Yin, L., Chen, Z., Soto, C. J., & John, O. P. (2022). The Big Five Inventory-2 in China: A comprehensive psychometric evaluation in four diverse samples. *Assessment*, 29(6), 1262–1284. <https://doi.org/10.1177/10731911211008245>
- Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for  $\omega_h$ . *Applied Psychological Measurement*, 30(2), 121–144. <https://doi.org/10.1177/0146621605278814>

## History

Received November 10, 2023

Revision received March 16, 2024

Accepted April 3, 2024

Published online July 12, 2024

EJPA Section / Category Personality

## Conflict of Interest

The authors have no known conflict of interest to disclose.

## Open Science

Open Data: The authors confirm that there is sufficient information for an independent researcher to reproduce all of the reported results, including codebook if relevant (OECD, 2018a, 2018b).

Open Materials: The authors confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology (Roemer, 2024; <https://osf.io/tydb7/>).

Open Analytic Code: The scripts, code, and outputs needed to reproduce all of the reported results are available at <https://osf.io/tydb7/> (Roemer, 2024).


Preregistration of Studies and Analysis Plans: This study was not preregistered.

## Funding

Open access publication enabled by GESIS Leibniz Institute for Social Sciences, Mannheim, Germany.

## ORCID

Beatrice Rammstedt

 <https://orcid.org/0000-0002-6941-8507>

Lena Roemer

 <https://orcid.org/0000-0002-5885-4426>

Chris Soto

 <https://orcid.org/0000-0002-2875-8919>

## Beatrice Rammstedt

GESIS – Leibniz Institute for the Social Sciences

PO Box 12 21 55

68072 Mannheim

Germany

[beatrice.rammstedt@gesis.org](mailto:beatrice.rammstedt@gesis.org)