

Consistency of the Structural Properties of the BFI-10 Across 16 Samples From Eight Large-Scale Surveys in Germany

Rammstedt, Beatrice; Roemer, Lena; Lechner, Clemens

Veröffentlichungsversion / Published Version

Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:

GESIS - Leibniz-Institut für Sozialwissenschaften

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 491156185 / Funded by the German Research Foundation (DFG) - Project number 491156185

Empfohlene Zitierung / Suggested Citation:

Rammstedt, B., Roemer, L., & Lechner, C. (2023). Consistency of the Structural Properties of the BFI-10 Across 16 Samples From Eight Large-Scale Surveys in Germany. *European Journal of Psychological Assessment*, 1-12. <https://doi.org/10.1027/1015-5759/a000765>

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>

Consistency of the Structural Properties of the BFI-10 Across 16 Samples From Eight Large-Scale Surveys in Germany

Beatrice Rammstedt^{ID}, Lena Roemer^{ID}, and Clemens M. Lechner

GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

Abstract. The assessment of the Big Five personality domains is standard practice in most large-scale social surveys nowadays. The instrument most widely used for this purpose is the BFI-10, an ultra-short measure assessing each Big Five domain with two items. Recent studies have identified issues with the structural properties of the BFI-10, especially its factorial validity. To investigate whether these issues arise from the instrument itself or biases due to translation or sampling, we examined the extent to which the structural properties of the BFI-10 in terms of descriptive statistics, intercorrelations, reliability, and factorial validity vary when keeping the target population and language constant. Results revealed that, across 16 independent samples (total $N \sim 60,000$) from eight large-scale surveys representative of the adult population in Germany, the structural properties of the BFI-10 were (a) largely consistent and (b) mostly adequate. Most importantly, in nearly all samples, patterns of loading were congruent with an idealized Big Five structure, thereby supporting factorial validity. These results demonstrate that the structural properties of the BFI-10 are highly stable and replicable in large-scale samples. Especially given its brevity, the BFI-10 can thus be regarded as adequate for use in large-scale survey settings.

Keywords: BFI-10, Big Five, psychometric properties, large-scale surveys, factorial validity



In recent decades, the Big Five – Extraversion, Agreeableness, Conscientiousness, Neuroticism/Emotional Stability, and Openness to Experience – have become increasingly established as a comprehensive description of personality (e.g., John et al., 2008; McCrae & Costa, 2008). This led to increased interest in their assessment in studies in fields outside core personality research, such as sociology, economics, and epidemiology. Most of these studies were characterized by the fact that (a) they usually focused on topics other than personality, and (b) they aimed to assess the Big Five domains as a correlate rather than for an individual diagnostic purpose. Thus they needed an ultra-efficient measure with psychometric properties that allowed group comparisons.

To meet this need, Rammstedt and John (2007) developed the 10-item Big Five Inventory (BFI-10), an ultrashort version of the established Big Five Inventory (BFI-44; John et al., 1991, 2008). As the BFI-10 assesses each Big Five

domain with two items, it allows the assessment of personality in research settings with severe time constraints. A host of studies have found evidence supporting the psychometric quality of the BFI-10, including its reliability, structural or factorial validity, and – most importantly – its criterion validity, which is often similar to that of much longer inventories (Rammstedt et al., 2021; see also Thalmayer et al., 2011).

Since its publication in 2007, the BFI-10 has become one of the most widely used measures to assess the Big Five in large-scale surveys, especially in Germany but also beyond, with nearly 5,000 citations as per Google Scholar of the original articles at the time of writing (see also Rammstedt et al., 2023). It has been implemented in numerous national and international comparative studies in various disciplines and has been adapted to many different languages worldwide. Numerous analyses have been conducted on the associations of the Big Five with the various outcomes assessed in these studies (e.g., Fischer & Karl, 2021; Rammstedt, 2007; Rammstedt et al., 2015).

Many of the studies that have used the inventory for substantive research have not investigated its psychometric properties in their respective samples, but have instead

relied on the evidence presented in the original studies on the inventory (e.g., Rammstedt & John, 2007). However, some recent studies that did investigate the psychometric quality of the BFI-10 across multiple samples have raised concerns about its structural properties, in particular its factorial validity. Studies analyzing the psychometric properties of the BFI-10 in the Survey of Health, Ageing and Retirement (SHARE; Levinsky et al., 2019) and the World Values Survey (WVS; Ludeke & Larsen, 2017) suggest that factorial validity can vary substantially across studies, particularly when the BFI-10 is fielded in multiple countries and languages. In some countries in SHARE, and especially in WVS, these studies reported abnormal inter-item relationships and a lack of factorial validity for four of the five dimensions and concluded that the resulting personality data were highly problematic.

These studies raised concerns about the stability and replicability of the BFI-10's structural properties, and consequently the validity of interpretations based on its scores. Recent re-analyses of the WVS data suggest that the psychometric shortcomings of the BFI-10 in the WVS stem primarily from translation problems, and can be partially remedied by focusing on a subset of countries and items (Lu & Cui, 2022). Overall, however, it remains unclear whether the variation in the structural properties of the BFI-10 across different survey programs and samples reflects variation caused by (controllable) methodological factors such as cultural adaptations or different sampling procedures, or whether it is due instead to random and uncontrollable fluctuations. As a first step toward resolving this important question, one needs to establish a baseline of how strongly the structural properties of the BFI-10 vary across survey samples when the surveys are administered in the same language to the same target population – that is, in the absence of cultural differences, translation issues, and sample selectivity that may have led to the differences in the psychometric quality of the instrument observed in earlier work (Levinsky et al., 2019; Ludeke & Larsen, 2017).

The present study provides one of the most extensive analyses of the BFI-10's structural properties (i.e., descriptive statistics, reliabilities, item and scale intercorrelations, and factorial validity) to date. Our goal was to establish a baseline of how strongly these structural properties vary when keeping the language and target population constant. We combined data from eight large-scale surveys conducted in Germany to gauge the stability and replicability of the BFI-10's structural properties across a total of 16 adult samples drawn from the same population.

We defined the database for our investigation as comprehensively as possible, while at the same time holding constant central aspects that have been shown to affect structural aspects of personality scales. Thus, to avoid potential methodological biases, we applied the following

selection criteria: First, to avoid language and/or cultural biases, we included only studies conducted in Germany and German. Second, to avoid sample selectivity and possible range restriction, we included only studies based on random samples covering a broad age range of the general adult population. Third, we included only studies whose data were accessible to researchers.

Method

Samples

In order to identify studies meeting our selection criteria we followed a three-step search strategy: In the first step, we included surveys that we – having worked extensively with several of these data sources – were personally aware of. Second, we searched the databases of several German data research centers for the keywords personality and BFI-10. Third, we screened international survey programs we were aware of for their inclusion in the BFI-10.

Based on the aforementioned criteria, we included eight study programs that had administered the BFI-10: (1) the German General Social Survey (ALLBUS; GESIS, 2011a, 2011b, 2015); (2) the German Internet Panel (GIP; Blom et al., 2015); (3) the German Longitudinal Election Study (GLES, 2019a, 2019b, 2022); (4) the GESIS Panel (Bosnjak et al., 2018; GESIS, 2022); (5) the “Older Adults” sample of the Jena Study on Social Change and Human Development (Silbereisen et al., 2008; see also Lechner & Rammstedt, 2015); (6) the adult cohort of the German National Educational Panel Study (NEPS; Blossfeld & von Maurice, 2011; NEPS Network, 2021); as well as the German surveys within the framework of (7) the Study on Health and Retirement (SHARE; Bergmann et al., 2019; Börsch-Supan, 2022; Börsch-Supan et al., 2013) and (8) the World Values Survey (WVS; Inglehart et al., 2018). Whereas most of these programs focus on the general adult population, SHARE, and the Jena Study focus on older adults (see Table 1 for an overview of the different programs and samples). In the Open Science section, we provide links to the program websites, where further information about the studies can be accessed.

Because some of these programs included the BFI-10 multiple times based on separate samples, these samples were treated separately in the present study. This yielded 16 samples ranging in size from $N = 868$ to $N = 11,689$ respondents and comprising $N = 57,986$ respondents in total (see Table 1).

Instruments

The BFI-10 consists of 10 phrase-like items (e.g., “I see myself as someone who ... is outgoing, sociable”) – one

Table 1. Overview of the eight study programs and 16 samples

Study program	Subsample/ Substudy (if applicable)	Year of (first) BFI-10 assessment	Sampling procedure	N	Female (%)	Age M (SD)	Age range	Average nonresponse on BFI-10 items (%)
ALLBUS	ISSP 2003/04	2004	Register	2,609	51	47.56 (17.29)	18–91	1.59
	ISSP 2005/06	2006	Register	6,664	52	49.28 (17.21)	18–94	1.45
	ISSP 2007/08	2008	Register	6,814	51	50.21 (17.78)	18–97	1.62
GLES	Post-election cross-section	2017	Register	2,105	48	50.10 (19.19)	16–95	0.22
	Pre-election cross-section	2017	Register	2,175	50	51.23 (19.07)	16–96	0.29
	Pre-election cross-section	2021	Register	5,036	49	52.04 (18.10)	16–89	0.70
WVS	Wave 6	2013	Register	2,041	50	49.47 (17.70)	17–95	0.90
GIP ¹	Wave 1	2012	Random route	1,465	50	45.14 (15.34)	15–75	0.18
	Wave 13	2014	Random route	3,189	50	45.79 (15.66)	17–77	0.13
	Wave 37	2018	Register	2,892	49	45.75 (16.01)	18–76	0.05
GESIS Panel	Cohort 1	2014	Register	4,020	53	47.54 (14.21)	19–71	1.08
	Cohort 2	2016	Register	1,143	51	51.52 (15.83)	21–73	1.01
	Cohort 3	2018	Register	868	52	51.93 (15.76)	23–75	1.59
Jena Study	Older adults survey	2009	Random route	1,508	52	65.59 (5.87)	56–76	0.11
NEPS	SC6	2013	Register	11,689	51	49.57 (10.98)	27–69	0.06
SHARE	Wave 7	2017	Register	3,768	53	67.28 (9.41)	35–95	0.26

Note. ¹In the GIP samples, information on age was provided in categories, each spanning 5 years. To analyze the average age of the samples, we used the midpoint as a proxy for the participants' age to analyze the samples' average age.

positively coded and one negatively coded item for each Big Five domain. The items are answered on a 5-point scale ranging from 1 = *disagree strongly* to 5 = *agree strongly*. The BFI-10 was developed simultaneously in English and German as an abbreviated version of the BFI-44 (Rammstedt & John, 2007). The two items per domain were selected to cover the maximum bandwidth of the underlying Big Five domains. Thus, the intention was to achieve within-trait inter-item heterogeneity rather than homogeneity.

The initial validation study of the BFI-10 (Rammstedt & John, 2007) was based on several – highly selective – samples of university (mostly psychology) students. Results indicated that in all samples investigated (a) the five BFI-10 domain scales were largely independent, with a mean inter-correlation of .11 and with none of the correlations reaching [.25]; and (b) the 10 items represented based on principal component analyses (PCA) followed by Varimax rotation the intended five-factor structure, with all items showing clear simple-structure solutions.

For some of the study programs included in our study, the implementation of the BFI-10 differed slightly from the original validation study. Specifically, item formulations were different for ALLBUS 2004 and the WVS: In ALLBUS 2004, an alternative negatively coded Agreeableness item was administered, namely, “can be cold and aloof” instead of “tends to find fault with others.” The German WVS survey used an alternative translation of the BFI-10, with slight differences in all 10 items (see, e.g., the methods

report for Germany: <https://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp>). In all ALLBUS samples and the Jena Study on Social Change and Human Development, different response formats were used, mainly to align the BFI-10 assessment with the overall questionnaire. In the ALLBUS samples, the response scale was inverted, thus ranging from 1 = *agree strongly* to 5 = *disagree strongly*. In the Jena Study, a 7-point response format was used instead of the original 5-point scale.

Results

We investigated the structural aspects of the psychometric quality of the BFI-10 in and across all 16 samples in terms of (a) the scale descriptives, (b) their reliabilities, and, most importantly, (c) the factorial validity of the instrument. On the project's Open Science Framework (OSF) web page (<https://osf.io/bup8s/>), we provide the analysis code and a more detailed presentation of the results – also on the item level. As our central aim was to compare the results across samples and to provide an aggregated estimate of the psychometric quality of the BFI-10 across all samples, we present the following results with a focus on their comparability – that is, their similarities and differences across samples. To this aim, we pooled our data, merging the datasets from the single samples into one. Pooled analyses assume homogeneity across datasets and produce one overarching set of results, assumed to be the effect size

underlying the single studies. Accordingly, pooled analyses have strict harmonization requirements (e.g., identical target population, identical measures; Curran & Hussong, 2009), which were fulfilled in the current study (note that the Jena Study used a different response scale such that these data were not pooled with the other data). This pooled dataset can hence be used to conduct a comprehensive approximation of the psychometric structure of the BFI-10 in the German adult population, which the single studies can be judged against. In Table S1, we provide the pooled correlation matrix to facilitate the reuse of these results.

In the first step, we investigated to what degree item non-response – as a proxy for the acceptance of the instrument among respondents – differed across samples. We, therefore, compared the average item nonresponse for the 10 items across samples. Results reveal generally very low item nonresponse, with an average of less than 1% per item (for the average nonresponse across all BFI-10 items by sample, see Table 1).

Figure 1 shows the distribution of the means and their 95% confidence intervals (CI) for the five BFI-10 scales across the samples, as well as for the pooled data. Not surprisingly, CIs based on a pooled sample of $N \sim 60,000$ respondents were very small (between 0.01 and 0.02). Tables S2 and S3 in the supplemental material display the means and the standard deviations, too, also at the item level. Overall, the means were highly homogeneous. The narrow CIs were partially overlapping, indicating some significant differences to the overall, pooled mean. In terms of effect sizes, these differences were rather small. The average absolute Cohen's d of the differences between the single studies and the pooled data means ranged between $d = .04$ for Openness and $d = .18$ for Agreeableness (see Table S4). The latter was due mainly to the comparatively high average Agreeableness score of 3.79 in the WVS, which, as noted above, used slightly different item formulations.

In a second step, we investigated the convergent item correlations (i.e., the correlation between the two items of each Big Five domain) and the mean divergent item correlations (i.e., the Fisher's z and back-transformed average of the absolute correlations of the two items of one Big Five domain with all eight items of the other four domains), as

well as the 95% CIs of these (aggregated) correlations (Borenstein et al., 2009). Relatedly, on the scale level, we computed standardized Cronbach's α (see Eisinga et al., 2013) and McDonald's Ω . However, in the present case, the overall size of these internal consistency estimates cannot be interpreted as an indicator of the scale's quality, as the two items per domain were selected to represent a domain's heterogeneity, not its homogeneity. Figure 2 and Table S5 show the convergent and mean absolute divergent correlations by BFI-10 domain and averaged across all domains, and Figure 3 and Table S6 show the internal consistency estimates.

Here, too, results are rather homogeneous across the samples. In nearly all cases, convergent correlations exceeded mean absolute divergent correlations. Generally, this difference between convergent and divergent correlations was highest for Extraversion (in the pooled sample, $\Delta = .34$) and lowest for Agreeableness ($\Delta = .04$). Pooled across all samples, internal consistency averaged across domains at $\alpha = .44$, $\Omega = .53$ with mean coefficients for the five scales ranging from .18, respectively .36 for Agreeableness to .60, respectively .63 for Extraversion.

In the third and most crucial step, we investigated the extent to which the 10 BFI-10 items represented the intended five-dimensional structure. In line with the BFI literature (e.g. Rammstedt & John, 2007) we ran separate PCAs for the 16 samples, with a forced extraction of five components.¹ All loading matrices were subsequently rotated to an optimal fit, with (a) the “simple structure” criterion (using Varimax rotation) and (b) an idealized 10-item 5-dimensional Big Five factor structure (with 1 and 0 loadings) using orthogonal target rotation (as advocated by Allik & McCrae, 2004; McCrae et al., 1996). We then assessed the similarity of the resulting factorial structures with the idealized structure using congruence coefficients c , as suggested by Lorenzo-Seva and ten Berge (2006; see also Rammstedt et al., 2013).²

To identify a level of congruence that is significantly greater than what one can expect for random configurations, Rammstedt et al., 2013; see also Lechner & Rammstedt, 2015) provided simulation norms. According to these a coefficient greater than .78 is the critical benchmark value of random congruence. This finding is also in line with Lorenzo-Seva and ten Berge (2006), stating that

¹ In contemporary Big Five research, especially using more comprehensive questionnaires, Exploratory Factor Analysis (EFA) is also used to investigate the dimensional structure. However, most prior studies using the BFI-10 used PCA partly due to the fact that the BFI-10, with two items per Big Five domain, does not fulfill the requirement for stable EFA solutions of having at least four, and ideally more, items per factor (e.g., MacCallum et al., 1999; Mundfrom et al., 2005).

² Within the Big Five context some researchers also use oblique rotation to account for the correlations among the Big Five factors. In the present context we chose to use orthogonal (Varimax) rotation, in order to (a) stay methodologically in line with the BFI-10 literature (e.g., Rammstedt & John, 2007), (b) to provide future researchers with respective benchmarks for their factorial results, and (c) to be able compare the resulting factorial solutions to an ideal-typical Big Five structure, which is traditionally and theoretically assumed to have five uncorrelated domains (e.g., John, 1990). For testing the match with the idea-typical structure, orthogonal rotation is a more conservative approach in that it does not allow adaptation to the idealized Big Five structure by permitting arbitrarily large factor intercorrelations.

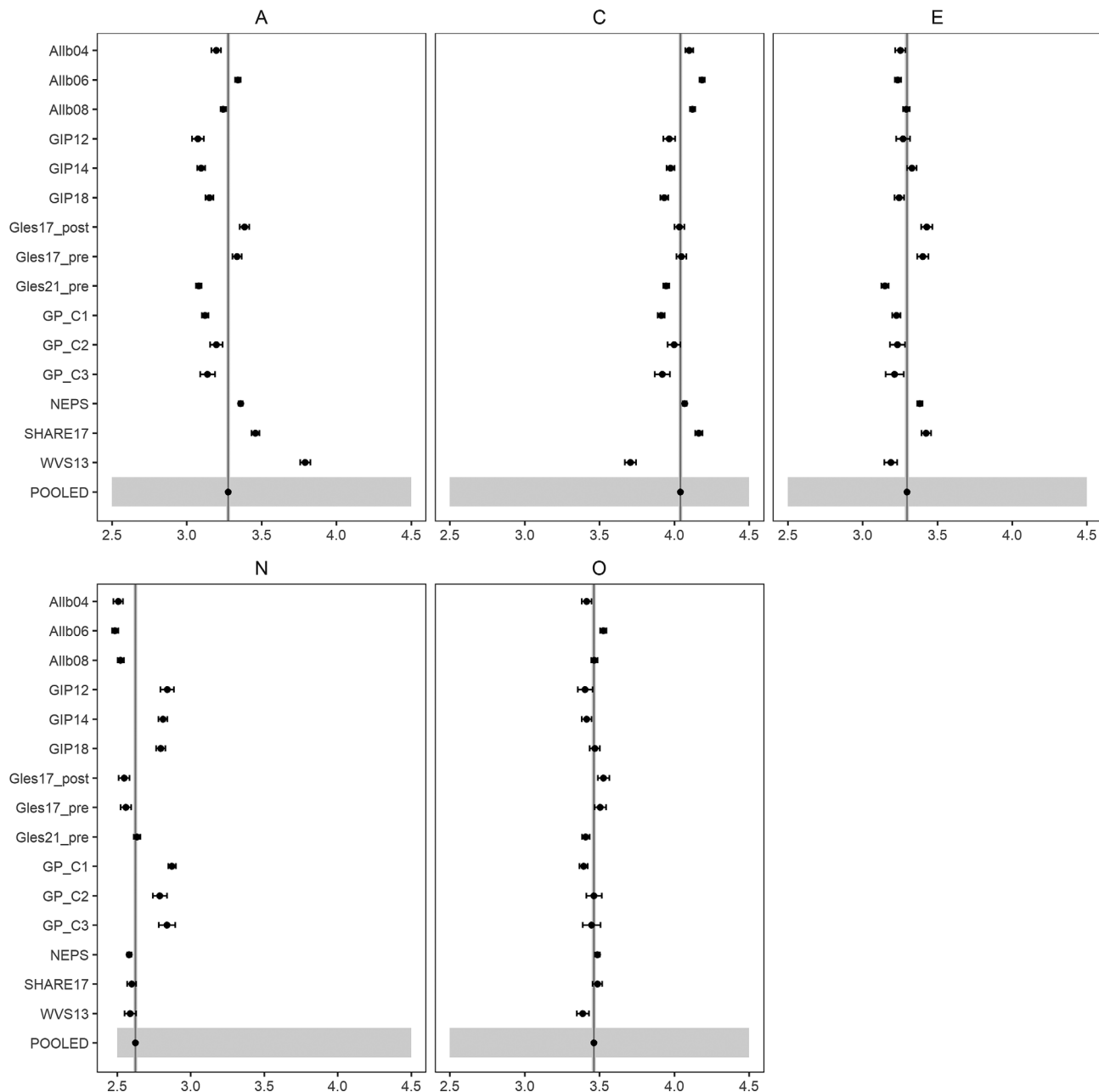


Figure 1. Overview of means and 95% CIs for the five BFI-10 scales across 15 samples and for the pooled sample. The (very narrow) shaded area around the vertical line denotes the 95% CI for the mean of the pooled sample. The 2009 Jena Study sample had a different scaling and is not shown here. A = Agreeableness; C = Conscientiousness; E = Extraversion; N = Neuroticism; O = Openness to Experience.

congruence of $> .78$ can be expected in less than 1% of the cases, and is therefore considered statistically significant.

Figure 4 and Table S7 in the supplemental material show the average primary and absolute secondary loadings by the domain (as well as across all domains) and by sample as well as the 95% CIs of these average loadings. Due to the orthogonal rotation, the loadings can be interpreted as correlations such that we used the standard error for aggregated correlations (e.g., Borenstein et al., 2009) to compute

the CIs of the aggregate loadings. In Figure 5 (and also in Table S8) the congruence coefficients are displayed by sample and by domain (as well as across all domains). Overall, the resulting picture for the factor structure was again very homogeneous across samples. In all studies mean primary loadings were at least $.37$ higher than the average secondary loadings (see Table S7). Compared to the pooled data, mean primary and secondary loadings were rather similar in size (see Table S9). Averaged across

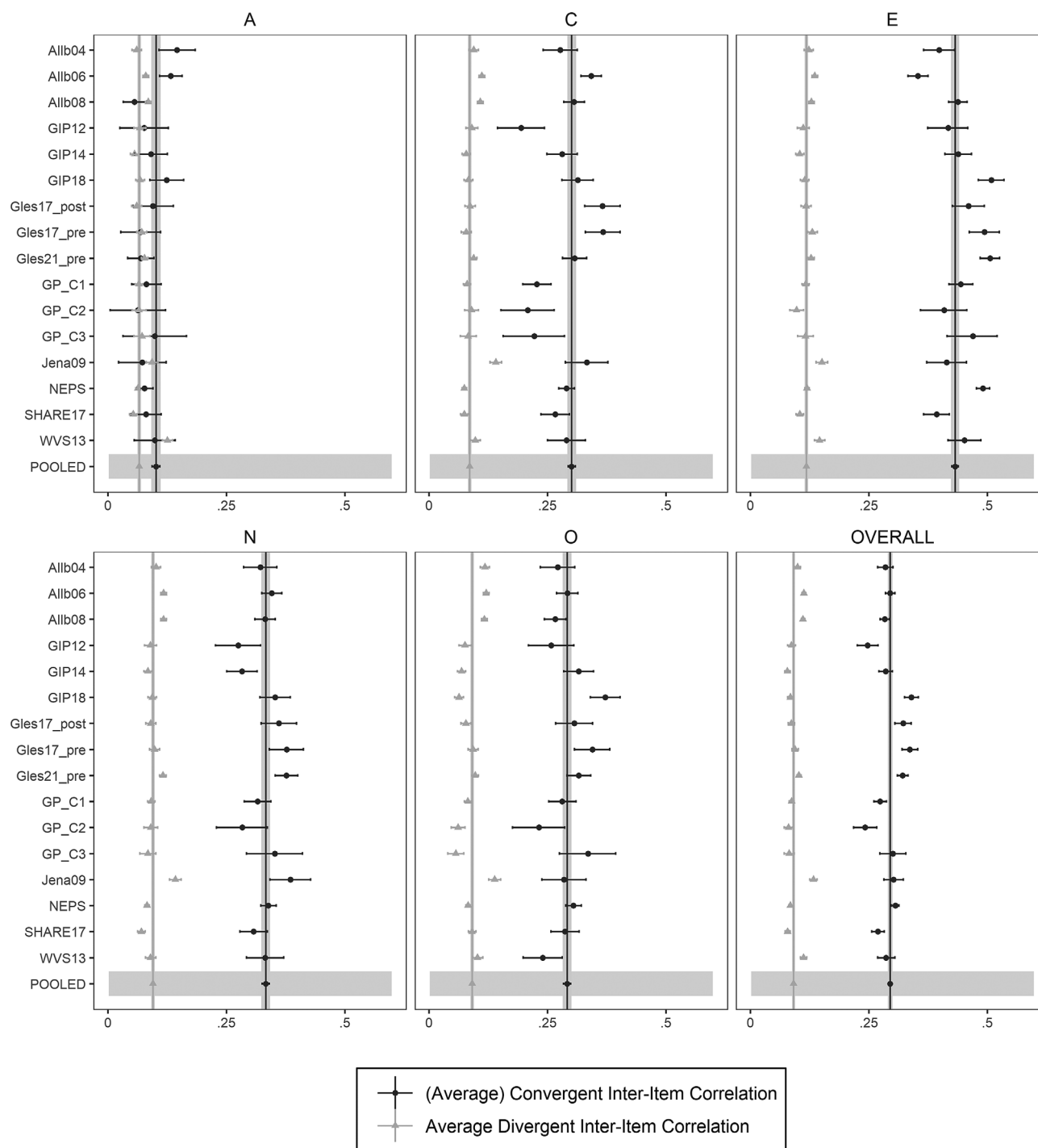


Figure 2. Convergent inter-item correlations (correlations of the two items of the same domain) and average divergent inter-item correlations (correlations with the eight items of the other domains) as well as their 95% CIs. For all 16 samples, these values are displayed separately by domain and averaged across all domains (panel “OVERALL”). The (very narrow) shaded area around the vertical line denotes the 95% CI for the respective correlations of the pooled sample. A = Agreeableness; C = Conscientiousness; E = Extraversion; N = Neuroticism; O = Openness to Experience.

studies, the absolute differences to the pooled loadings ranged from $\Delta_{\text{mean}} = .02\text{--}.11$. Largest deviations from the pooled loading structure were detected for the ALLBUS 2004 and ALLBUS 2006 with, for example, lower primary loadings on Extraversion of $\Delta = -.33$ and $\Delta = -.29$.

Moreover, congruence with an idealized factorial structure was highly similar, reaching .91 across all samples and domains. Thus, in all 16 samples averaged congruence coefficients can be regarded as statistically significant regarding the benchmark set by Lorenzo-Seva and ten

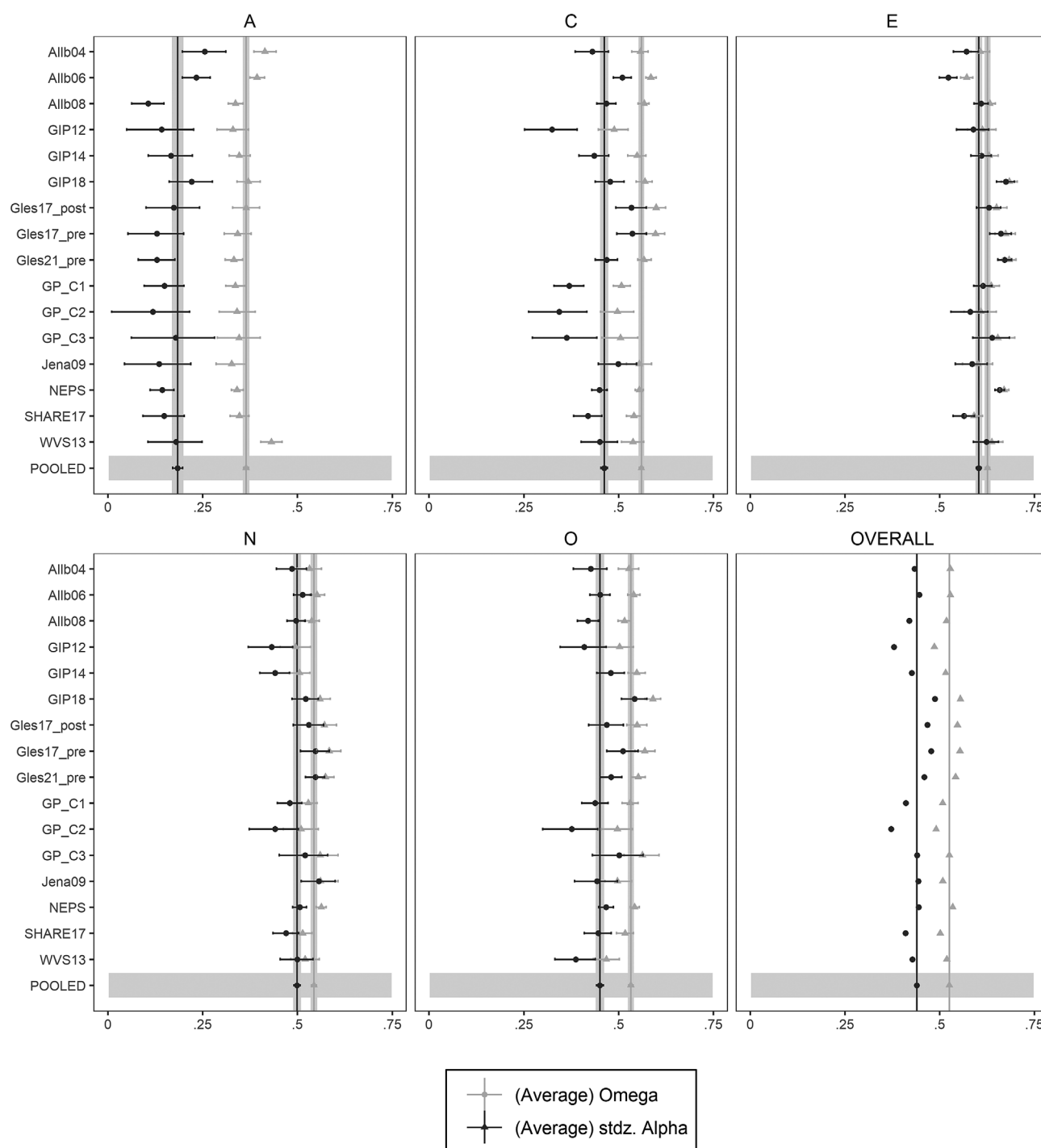


Figure 3. Standardized α s and McDonald's Ω s as well as their 95% CIs. For all 16 samples, these values are displayed separately by domain and averaged across all domains (panel "OVERALL"). The (very narrow) shaded area around the vertical line denotes the 95% CI for respective estimate of the pooled sample. A = Agreeableness; C = Conscientiousness; E = Extraversion; N = Neuroticism; O = Openness to Experience.

Berge (2006). However, some exceptions were observed within the individual domains: All three ALLBUS samples showed weaker congruence for at least one domain, and the GIP 2012 sample did not meet the congruence criterion for the domain Conscientiousness (see Table S8).

Discussion

Previous research based on WVS data from 25 countries indicated that the psychometric properties of the BFI-10 do not always meet psychometric quality standards with

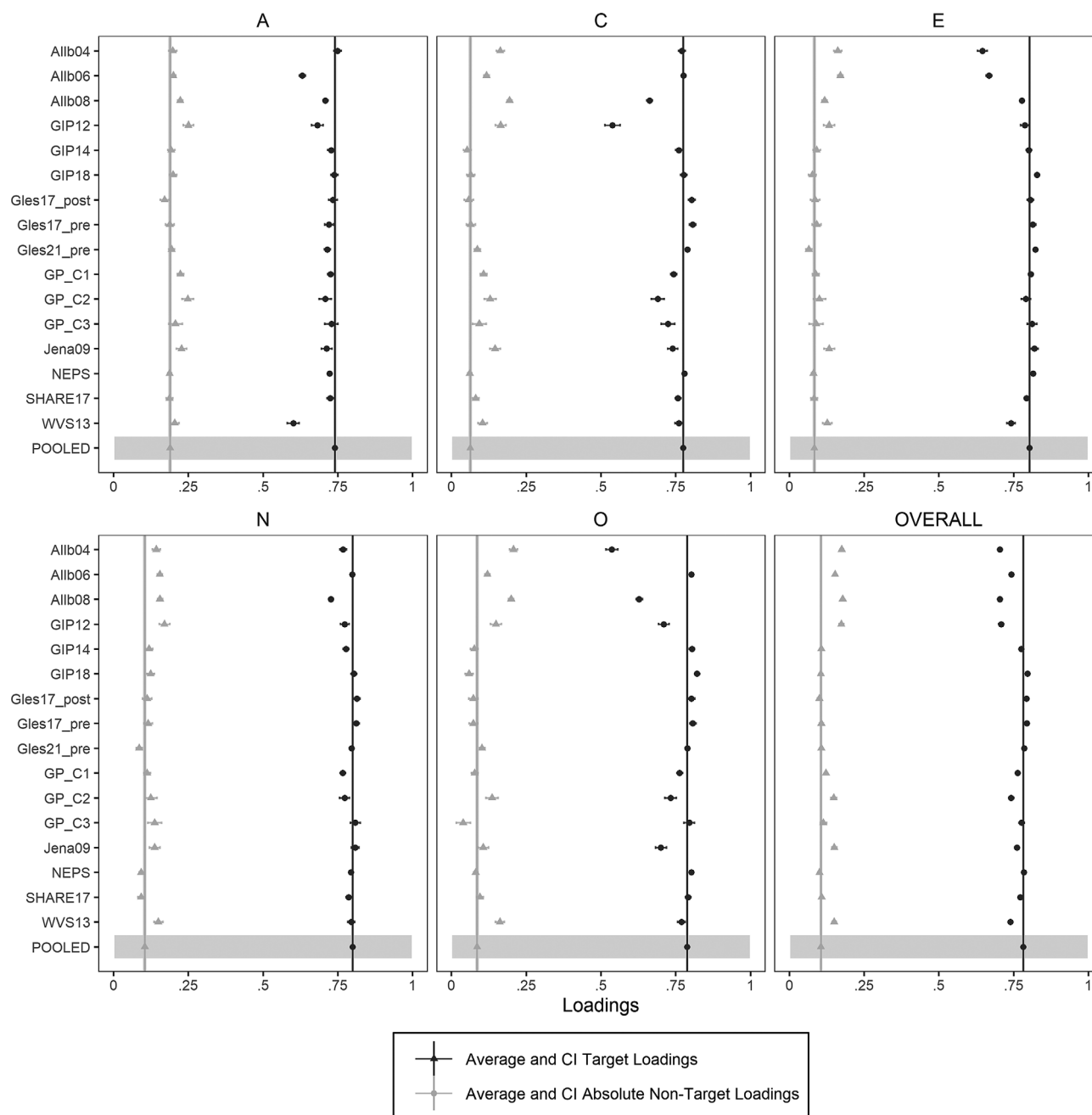


Figure 4. Average of the target loadings and absolute non-target loadings and their 95% CI by sample (separately by domain and averaged across domains) and averaged across domains (panel "OVERALL"). The (very narrow) shaded area around the vertical lines denotes the 95% CI for the respective average loadings of the pooled sample. A = Agreeableness; C = Conscientiousness; E = Extraversion; N = Neuroticism; O = Openness to Experience.

regard to its structural properties, especially its factorial validity, in international surveys (Ludeke & Larsen, 2017; see also Levinsky et al., 2019). The extent to which these issues reflect "natural" variation in the structural properties of the instrument (e.g., due to sampling variation), or factors such as language/translation and culture, is unclear because the extent to which the structural properties of the BFI-10 can be expected to vary even within the same

language and target population has not been thoroughly investigated to date.

To contribute to filling this research gap, we carried out a comprehensive evaluation of the stability and replicability of the BFI-10's structural properties by investigating the extent to which these properties vary across samples when keeping the target population and language constant (i.e., in the absence of any translation issues, cultural differences,

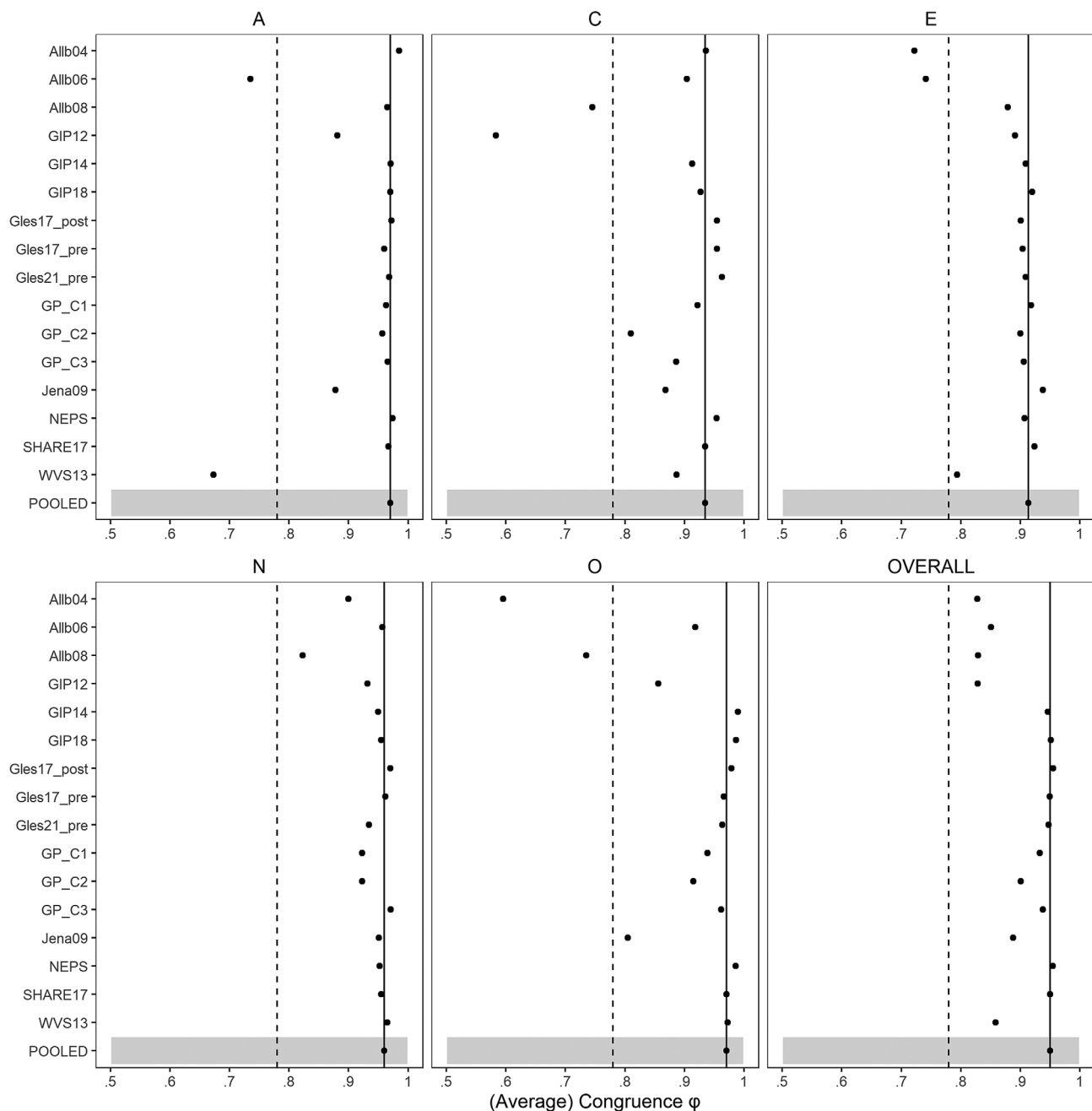


Figure 5. Congruence (Tucker's ϕ) of the component loadings with an idealized Big Five target loading structure by domain and across all domains, by sample. A = Agreeableness; C = Conscientiousness; E = Extraversion; N = Neuroticism; O = Openness to Experience.

sample selectivity, or other characteristics that may otherwise impact structural properties of the inventory). We examined several key structural properties of the BFI-10 – namely, intercorrelations, descriptive statistics, reliability, and factorial validity – across 16 representative large-scale samples of the German adult population with a total of approximately $N = 60,000$ respondents. Our results indicate that the BFI-10 shows quite comparable structural properties across the 16 samples. All investigated structural

features were largely homogeneous in size across Big Five domains and samples. Further, in nearly all cases, the factorial structure was sufficiently congruent with an idealized Big Five structure. This is all the more true considering that the test of congruency is rather strict – the target matrices contain only unit or zero loadings, whereas empirical loadings are rarely exactly zero or one and can never exceed one. Hence, ruling out biasing influences such as translation difficulties, cultural differences, or sample selectivity, the

BFI-10 provides a relatively stable and replicable measure of the Big Five domains.

For all indicators, some slight deviations in performance were identified across samples. As some samples differed in their implementation of the BFI-10, it seems plausible that the deviations found are often due to these differences: Not surprisingly, scale means in the WVS – in which alternative item formulations were used – differed from those in all other samples. For all ALLBUS samples, the factor structure for at least one Big Five domain was not congruent with the idealized factor structure. Here, the inverted response format might have had a negative impact on the structural validity of the items.

These two examples indicate that differences in the administration of the inventory – even when controlling more crucial aspects such as the target population or the assessment language – can affect the resulting psychometric quality.

Our focus in the present study was descriptive. We aimed to gauge the extent to which the psychometric properties of the BFI-10 vary, or are rather consistent, across large-scale surveys in Germany. An important next step would be to identify the potential sources of variation in the psychometric properties, such as survey mode, the presence or absence of interviewers, the position of the BFI-10 in the surveys, or respondent burden. Such analyses were beyond the scope of our present paper, and the present set of studies did not permit us to conduct such analyses because they showed little variation in such study characteristics that would be needed to point to sources of variation through integrative data analyses. Future research may gain additional insights by extending the range of surveys and clarifying whether variation in the psychometric properties is random (i.e., sampling variation) or can be traced back to study characteristics. From a general point of view, researchers using established inventories should be cautious when adapting them and should carefully consider every change to the standard application.

In sum, our study demonstrates that the structural properties of the BFI-10 are highly stable and replicable in German large-scale surveys. Keeping sample selection and administration language constant, the psychometric properties of the BFI-10 were highly comparable across numerous samples.

References

- Allik, J., & McCrae, R. R. (2004). Escapable conclusions: Toomela (2003) and the universality of trait structure. *Journal of Personality and Social Psychology*, 87(2), 261–265. <https://doi.org/10.1037/0022-3514.87.2.261>
- Bergmann, M., Kneip, T., De Luca, G., & Scherpenzeel, A. (2019). *Survey participation in the Survey of Health, Ageing and Retirement in Europe (SHARE), Wave 1-7* (SHARE Working Paper Series 41-2019). SHARE-ERIC. http://www.share-project.org/uploads/tx_sharepublications/WP_Series_41_2019_Bergmann_et_al.pdf
- Blom, A. G., Bossert, D., Funke, F., Gebhard, F., Holthausen, A., Krieger, U., & SFB 884 “Political Economy of Reforms” Universität Mannheim (2016). *German Internet Panel, Wave 1 – Core Study (September 2012)* (Version 2.0.0) [Dataset]. GESIS Data Archive. <https://doi.org/10.4232/1.12607>
- Blom, A. G., Bossert, D., Gebhard, F., Funke, F., Holthausen, A., Krieger, U., & SFB 884 “Political Economy of Reforms” Universität Mannheim (2016). *German Internet Panel, Wave 13 – Core Study (September 2014)* (Version 2.0.0) [Dataset]. GESIS Data Archive. <https://doi.org/10.4232/1.12619>
- Blom, A. G., Fikel, M., Friedel, S., Höhne, J. K., Krieger, U., Rettig, T., Wenz, A., & SFB 884 “Political Economy of Reforms” Universität Mannheim (2020). *German Internet Panel, Wave 37 – Core Study (September 2018)* (Version 2.0.0) [Dataset]. GESIS Data Archive. <https://doi.org/10.4232/1.13584>
- Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting up an online panel representative of the general population: The German Internet Panel. *Field Methods*, 27(4), 391–408. <https://doi.org/10.1177/1525822X15574494>
- Blossfeld, H.-P., & von Maurice, J. (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft*, 14(2), 19–34. <https://doi.org/10.1007/s11618-011-0179-2>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis* (1st ed.). Wiley. <https://doi.org/10.1002/9780470743386>
- Börsch-Supan, A. (2022). *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 7* (Version 8.0.0) [Dataset]. SHARE-ERIC. <https://doi.org/10.6103/SHARE.W7.800>
- Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbacher, J., Malter, F., Schaan, B., Stuck, S., & Zuber, S. (2013). Data resource profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, 42(4), 992–1001. <https://doi.org/10.1093/ije/dyt088>
- Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). Establishing an open probability-based mixed-mode panel of the general population in Germany: The GESIS Panel. *Social Science Computer Review*, 36(1), 103–115. <https://doi.org/10.1177/0894439317697949>
- Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, 14(2), 81–100. <https://doi.org/10.1037/a0015914>
- Eisinga, R., te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, 58(4), 637–642. <https://doi.org/10.1007/s00038-012-0416-3>
- Fischer, R., & Karl, J. A. (2021). *Niche diversity effects on personality measurement – Evidence from representative samples during the COVID-19 pandemic*. PsyArXiv. <https://doi.org/10.31234/osf.io/5jva9>
- GESIS. (2011a). *ALLBUS/GGSS 2004 (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 2004)* (ZA3762 Data file Version 2.0.0) [Dataset]. GESIS Data Archive. <https://doi.org/10.4232/1.10977>
- GESIS. (2011b). *ALLBUS/GGSS 2006 (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 2006)* (ZA4500 Data file Version 2.0.0) [Dataset]. GESIS Data Archive. <https://doi.org/10.4232/1.10832>
- GESIS. (2015). *ALLBUS/GGSS 2008 (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey 2008)* (ZA4600 Data file Version 2.1.0) [Dataset]. GESIS Data Archive. <https://doi.org/10.4232/1.12345>

- GESIS. (2022). *GESIS Panel – Standard edition* (Version 43.0.0) [Dataset]. GESIS Data Archive. <https://doi.org/10.4232/1.13880>
- GLES. (2019a). *Post-election cross-section (GLES 2017)* (Version 4.0.1) [Dataset]. GESIS Data Archive. <https://doi.org/10.4232/1.13235>
- GLES. (2019b). *Pre-election cross-section (GLES 2017)* (5.0.1) [Dataset]. GESIS Data Archive. <https://doi.org/10.4232/1.13234>
- GLES. (2022). *GLES pre-election cross-section 2021* (Version 2.0.0) [Dataset]. GESIS Data Archive. <https://doi.org/10.4232/1.13860>
- Ingelhart, R., Haerpfer, C. W., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., & Puranen, B. (2018). *World Values Survey Wave 6 (2010-2014)* (Version 20201117) [Dataset]. World Values Survey Association. <https://doi.org/10.14281/18241.8>
- John, O. P. (1990). The “Big Five” factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66–100). The Guilford Press.
- John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *Big Five Inventory (BFI)*. APA PsycTests. <https://doi.org/10.1037/t07550-000>
- 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). Guilford Press.
- Lechner, C. M., & Rammstedt, B. (2015). Cognitive ability, acquiescence, and the structure of personality in a sample of older adults. *Psychological Assessment*, 27(4), 1301–1311. <https://doi.org/10.1037/pas0000151>
- Levinsky, M., Litwin, H., & Lechner, C. M. (2019). Personality traits: The ten-item Big Five Inventory (BFI-10). In M. Bergmann, A. Scherpenzeel, & A. Börsch-Supan (Eds.), *SHARE Wave 7 Methodology: Panel innovations and life histories* (pp. 29–34). MEA, Max Planck Institute for Social Law and Psychology. http://www.share-project.org/fileadmin/pdf_documentation/MFRB_Wave7/SHARE_Methodenband_A4_WEB.pdf
- 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>
- Lu, Q., & Cui, S. (2022). Not completely unusable: Procedures to rescue the Big-Five personality data in the World Values Survey wave 6. *Personality and Individual Differences*, 199, Article 111832. <https://doi.org/10.1016/j.paid.2022.111832>
- 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>
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4(1), 84–99. <https://doi.org/10.1037/1082-989X.4.1.84>
- McCrae, R. R., & Costa, P. T. (2008). The five-factor theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 159–181). Guilford Press.
- McCrae, R. R., Zonderman, A. B., Costa, P. T., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology*, 70(3), 552–566. <https://doi.org/10.1037/0022-3514.70.3.552>
- Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, 5(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4
- NEPS Network. (2021). *National educational panel study, scientific use file of starting cohort adults* (12.1.0) [Dataset]. IfBi Leibniz Institute for Educational Trajectories. <https://doi.org/10.5157/NEPS:SC6:12.1.0>
- Rammstedt, B. (2007). *Welche Vorhersagekraft hat die individuelle Persönlichkeit für inhaltliche sozialwissenschaftliche Variablen?* [What is the predictive power of individual personality for substantive social science variables?] (ZUMA-Arbeitsbericht No. 2007/01). Zentrum für Umfragen, Methoden und Analysen (ZUMA). <https://nbn-resolving.org/urn:nbn:de:0168-ssaoar-200543>
- 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., Lechner, C., & Danner, D. (2021). Short forms do not fall short: A comparison of three (extra-)short forms of the Big Five. *European Journal of Psychological Assessment*, 37(1), 23–32. <https://doi.org/10.1027/1015-5759/a000574>
- Rammstedt, B., Mutz, M., & Farmer, R. F. (2015). The answer is blowing in the wind: Weather effects on personality ratings. *European Journal of Psychological Assessment*, 31(4), 287–293. <https://doi.org/10.1027/1015-5759/a000236>
- Rammstedt, B., Roemer, L., Mutschler, J., & Lechner, C. (2023). *The Big Five personality dimensions in large-scale surveys: An overview across 25 publicly available data sets for personality research*. Manuscript submitted for publication.
- Roemer, L., Lechner, C. M., & Rammstedt, B. (2023, April 13). *Data and supplementary materials for “Consistency of the structural properties of the BFI-10 across 16 samples from eight large-scale surveys in Germany”*. <https://osf.io/bup8s>
- Silbereisen, R. K., Pinquart, M., Reitzle, M., Tomasik, M. J., Fabel, K., & Grümer, S. (2008). *Psychosocial resources and coping with social change* (SFB-580 Reports, Volume 19). https://www.researchgate.net/publication/37367356_Psychosocial_Resources_and_Coping_with_Social_Change
- Thalmayer, A. G., Saucier, G., & Eigenhuis, A. (2011). Comparative validity of brief to medium-length Big Five and Big Six personality questionnaires. *Psychological Assessment*, 23(4), 995–1009. <https://doi.org/10.1037/a0024165>

History

Received September 9, 2022

Revision received January 25, 2023

Accepted February 1, 2023

Published online May 23, 2023

EJPA Section / Category Personality

Conflict of Interest

We have no known conflict of interest to disclose.

Open Science

We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses, including all tested models. If we used inferential tests, we report exact *p* values, effect sizes, and 95% confidence or credible intervals.

Open Data: All analyzed datasets are either openly accessible for research purposes (i.e., Jena Study: OSF; WVS: <https://www.worldvaluessurvey.org/WVSContents.jsp>; accessible subject to registration (i.e., ALLBUS: <https://www.gesis.org/allbus/download>; GLES: <https://www.gesis.org/en/elections-home/gles>; or accessible

subject to conclusion of a data use agreement (GESIS Panel: <https://www.gesis.org/en/gesis-panel/gesis-panel-home>; GIP: <https://www.uni-mannheim.de/en/gip/for-data-users/>; NEPS: <https://www.neps-data.de>; SHARE: <http://www.share-project.org/data-access.html>). Open Materials: We confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology. The analysis code and output are provided on the project's OSF page: <https://osf.io/bup8s/> (Roemer et al., 2023). Preregistration of Studies and Analysis Plans: This study was not preregistered.

Funding

This paper uses data from SHARE Wave 7 (DOI: 10.6103/SHARE.w7.800), see Börsch-Supan et al. (2013) for methodological details. The SHARE data collection has been funded by the European Commission, DG RTD through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N° 211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982, DASISH: GA N°283646) and Horizon 2020 (SHARE-DEV3: GA N° 676536, SHARE-COHESION: GA N°870628, SERISS: GA N°654221, SSHOC: GA N°823782, SHARE-COVID19: GA N°101015924) and by DG Employment, Social Affairs & Inclusion through VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332, and VS 2020/0313. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the US National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C, RAG052527A) and from various national funding sources is gratefully acknowledged (see <https://www.share-project.org>).


This paper uses data from Waves 1, 13, and 37 of the German Internet Panel (GIP; DOIs: 10.4232/1.12607, 10.4232/1.12619, 10.4232/1.13584; Blom, Bossert, Funke, et al., 2016; Blom, Bossert, Gebhard, et al., 2016; Blom et al., 2020). A study description can be found in Blom et al., 2015. The GIP is funded by the German Research Foundation (DFG) as part of the Collaborative Research Center 884 (SFB 884; project number 139943784; Project Z1).

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Adults, doi: 10.5157/NEPS:SC6:12.1.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Open access publication enabled by GESIS Leibniz Institut für Sozialwissenschaften, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 491156185.

ORCID

Beatrice Rammstedt

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

Lena Roemer

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

Beatrice Rammstedt

GESIS – Leibniz Institute for the Social Sciences

PO Box 12 21 55

68072 Mannheim

Germany

beatrice.rammstedt@gesis.org