

# Measurement Invariance and Sex and Age Differences of the Big Five Inventory–2: Evidence From the Russian Version

Assessment  
2020, Vol. 27(3) 472–486  
© The Author(s) 2019  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/1073191119860901  
journals.sagepub.com/home/asm



Sergei Shchebetenko<sup>1</sup>, Aleksey Y. Kalugin<sup>2</sup>,  
Arina M. Mishkevich<sup>3</sup>, Christopher J. Soto<sup>4</sup>, and Oliver P. John<sup>5</sup>

## Abstract

The Big Five Inventory–2 (BFI-2) is a recently published 60-item questionnaire that measures personality traits within the five-factor model framework. An important aspect of the BFI-2 is that it measures the traits at both the domain and facet levels and also controls acquiescence bias via the balanced number of true- and false-keyed items across the domains and facets. The current research evaluates factorial measurement invariance of a Russian version of the BFI-2 across sex and age within samples of 1,024 university students (Study 1) and 1,029 Internet users (Study 2). Across these samples, men scored lower on the domains of negative emotionality and agreeableness and slightly higher on extraversion. Sex differences were also obtained on various facets. In the Internet sample, age correlated modestly with several Big Five domains in accordance with the well-documented maturity principle. The newly developed Russian version of BFI-2 showed good reliability and validity across both samples. Moreover, random intercept exploratory factor analyses showed that the BFI-2 displayed a hierarchical five-domain-15-facet structure that demonstrated strict measurement invariance across sex and age.

## Keywords

personality assessment, measurement invariance, Big Five Inventory, sex differences, age

Adequate measurements of traits are of critical importance to contemporary personality research. In this regard, scientists are faced with at least two inherently conflicting tasks: on the one hand, they must provide reliable, in-depth measurement of personality traits while, on the other hand, they must ensure the measurement remains brief and cost-effective. The lack of shared vision on how to reconcile these problems generates diversity in the measures of personality traits. An optimal solution would be a balanced compromise such that the accuracy and depth of analysis are coupled with a relatively high cost-effectiveness. The Big Five Inventory (BFI; John, Donahue, & Kentle, 1991; John, Naumann, & Soto, 2008) was developed with this goal in mind. This 44-item questionnaire has consistently demonstrated a high degree of reliability and validity and is freely accessible for scientific and educational purposes. The multiple advantages of this test have made it one of the primary tools utilized by Big Five investigators over the past three decades.

## Big Five Inventory

Despite its popularity, the original BFI (BFI-1) possessed several shortcomings. A prominent issue identified in the

BFI-1 was that it addressed solely the domain level of traits. However, the hierarchical nature of personality remains a consensus among personality psychologists and is at the core of the idea of personality traits (Eysenck, 1947; Goldberg, 1993). This suggests that it is viable to measure not only broad trait domains like the Big Five but also more specific facets—personality substructures that constitute the domains. The BFI-1 was not developed to measure facets; therefore, while the authors later resolved to analyze the items of the BFI-1 on the facet level (Soto & John, 2009), the resulting, post hoc facet scales did not achieve the same

---

<sup>1</sup>National Research University Higher School of Economics, Moscow, Russia

<sup>2</sup>Perm State Humanitarian Pedagogical University, Perm, Russia

<sup>3</sup>Perm State University, Perm, Russia

<sup>4</sup>Colby College, Waterville, ME, USA

<sup>5</sup>University of California, Berkeley, CA, USA

## Corresponding Author:

Sergei Shchebetenko, Department of Psychology, National Research University Higher School of Economics, 20 ulitsa Myasnitskaya, 101000 Moscow, Russia.

Email: sshchebetenko@hse.ru

high standard of reliability and validity established by the BFI-1 domain scales. These limitations ultimately encouraged the authors to develop an improved version of the questionnaire.

The new BFI-2 (Soto & John, 2017b) provides facet-level personality assessment in a systematic yet cost-effective manner. Each BFI-2 domain (extraversion, agreeableness, conscientiousness, negative emotionality, and open-mindedness<sup>1</sup>) includes three facets (cf. Goldberg, 1993). Each facet is measured using four items, with two true-keyed and two false-keyed items per facet. The rigorous balance in the number of true- and false-keyed items across the domain scales allows for the control of acquiescence bias, the tendency of a respondent to consistently agree (yea-saying) or consistently disagree (nay-saying) with items, regardless of their content. Individual differences in acquiescence typically emerge in the use of unbalanced questionnaires like the BFI-1 (Rammstedt & Farmer, 2013; Soto, John, Gosling, & Potter, 2008). Regarding the BFI-2, the control of acquiescence bias considerably helps improve the five-factor structure of the questionnaire, while also providing a more precise evaluation of the relative contribution of traits to various criteria of interest (Soto & John, 2017b).

The original success of the BFI-1 led to an immediate surge in the development of non-English translations of the BFI-2 on its release. German (Danner et al., 2019) and Dutch (Denissen, Geenen, Soto, John, & van Aken, 2019) versions of the BFI-2 were recently published, and more than a dozen other translations are currently being developed. The BFI-2, including its non-English versions, has already seen widespread use in various research endeavors (e.g., Aichholzer, Danner, & Rammstedt, 2018; Margolis, Schwitzgebel, Ozer, & Lyubomirsky, 2018).

The BFI has long been involved in worldwide research on sex differences and age trends in personality. These studies, conducted in various cultural and linguistic settings, have reported sex differences such that women tend to score higher than men in negative emotionality, extraversion, agreeableness, and conscientiousness (Chiorri, Marsh, Ubbiali, & Donati, 2016; Schmitt, Realo, Voracek, & Allik, 2008; Srivastava, John, Gosling, & Potter, 2003). These findings closely correspond to those obtained with various alternative measures of the five-factor model (e.g., Costa, Terracciano, & McCrae, 2001; Kajonius & Johnson, 2018; Samuel, South, & Griffin, 2015).

Numerous studies using the BFI-1 have also addressed age-related variation in personality (Bleidorn et al., 2013; Soto, John, Gosling, & Potter, 2011; Specht, Egloff, & Schmukle, 2011; Srivastava et al., 2003). These studies have mostly demonstrated decreases in agreeableness and conscientiousness during the transition to adolescence, followed by increases in these domains and decreases in negative emotionality across adulthood. They have also revealed

that the related but distinguishable facets within each Big Five domain often show distinctive age trends.

## Measurement Invariance of the BFI-1

The issue of measurement invariance (MI; Meredith, 1993) has become a focus of researchers' attention in the past two decades. MI determines if the items used in a questionnaire mean the same things to members of different groups. If MI cannot be established, a between-group observed mean difference cannot be straightforwardly interpreted (Cheung & Rensvold, 2002). This raises difficulties for drawing conclusions about traditional observed mean differences in traits on various aspects including sex and age groups (Church et al., 2011; Suzuki et al., 2019). With respect to the BFI-1, multiple studies explicitly addressed MI across various groups such as sex (Chiorri et al., 2016), survey methods (Lang, John, Lüdtkke, Schupp, & Wagner, 2011), age moments (Nye, Allemand, Gosling, Potter, & Roberts, 2016; Specht et al., 2011), data collection methods (Feitosa, Joseph, & Newman, 2015), or geographic regions like countries or urban areas (Gebauer et al., 2014).

Chiorri et al. (2016) examined MI across sex employing an Italian version of the BFI-1 by means of exploratory structural equation modeling (ESEM), an approach designed to integrate the best features of confirmatory factor analysis (CFA) and exploratory factor analysis (EFA), two traditional tools for factor structure assessment. They found that the BFI-1 is strictly gender-invariant, while the pattern of latent scores reproduced those sex differences normally obtained in observed means. Nye et al. (2016), using the item response theory approach, demonstrated that the BFI-1 is invariant between individuals at age 20 and at age 50, such that the latter were more conscientious and less neurotic than the former. Specht et al. (2011), using latent change models, obtained strict invariance across personality traits measured with the BFI-1. They showed that a 4-year change in traits led to changes in the latent factors instead of changes in measurements.

With regard to the new BFI-2, to our knowledge, MI has been investigated only in a single cross-cultural study that compared groups of American English and German language origin (Rammstedt, Danner, Soto, & John, 2018). This study showed that German adaptations of two short forms of the BFI-2 were approximately invariant to the Anglo-American original (Soto & John, 2017a). Therefore, research on the MI of BFI-2 across sex and age groups is currently lacking but strongly required, given the growing popularity of this test.

## Russian Measures of Personality Traits

Another main objective of the present research was to develop a Russian version of the BFI-2. The development

of international questionnaires in the Russian language started in the 1980s as the Soviet state and society became increasingly open. Parts of this work were published and gained prominence to varying extents (e.g., Hanin, Eysenck, Eysenck, & Barrett, 1991; Slobodskaya, Knyazev, Safronova, & Wilson, 2003).<sup>2</sup> Numerous Russian versions of five-factor measures have appeared since the early 2000s (e.g., Kniazev, Mitrofanova, & Bocharov, 2010; Martin, Costa, Oryol, Rukavishnikov, & Senin, 2002) including a Russian version of the BFI-1 (Shchebetenko, 2014). All these Russian measures share the pros and cons of their international counterparts. They either lack brevity and cost-efficiency or measure solely the domain level of traits. In this context, a validated Russian version of BFI-2 would have substantial benefits for Russian language personality assessment.

## The Present Study

In sum, we conducted a series of studies to (a) develop and validate a Russian version of the BFI-2, and (b) examine MI and group differences of the BFI-2 across sex and age (see Shchebetenko, Kalugin, Mishkevich, Soto, & John, 2019, for the codes and data sets of this study). First, a preliminary, extended pool of candidate item translations for the Russian BFI-2 was developed. This pool was then assessed in a pilot study, and refined into an updated Russian BFI-2 (see supplementary materials for the full list of the Russian BFI-2 items available online). Next, in Study 1, we evaluated the Russian BFI-2's reliability, validity, and MI across sex in an independent student sample. Finally, in Study 2, we used an Internet sample to replicate these analyses and also evaluate the BFI-2's MI across age groups.

## Pilot Study

The primary goal of the pilot study was to develop an extended pool of Russian translations for the BFI-2 items and, furthermore, to preliminarily examine whether the best translations of the 60 BFI-2 items would demonstrate appropriate psychometric characteristics.

## Method

**BFI-2 Translation.** Four expert bilinguals fluent in both English and Russian, familiar with both cultures and with personality measurement, performed the initial translations of each English BFI-2 item, as well as the 12 original BFI-1 items that were not retained for the BFI-2, into Russian. Each translator first read through the entire set of English items, organized by domain and facet, to get a sense of each trait's overall meaning, as well as how the domains and facets differ from each other. The experts were asked to translate each facet scale's items as a group, keeping in mind the

facet's overall definition. They were instructed to avoid using the same trait-descriptive adjective or phrase in multiple items and to draw clear distinctions between the facets and domains. They were asked to focus on capturing the psychological meaning of each item and expressing this meaning in Russian language in a way that is clear and easy to understand. As a result, the translators occasionally produced various alternatives to the same item. This procedure gave an extended list of candidate Russian translations which was then given to another group of four expert bilinguals. The members of this latter group independently translated each Russian item back into English. Afterward, the original English items and back-translated English items were compared and discussed among the authors and experts during an online meeting.

This process resulted in an extended list of 101 Russian-translated BFI-2 candidate items that were subsequently administered in a student sample. This superset included the 60 best (as initially judged by the authors) conceptual translations of the BFI-2 items, alternative candidate translations for items that proved difficult to translate, and translations of the 12 BFI-1 items not included in the BFI-2.

**Participants and Measures.** The participants were 311 undergraduate students from a large Russian university aged between 18 and 51 years ( $M = 23.12$ ,  $SD = 5.26$ ); among them were 56 males (18%). All participants were Russian native speakers. Along with the BFI-2, we administered a number of additional measures not relevant to the present research (see supplementary material section for the entire list of questionnaires available online).

## Results

Following the procedure applied in the original BFI-2 psychometric study (Soto & John, 2017b), we first centered each individual's set of individual item responses around their within-person mean response, without reversing the false-keyed items. This procedure helps minimize the influence of acquiescence bias on individual item responses (cf. Rammstedt & Farmer, 2013; Soto et al., 2008; Soto & John, 2017b). We then estimated reliability and validity of the 60 best candidate Russian BFI-2 items using raw and centered scores.

**Internal Consistency.** The average Cronbach's alpha (see Table S1.1, supplementary materials available online, for details) across the Big Five domain scales was .82 for both the raw and centered items, with alphas ranging from .76 (agreeableness) to .86 (extraversion and conscientiousness) for the raw items, and from .77 (agreeableness) to .87 (extraversion) for the centered items. At the facet level, for the raw items, alpha reliabilities ranged from .46 to .82, averaging .66, whereas for the centered items alpha ranged

from .49 to .84, averaging .68. Among the 15 facet scales, two (trust and intellectual curiosity) demonstrated poor internal consistency (for the raw items, .46 and .47; for the centered items, .50 and .49, respectively). Overall, these results indicate good reliability for the domain scales and adequate reliability for the majority of the facet scales.

**Factorial Validity.** We conducted principal component analyses (PCA) separately using the raw and centered item scores. In both instances, we extracted and varimax-rotated five components across the 60 best candidate items (see Table S1.2, supplementary materials available online, for details). These results indicate that the within-person centering procedure adjusted the item structure to five components quite effectively, presumably by eliminating the effects of acquiescence variance on interitem correlations (cf. Soto & John, 2017b). In particular, the raw items had their strongest loading on an unintended component 12 times, in contrast to only four times for the centered items.

We then conducted a series of PCAs separately for each domain using the 12 domain items and tested whether they would properly distribute across three within-domain facets. Although an overall consistency to the intended three-facet structures within each domain was indeed revealed, several items, especially in the agreeableness and conscientiousness domains, produced blurred and uninterpretable structures that we noted for further investigation (see Tables S1.3-S1.7, supplementary materials available online).

## Discussion

In the pilot study, we developed an extended pool of Russian BFI-2 items with the help of eight bilingual experts. Afterward, we presented the Russian BFI-2 to a sample of university students. This generated a set of 60 candidate items that demonstrated adequate internal consistency and factorial validity. However, some items from this version remained problematic in that they had low loadings on their intended factors at both the domain and facet levels. Therefore, we next developed an updated pool of items to improve the psychometric quality of the Russian BFI-2.

## Study 1

In this study, we created a number of additional items to replace those items found to be problematic in the pilot study. Our main goal of Study 1 was to produce an improved version of the Russian BFI-2 and investigate its psychometric properties. Our second main goal was to test whether the BFI-2 would be measurement invariant across men and women.

## Method

**Development of New Items.** The first and second authors created an additional pool of 28 items while considering the psychological meaning of the respective facets and domains. The total pool of the Russian BFI-2 for Study 1 was thus composed of 98 items, including 70 candidate items taken from the pilot study and the 28 new items. This item pool was then presented to a new university student sample.

**Participants.** The sample composed of 1,024 undergraduates from two universities in a large region of Russia. They were aged from 17 to 44 years ( $M = 21.14$ ,  $SD = 4.54$ ); among them were 271 males (26.5%). The undergraduates took part in the study in exchange for partial course credit. To assess retest reliability, approximately 6 weeks after initially completing BFI-2, a subsample of 90 participants completed it a second time.

**Measures.** To assess the nomological network of the BFI-2, we also included a number of additional measures. We administered a 100-item Russian version (Kniazev et al., 2010) of the Big Five factor markers taken from the International Personality Item Pool (IPIP; Goldberg et al., 2006), which is a public-domain Big Five measure. The IPIP scales showed good internal consistency,  $\alpha_s = .93, .86, .91, .93, .86$ , for extraversion, agreeableness, conscientiousness, emotional stability, and intellect,<sup>3</sup> respectively. The items were anchored from 1 (*absolutely wrong*) to 5 (*absolutely right*). We expected that within-domain correlations assessed by the two Big Five measures would be strong, while the size of between-domain correlations would vary from negligible to moderate.

An 18-item Russian version (Shchebetenko, 2011) of the Need for Cognition Scale (Cacioppo, Petty, & Kao, 1984) was also administered. The participants rated each item on a 5-point scale ranging from 1 (*extremely uncharacteristic of me*) to 5 (*extremely characteristic of me*). According to extant literature (Madrid & Patterson, 2016; Sadowski & Cogburn, 1997), we expected need for cognition to show positive, moderate correlations with open-mindedness and conscientiousness.

A Russian version (Osin, 2012) of the 20-item Positive and Negative Affect Schedule (Watson, Clark, & Tellegen, 1988) was used to measure the extent of positive and negative affect that participants experienced during the previous few weeks. The participants rated each item on a 5-point scale ranging from 1 (*very slightly or not at all*) to 5 (*extremely*). The items were aggregated into two scales measuring positive and negative affect separately. We expected positive affect to correlate strongly with extraversion and moderately with the remaining Big Five domains, and expected negative affect to strongly correlate with negative emotionality (cf. Clark, Lelchook, & Taylor, 2010; Haslam, Whelan, & Bastian, 2009).

The participants completed a five-item Russian version (Osin & Leontiev, 2008) of the Satisfaction With Life Scale (Diener, Emmons, Larsen, & Griffin, 1985), using a 7-point rating scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). We expected this criterion to relate positively with extraversion and negatively with negative emotionality (cf. Pavot & Diener, 2008). Finally, the participants completed a Russian version (Shchebetenko, unpublished) of the 10-item Rosenberg Self-Esteem Scale (Rosenberg, 1965), which measured self-esteem using a 4-point rating scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). We expected self-esteem to relate positively with extraversion and conscientiousness, and negatively with negative emotionality (cf. Marshall, Lefringhausen, & Ferenczi, 2015; Robins, Tracy, Trzesniewski, Potter, & Gosling, 2001). All the criterion scales displayed good internal consistency,  $\alpha < .82$  (see Table S2.1, supplementary materials available online).

## Results

As a first step, we conducted analyses using the set of 60 best candidate BFI-2 items taken from the pilot study. Reliability and PCA validity analyses were performed to identify problematic items in this set (see Tables S2.2-S2.3, supplementary materials available online). Once such items were identified, they were replaced with more suitable candidate items retrieved from the extended item pool. Overall, we identified 12 problematic items in the initial version and replaced them (Table 1); by doing so, we preserved the overall psychological meaning of the replaced items and thus a revised version of the Russian BFI-2 was obtained. From this point forward, we discuss findings obtained with the revised BFI-2 version.<sup>4</sup>

### Reliability

**Internal consistency.** The average Cronbach's alpha across the five domain scales using raw and centered scores was good (.84 in both cases). Across 15 facet scales, the average alpha was .73 with the raw scores and .74 with the centered ones (see Table S2.2, supplementary materials available online).

**Test-retest assessment.** Test-retest reliability was assessed by Green's alpha coefficient which is less susceptible to effects of item recall than a regular Pearson correlation (Green, 2003). All five domain scales had Green's alpha of at least .72 ( $M = 0.76$ ) with both the raw and centered scores. For the 15 facet scales, Green's alpha ranged from .54 to .79 ( $M = 0.66$ ) for the raw scores and from .55 to .79 ( $M = 0.67$ ) for the centered scores. These results indicate adequate to strong retest reliability for the BFI-2 domain and facet scales (see Table S2.4, supplementary materials available online, for details).

### Factorial Validity

**PCA.** We conducted separate PCAs of the raw and centered item responses to the initial and revised versions of the Russian BFI-2. Variance explained by the first five components increased from the initial set of raw scores (41.30%) to the revised set of centered scores (44.27%). A scree test of the initial set of raw items supported a six-component solution, whereas scree tests of both sets of centered items supported a five-component solution (see Figure S2.1, supplementary materials available online). The absolute primary loadings ranged from .20 or stronger ( $M = 0.55$ ) within the initial set of raw item scores to .37 or stronger ( $M = 0.57$ ) within the revised set of centered item scores. The number of items that loaded most strongly on an unintended component decreased from three in the initial item sets to one in the revised sets (see Table S2.3, supplementary materials available online).

To assess the facet-level structure across the domains, we conducted a series of PCAs with three-component solution for the 12 within-domain items. The three-facet consistency at the item level increased from occasional within the initial set of raw items to theoretically consistent within the revised set of centered items (see Tables S2.5-S2.9, supplementary materials, available online, for details). In particular, there were 11 items in the former case that loaded most strongly on an unintended component, whereas there was just one such item in the latter case. The average absolute primary loading was .59 in the initial set of raw items, and .70 in the revised set of centered items.

**CFA and RI-EFA.** Since the limitations of exploratory PCA have been widely discussed in the literature (e.g., Marsh et al., 2009), to further examine the factorial structure of BFI-2, we subsequently conducted CFAs. We thereby assessed three nested models using the centered scores of BFI-2 and conducting analyses with maximum likelihood estimation. The first, *five orthogonal domains* model allowed 12 items within a particular domain to load on that domain's latent factor. Covariations between the latent factors were set to zero. The second, *five correlated domains* model was similar to Model 1 but the domain factors were allowed to correlate between each other. Finally, the third, *domain and facet* model was similar to Model 2, but item uniquenesses were allowed to correlate within a facet (i.e., between the items belonging to the same facet), thus accounting for the BFI-2's intended facet-level structure (Marsh, Morin, Parker, & Kaur, 2014).

A crucial limitation of applying CFA to personality data is that it requires cross-loadings of the items to be set to zero. Both adjective and questionnaire personality data show complex rather than simple structure, thereby constraining the ability of CFA to provide support for the five-factor model regardless of the specific measure used (e.g., Chiorri et al., 2016). One increasingly common way to

**Table 1.** The Replaced and Replacing Items Employed in the Revised Version of the Russian BFI-2.

Replaced items				Replacing items		
<i>n</i>	Facet/domain	Russian text	English text	<i>n</i>	Russian text	English text
8	Productiveness/C	Склонен к лени.	Tends to be lazy.	115	Предпочитает отдыхать, а не работать.	Prefers to have a rest, not to work.
10	Intellectual curiosity/O	Имеет много разных интересов.	Is curious about many different things.	93	Интеллектуал.	An intellectual.
11	Energy/E	Редко испытывает воодушевление и радостное волнение.	Rarely feels excited or eager.	117	Часто чувствует себя уставшим.	Often feels tired.
19	Anxiety/N	Может быть напряженным.	Can be tense.	89	Нервничает по любому поводу.	Nervous for any reason.
22	Respectfulness/A	Вступает в споры с другими людьми.	Starts arguments with others.	103	Бестактный в общении.	Tactless in social situations.
24	Depression/N	Чувствует себя защищенным, ему комфортно с собой.	Feels secure, comfortable with self.	87	Гармоничен и доволен жизнью.	Harmonious and pleased with life.
26	Energy/E	Физически менее активен, чем другие.	Is less active than other people.	110	Пассивный, вялый.	Passive, sluggish.
27	Trust/A	Умеет прощать, отходчивый.	Has a forgiving nature.	75	В целом, доверяет другим людям.	Generally trusts others.
28	Responsibility/C	Бывает легкомысленным.	Can be somewhat careless.	82	Нарушает обязательства.	Violates commitments.
30	Creative/O	Не очень творческий.	Has little creativity.	120	Мыслит шаблонно, стереотипно.	Thinks clichés, stereotypically.
45	Creative/O	Бывает трудно что-то вообразить, представить.	Has difficulty imagining things.	116	Неизобретательный.	Unresourceful.
47	Compassion/A	Может быть холодным и равнодушным.	Can be cold and uncaring.	125	Помогает только если ему выгодно.	Helps only if has benefits from it.

Note. BFI-2 = Big Five Inventory-2. BFI-2 items copyright 2015 by Oliver P. John and Christopher J. Soto. Reprinted with permission.

mitigate this problem is to employ ESEM (Asparouhov & Muthén, 2009), which integrates the key features of CFA and EFA. Since we also intended to account for acquiescence bias, we employed random intercept EFA (RI-EFA; Aichholzer, 2014) which is a special case of ESEM that explicitly estimates individual differences in acquiescence as a random variable. We thus reassessed Models 2 and 3 via RI-EFA, conducting analyses with Quartimin rotation using maximum likelihood with robust standard errors estimation. The covariations between the latent acquiescence factor and each of the Big Five domains were set to zero. Since the acquiescence variable was explicitly estimated with RI-EFA, the raw item scores instead of the centered scores were used.

With the use of CFA, Models 1 to 3 all provided poor fit according to chi-square, comparative fit index (CFI), and Tucker-Lewis index (TLI), although root mean square error of approximation (RMSEA) was acceptable for Model 3 (Table 2). With the use of RI-EFA, Model 2 also did not fit the data according to chi-square, CFI, and TLI. However, Model 3 yielded a considerable increment in fit and overall

provided good fit according to CFI, TLI, and RMSEA. In particular, CFI changed by .10, TLI by .11, and RMSEA by .02, which provides strong evidence for a difference between the models. The difference in chi-square between the RI-EFA models was also significant,  $\Delta\chi^2(90) = 2,091.23, p < .001$ , while the Bayesian information criterion (BIC) values greatly differed by 1,755. Thus, across 60 items of the Russian BFI-2, a model with five correlated domains and 15 facets provided a very good fit to the data while using RI-EFA (for factor loadings and factor correlations, see Table S2.10, supplementary materials available online). Notably, these findings show that the BFI-2 facets substantially contributed to item variance beyond the domains.

**Measurement Invariance Across Sex.** Building on the structural validity analyses described above, we tested MI with Model 3 (five correlated domains plus facets) using RI-EFA. As is common for testing MI (Meredith, 1993), we applied increasingly stringent equality constraints on the measurement parameters of the model between male and female participant

**Table 2.** Summary of Goodness-of-Fit Statistics for CFA and RI-EFA Models (Study 1).

Model and description	$\chi^2$	<i>df</i>	CFI	TLI	RMSEA	BIC
<i>CFA</i>						
1a. Five orthogonal domains	12,055.23	1,710	.597	.583	.077	12,887
2a. Five correlated domains	11,206.03	1,700	.630	.614	.074	12,107
3a. Five correlated domains plus facets	7,302.75	1,612	.778	.757	.059	8,814
<i>RI-EFA</i>						
2b. Five correlated domains	4,424.94	1,479	.856	.828	.044	168,832
3b. Five correlated domains plus facets	2,333.71	1,389	.954	.941	.026	167,077

Note. *N* = 1,024. *df* = degrees of freedom; CFI = confirmatory fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; BIC = Bayesian information criterion; CFA = confirmatory factor analysis; RI-EFA = random intercept exploratory factor analysis.

**Table 3.** Fit Statistics at Each Sex Measurement Invariance Level and Comparisons of the Indices Between Levels (Study 1).

Sex invariance models	$\chi^2$	<i>df</i>	RMSEA	CFI	TLI	SRMR	BIC	$\Delta\chi^2$	$\Delta df$	<i>p</i>	$\Delta CFI$
1. Configural	4,214.75	2,778	.032	.93	.91	.03	169,335				
2. Metric	4,502.22	3,053	.030	.93	.92	.04	167,831	309.10	275	.077	+.001
3. Strong	4,625.20	3,107	.031	.93	.92	.04	167,586	125.02	54	.000	-.003
4. Strict	4,818.12	3,167	.032	.92	.91	.05	167,381	190.97	60	.000	-.007

Note. *df* = degrees of freedom; CFI = confirmatory fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; BIC = Bayesian information criterion.

groups. The four levels of invariance tested, from least to most strict, were configural, metric, strong, and strict. With ESEM, configural invariance specifies the same items and number of factors for each group, while cross-loadings are allowed to vary for all items on all factors. Factor means were set to zero and factor variances to 1.0 in both men and women, and the RI factor was freely estimated in both groups. Next, metric invariance assumes configural invariance and adds a constraint of invariant factor loadings across groups. Factor means were set to zero in both groups and factor variances were set to 1.0 in men and freely estimated in women. Strong invariance adds to metric invariance a constraint of invariant intercepts across groups. Finally, strict invariance adds to strong invariance a constraint of invariant residual variances across groups. Decreases in CFI of more than .01, in RMSEA of more than .015, and in BIC of less than 2 provide evidence of measurement variance at a given level (e.g., Morin, Arens, & Marsh, 2016).

As shown in Table 3, the configural invariance model produced adequate model fit statistics in terms of RMSEA, CFI, TLI, and standardized root mean square residual (SRMR). These findings support a baseline of configural invariance across women and men. The test of metric invariance resulted in fit compatible with that obtained for configural invariance, and even better for RMSEA, CFI, and TLI. These results, including substantial decrease in BIC (1,504), support factor loading invariance across sex. The test of strong invariance resulted in roughly the same scores for RMSEA, TLI, SRMR, and CFI, which decreased only by

.003. These findings, along with BIC decrease of 245, support intercept invariance across men and women. Finally, the test of strict invariance resulted in a decrease in CFI of only .007, which, along with the large BIC decrease of 205, supports residual variance invariance across sex.

These findings clearly indicate that different mean scores on the domains and facets would characterize true differences in personality characteristics, rather than measurement artifacts. Thus, the sex-specific means presented in Table S2.13 (supplementary materials available online) indicate that women were substantially more negatively emotional ( $d = 0.80$ ), moderately more agreeable ( $d = 0.51$ ), and somewhat less extraverted ( $d = -0.26$ ) than men. At the facet-level, women had higher scores in all negative emotionality facets, especially anxiety ( $d = 0.87$ ). Women also had higher scores in all agreeableness facets, especially compassion ( $d = 0.58$ ). Regarding extraversion, it was only assertiveness ( $d = -0.44$ ) in which sex differences were observed. Despite negligible difference in open-mindedness domain ( $d = -0.07$ ), men and women substantially differed in its facets. Particularly, women were higher in aesthetic sensitivity ( $d = 0.47$ ) and lower in creative imagination ( $d = -0.33$ ) and intellectual curiosity ( $d = -0.41$ ) than men. Sex differences were not statistically significant for the conscientiousness domain and its facets.

#### Convergent and Discriminant Validity

*IPIP domain-level associations.* The BFI-2 domain scales correlated on average with the respective IPIP scales at .80

(Table S2.11, supplementary materials available online), while the average absolute correlation between different domains across these two instruments was .24. These results indicate strong convergent and discriminant validity of the Russian BFI-2.

**Nomological network.** We next examined the associations of the BFI-2 domain and facet scales with measures of other psychological constructs (see Table S2.12, supplementary materials available online, for details). As expected, open-mindedness correlated strongly ( $r = .60$ ) and conscientiousness moderately (.22) with need for cognition; unexpectedly, extraversion also correlated positively with need for cognition (.33). Positive affect during recent weeks correlated most strongly with extraversion (.56), and negative affect most strongly with negative emotionality (.62). Satisfaction with life showed the expected relations with extraversion (.37) and negative emotionality (−.30), and also correlated moderately with agreeableness (.27) and conscientiousness (.33). Finally, self-esteem strongly correlated with extraversion (.53), negative emotionality (−.45), and conscientiousness (.42). These findings largely correspond to our initial predictions based on the extant literature (see Method section above).

## Discussion

Study 1 demonstrates that the Russian BFI-2 has adequate reliability and validity, and that a revised version of the questionnaire obtained in this study outperforms the initial version developed from the pilot study in various important aspects. Twelve problematic items identified in the initial version were effectively replaced with new items while preserving the psychological meaning of the original items. As a result, the critical problem plaguing the initial version—low factor validity of the domain and facet scales—was successfully addressed.

The hierarchical domain and facet structure of the Russian BFI-2 version provided a very good fit to the data, which was shown with RI-EFA, a version of ESEM that accounts for acquiescence bias. Internal consistency and test–retest coefficients supported high reliability of the domain scales and adequate reliability of the facet scales. The nomological network of associations between the BFI-2 and various self-report criterion scales closely corresponded to previous findings and our expectations.

Building on the RI-EFA findings, we obtained support for the full range of MI across men and women, from weak configural to strict invariance. Our MI results indicate that the way in which men and women rated the BFI-2 items was close to identical in various structural aspects. This allows for the direct comparison of observed means across sex without risk of artificial differences. The pattern of mean sex differences across the domains closely corresponded to that

usually obtained with the BFI-1 (Chiorri et al., 2016; Schmitt et al., 2008; Srivastava et al., 2003). This result further demonstrates the validity of the Russian version of the BFI-2 in particular, and the BFI-2 in general. At the facet level, results indicated that different facet traits within the same Big Five domain sometimes showed distinctive gender differences. In particular, women differed from men on all facets of negative emotionality and agreeableness, while sex differences in extraversion reflected only a single facet, assertiveness. In the case of open-mindedness, trivial gender differences at the domain level concealed substantial differences at the facet level, with men scoring higher in intellectual curiosity and creative imagination, but women scoring higher in aesthetic sensitivity.

In sum, Study 1 is the first to show strict MI for the BFI-2 across sex. Moreover, the Russian BFI-2 demonstrated strong psychometrics, and its revised version outperformed the initial version. However, two clear limitations of this study were that it employed a university student sample, and that the advantages of the revised version were obtained post hoc. Also, the low variance in age typical for student samples precluded us from testing MI across age groups. With this in mind, we conducted a second study in which we employed a nonstudent sample with a wider age range.

## Study 2

The purpose of Study 2 was twofold. First, we aimed to examine if the BFI-2 maintains MI not only across sex but also across age groups. Second, we aimed to explicitly test whether the revised Russian BFI-2 maintains its psychometrics beyond the student population. To this end, we analyzed data from an Internet sample collected via an online questionnaire service.

## Method

### Participants

**Internet sample.** Participants were visitors to a website that asked them to complete the extended pool of Russian BFI-2 items in exchange for feedback. Specifically, their responses to the questionnaire would allow them to identify a character from the “Game of Thrones” television series (HBO Entertainment, 2012) that they resembled. Similar pop culture framings of personality tests have been used in previous research, and have produced high data quality, likely because respondents are intrinsically motivated to receive accurate feedback (e.g., Srivastava et al., 2003).

To develop this feedback, the first three authors first made a list of 36 main characters from the “Game of Thrones” series. These authors then generated a Big Five profile for each character by rating them on the 15 BFI-2 facets. Participants in Study 2 completed an extended version of the

**Table 4.** Summary of Goodness-of-Fit Statistics for CFA and RI-EFA Models (Study 2).

Model and description	$\chi^2$	<i>df</i>	CFI	TLI	RMSEA	BIC
<i>CFA</i>						
Five orthogonal domains	12,732.30	1,710	.590	.576	.079	13,565
Five correlated domains	11,934.69	1,700	.619	.604	.077	12,836
Five correlated domains plus facets	6,958.83	1,612	.801	.782	.057	7,422
<i>RI-EFA</i>						
Five correlated domains	5,217.81	1,479	.829	.795	.050	179,729
Five correlated domains plus facets	2,212.98	1,389	.962	.952	.024	177,009

Note. *N* = 1,029; *df* = degrees of freedom; CFI = confirmatory fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; BIC = Bayesian information criterion; CFA = confirmatory factor analysis; RI-EFA = random intercept exploratory factor analysis.

Russian BFI-2, and then were given automatically generated feedback about which Game of Thrones character their Big Five profile matched most closely. Each user received the feedback report immediately following completion of the questionnaire. The applied minimum threshold to stop data collection was to obtain valid responses from 1,000 participants. As of the end date, the sample comprised 1,029 valid cases. Two users indicated unrealistic ages of 0 and 827 years and these data were treated as missing. The remaining users reported ages ranging from 12 to 69 years ( $M = 27.23$ ;  $SD = 8.28$ ). Among the users were 312 males (30.3%), with one user failing to indicate her or his gender. The users indicated that they resided in various regions of Russia, although approximately 60% indicated residence within the Perm region. This likely reflects the fact that the first three authors, who resided in Perm, invited potential visitors using their online social network profiles.

**Measure.** Participants completed the same extended 98-item Russian BFI-2 pool that was administered in Study 1.

## Results

**Internal Consistency and Correlations.** The average alpha across the five raw and centered domain scales was good (.84 and .85, respectively). Across the facet scales, the average alpha was also slightly higher for the centered scores (.76) than for the raw scores (.75). Absolute correlations averaged .24 between the domain scales; absolute facet intercorrelations averaged .47 between pairs of same-domain facets, as compared with only .17 between different-domain facets (see Table S3.2, supplementary materials available online, for details).

### Factorial Validity

**PCA.** For the sake of comparison, we again estimated PCAs across the initial and revised versions of the questionnaire. Variance explained by the first five components increased steadily from the initial set of raw item scores (41.14%) to the revised set of centered item scores (44.51%). A scree test of the initial set of raw scores sup-

ported a six-component solution, whereas scree test of both sets of centered scores, along with the revised set of raw scores, supported a five-component solution (Figure S3.1, supplementary materials available online). The absolute primary loadings exceeded .10 ( $M = 0.55$ ) within the initial set of raw items and .32 ( $M = 0.58$ ) within the revised set of centered items. The number of items that loaded most strongly on an unintended component decreased from three in the initial sets of items to one in the revised sets of items. The average absolute secondary loadings dropped from .25 in the initial set of raw items to .23 in the revised set of centered items.

We also conducted a series of PCAs with three-component solutions for 12 within-domain items separately for each domain. In the initial set of raw items, there were 19 items that loaded most strongly on an unintended component. In the revised set of centered items there were no such items. The average absolute primary loading was .57 in the initial set of raw items, and .71 in the revised set of centered items (see Tables S3.4–S3.8, supplementary materials available online, for details).

**CFA and RI-EFA.** We examined factorial validity of the BFI-2 using the same model scheme as in Study 1. Again, the models were estimated, first, via CFA using the centered scores, and then by means of RI-EFA using the raw scores.

In general, the results of CFA and RI-EFA were similar to those obtained in Study 1 (Table 4). Again, with the use of CFA, all three models demonstrated poor fit indices with the exception of RMSEA. With the use of RI-EFA, Model 2 outperformed even CFA Model 3, although, again, it was only RMSEA that showed an acceptable fit. Finally, RI-EFA Model 3 revealed a strong increase in fit as compared with RI-EFA Model 2, and overall showed very good fit to the data (for factor loadings and factor correlations of Model 3 using CFA and RI-EFA, see Table S3.9, supplementary materials available online). In other words, the theoretically based five-domain-15-facet-model was strongly supported by the Internet sample data using the 60 items of the Russian BFI-2.

**Table 5.** Fit Statistics at Each Sex Measurement Invariance Level and Comparisons of the Indices Between Levels (Study 2).

Sex invariance models	$\chi^2$	df	RMSEA	CFI	TLI	SRMR	BIC	$\Delta\chi^2$	$\Delta df$	p	$\Delta CFI$
1. Configural	3,956.20	2,778	.029	.95	.93	.03	179,270				
2. Metric	4,245.62	3,053	.028	.95	.94	.04	177,724	300.43	275	.139	-.001
3. Strong	4,441.73	3,107	.029	.94	.93	.04	177,559	198.03	54	.000	-.006
4. Strict	4,594.95	3,167	.030	.94	.93	.05	177,457	155.93	60	.000	-.004

Note. df = degrees of freedom; CFI = confirmatory fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; BIC = Bayesian information criterion.

**Table 6.** Fit Statistics at Each Age Group Measurement Invariance Level and Comparisons of the Indices Between Levels (Study 2).

Age invariance models	$\chi^2$	df	RMSEA	CFI	TLI	SRMR	BIC	$\Delta\chi^2$	$\Delta df$	p	$\Delta CFI$
1. Configural	3,930.63	2,778	.028	.95	.93	.03	179,331				
2. Metric	4,207.22	3,053	.027	.95	.94	.04	177,764	287.09	275	.296	.000
3. Strong	4,365.72	3,107	.028	.94	.94	.04	177,554	165.89	54	.000	-.005
4. Strict	4,491.80	3,167	.029	.94	.93	.04	177,266	129.58	60	.000	-.002

Note. df = degrees of freedom; CFI = confirmatory fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; BIC = Bayesian information criterion.

### Measurement Invariance

**Sex.** First, we aimed to replicate our findings on sex invariance revealed in Study 1. We conducted the same MI analysis; specifically, four levels of invariance for RI-EFA Model 3 were tested (Table 5).

Again, the configural invariance model produced adequate, and even better than in Study 1, fit statistics in terms of RMSEA, CFI, TLI, and SRMR. The test of metric invariance resulted in no worse, and at times even better, estimates for RMSEA, CFI, TLI, and SRMR. Decrease in chi-square statistics was insignificant while decrease in BIC (1,547) was large. The test of strong invariance again resulted in roughly the same, as compared with the metric invariance test, estimates of RMSEA, TLI, SRMR, and CFI which decreased only by .006. These findings, along with BIC decrease of 165, support intercept invariance across sexes. Finally, the test of strict invariance resulted in only a trivial decrease in CFI (of .004), TLI, and a trivial increase in RMSEA (of .001). Together, these findings successfully replicate Study 1 by supporting strict invariance in the hierarchical structure of BFI-2 across men and women.

We could therefore examine mean-level sex differences in personality traits using the observed mean differences (Table S3.10, supplementary materials available online). Across the domains, women were substantially more negatively emotional ( $d = 0.63$ ), moderately more agreeable ( $d = 0.30$ ) and slightly less extraverted ( $d = -0.19$ ) than men. Sex differences in conscientiousness ( $d = -0.05$ ) and open-mindedness ( $d = -0.03$ ) were again negligible. At the facet level, women were uniformly more negatively emotional (all  $ds > 0.33$ ) and somewhat more agreeable (all  $ds > 0.16$ ) across all facets. Regarding extraversion,

men were relatively higher in assertiveness ( $d = -0.28$ ) and in energy level ( $d = -0.23$ ), while across the open-mindedness facets women were moderately higher in aesthetic sensitivity ( $d = 0.47$ ) but somewhat lower in intellectual curiosity ( $d = -0.28$ ) and creative imagination ( $d = -0.36$ ). Generally, this pattern closely corresponds to that revealed previously with the student sample (column-vector  $r = .96$ ).

**Age.** As this Internet-based sample substantially varied in age of participants, we also estimated MI across two age groups. To this end, we divided the sample into two subgroups: those who were aged 25 years or younger ( $n = 516$ ) and those who were aged 26 years or older ( $n = 511$ ).

We estimated MI across age using the same algorithm as previously used across sex (Table 6). Younger adults ( $M_{age} = 21.13$ ,  $SD = 2.65$ ) were opposed to older adults ( $M_{age} = 33.38$ ,  $SD = 7.45$ ). The configural invariance model produced adequate model fit in terms of RMSEA, CFI, TLI, and SRMR. Metric invariance was strongly supported by similar and in some cases better indices as compared with the configural invariance test, including insignificant increase in chi-square and large decrease in BIC (1,567). The test of strong invariance resulted in negligible decrease in fit as compared with the metric test (e.g.,  $\Delta CFI = -.005$ ), and the test of strict invariance revealed only negligible decrease in fit as compared with the strong invariance test (e.g.,  $\Delta CFI = -.002$ ). Overall, these findings demonstrate that the BFI-2's hierarchical structure was invariant across these two nonoverlapping age groups. It therefore allow the straightforward interpretation of age trends in observed BFI-2 scores.

In this study, the younger group showed no substantial mean-level difference from the older group across the domains (Table S3.11, supplementary materials available online). The largest difference was obtained for open-mindedness ( $d = 0.13$ ) such that older respondents were slightly more open-minded than younger ones. At the facet level, the only small difference was obtained for intellectual curiosity such that the older respondents were slightly more curious than the younger respondents ( $d = 0.22$ ). In terms of Pearson correlations, however, age of users showed small correlations with two domains (Table S3.9, supplementary materials available online): older users were slightly more agreeable ( $r = .09, p < .01$ ) and conscientious ( $r = .08, p < .05$ ). At the facet level, some associations were somewhat stronger: older users demonstrated higher levels of trust ( $r = .14, p < .001$ ), productiveness ( $r = .10, p < .01$ ), and responsibility ( $r = .10, p < .01$ ).

## Discussion

In an independent Internet validation sample, Study 2 successfully replicates the strong psychometric properties of the Russian BFI-2, and also shows MI of the BFI-2 across sex and age. By means of RI-EFA, the Russian BFI-2 showed good factorial validity at both the domain and facet levels.

In this study, strict MI of the BFI-2 across sex was replicated from Study 1. This result strongly supports the possibility of using observed mean scores on the BFI-2 to investigate sex differences at the domain and facet levels. The pattern of mean sex differences obtained in Study 2 closely corresponded to that obtained in Study 1. In both the student and Internet samples, women showed higher negative emotionality and agreeableness along with lower extraversion than men. These differences were mirrored in facets of these domains. Across the open-mindedness domain, opposite differences were obtained for different facets in both studies: women scored higher than men in aesthetic sensitivity while men had higher intellectual curiosity and creative imagination. Finally, in both studies, sex differences in conscientiousness and its facets were negligible.

In this study, for the first time, strict MI of BFI-2 was also revealed across two nonoverlapping age groups. These findings support the possibility to investigate age differences using observed scores on the BFI-2. Moreover, the positive age trends in agreeableness, conscientiousness, and their facets obtained in Study 2 replicate previous findings regarding life span personality development (e.g., Bleidorn et al., 2013; Soto et al., 2011; Specht et al., 2011; Srivastava et al., 2003).

## General Discussion

The BFI questionnaire has gained popularity in the past three decades among personality researchers all over the

world, and the BFI-2 represents a major update and extension of this measure. This new version provides a number of new characteristics which make BFI-2 an important instrument for personality measurement. Given the BFI-2's increasing popularity, the present research is critically important in two aspects. First, we have shown that the BFI-2 is able to maintain strict MI across two characteristics that are pivotal in personality research, namely sex and age. Second, this research has shown that the BFI-2 is able to effectively function well beyond English language, and even beyond Germanic languages and Western cultures. Particularly, we have demonstrated that the BFI-2 is able to adequately operate in a Slavic, Russian language, in the context of the Russian culture.

## The Russian BFI-2

The good psychometric quality of the Russian BFI-2 was demonstrated in various aspects. Across two different samples, the domain scales showed high internal consistency, while facet-level internal consistency had been initially, in a pilot study, quite problematic. Nonetheless, after a revision of the test this concern has been effectively addressed, and across two main studies internal consistency of the facet subscales consistently exceeded .60, despite the brevity of these four-item scales. Moreover, test-retest reliability assessed in Study 1 and measured with the item-recall-resistant Green's alpha was adequate for both the domain and facet scales. These results are in line with those obtained before with other BFI-2 versions (Denissen et al., 2019; Soto & John, 2017b). In general, the Russian BFI-2 has proved to be a reliable and valid measure of personality at both the domain and facet levels.

Soto and John (2017b) recommended the centering of scores around each person's mean across the entire set of 60 items as a way to reduce the acquiescence bias effect. We administered this procedure as well, which gave us improved factorial validity in the form of higher weights on intended components, lower weights on unintended components, larger variance explained by the presumed components, and scree tests that clearly suggested five meaningful components (with PCA).

As in the vast majority of past research using multifactor personality questionnaires, in this study, we obtained poor model fit while using simple structure CFA. In this sense, our study confirms that the restrictions of CFA poorly fit the complex nature of multifaceted questionnaires like the BFI-2. The use of methods that alleviate these restrictions, particularly which allow loadings across different factors to be nonzero, gives a more adequate estimation of personality structure. In this study, the method of RI-EFA (Aichholzer, 2014), a special case of ESEM, has proved useful for simultaneously addressing the cross-loading and acquiescence bias problems.

Chiorri et al. (2016), employing ESEM, showed that including facets into the model by allowing correlated uniquenesses (CUs) may substantially improve model fit of the BFI-1 structure. In this research, we demonstrate that such an approach is relevant for BFI-2 as well. The model with CUs greatly improved fit as compared with the five-domain model with uncorrelated uniquenesses. Given that, unlike BFI-1, the facets have been included into the BFI-2 in a systematic and theoretically driven fashion (Soto & John, 2017b), our findings demonstrate the promise of facet-level analyses of the BFI-2 for future research.

The domain scales of the Russian BFI-2 showed good convergence with the corresponding Big Five scales, as well as discrimination from irrelevant scales, of the IPIP. The nomological network of correlations between the BFI-2 and a set of various self-report measures was consistent with the extant personality literature. After controlling for the remaining domains, certain domains correlated strongly—and repeatedly—with certain criterion measures: open-mindedness correlated with need for cognition, extraversion with positive affect and satisfaction with life, negative emotionality with negative affect, and extraversion and (low) negative emotionality with self-esteem.

At the facet level, we obtained a more detailed picture of the manner in which personality traits relate to these constructs. Since there are few studies that previously addressed the associations between these criteria and Big Five facets, our findings in this regard were mostly exploratory. Need for cognition was a correlate of two facets of open-mindedness, namely intellectual curiosity and creative imagination, along with an extraversion facet, assertiveness. Positive affect was a correlate of only one extraversion facet, energy level, as well as a conscientiousness facet, productiveness, and a negative affectivity facet, (low) depression. The association between negative emotionality and satisfaction with life was largely accounted for by the depression facet. Finally, depression showed the strongest association, whereas the remaining facets showed weak associations, with low self-esteem.

Taken together, these findings support the reliability and validity of the Russian BFI-2. We therefore recommend the revised Russian BFI-2 for use in future Russian language and cross-cultural personality research.

### *Measurement Invariance and Group Differences*

Strict MI of the BFI-2's hierarchical structure in both Study 1 and Study 2 allowed us to compare observed mean sex differences in domains and facets. At the domain level, we found that women scored much higher on negative emotionality and moderately higher in agreeableness, whereas men scored slightly higher in extraversion. This pattern closely corresponds to the extant literature (Chiorri et al., 2016; Kajonius & Johnson, 2018; Samuel et al., 2015;

Schmitt et al., 2008) except for extraversion, which has been normally reported to be somewhat higher among women. This unexpected result may be due to the energy level facet, which in our case was higher among men than women (cf. Denissen et al., 2019; McCrae & Terracciano, 2005; Soto & John, 2017b).

For the most part, however, our findings at the facet level correspond to the extant literature (e.g., Costa et al., 2001; Denissen et al., 2019; Soto & John, 2017b) showing sex differences in several facets. At the same time, across two different Russian samples, we did not confirm sex differences in sociability, conscientiousness and its facets. Across two studies, we found that a lack of sex differences in overall open-mindedness masked differences at the facet level, with higher scores in two facets among men (intellectual curiosity and creative imagination) but higher scores in a third facet among women (aesthetic sensitivity). A similar pattern has previously been observed with the BFI-1 (Schmitt et al., 2008) and the NEO Personality Inventories (Costa et al., 2001), and the present research confirms that this pattern extends to the BFI-2. These findings illustrate how facet-level analysis may contribute to our understanding of personality, beyond the domain level.

In Study 2, we also found strict MI across age, between groups of adolescents and emerging adults versus older adults. Moreover, correlational analyses provided further support for the maturity principle of adult personality development by successfully replicating the gradual, positive age trends in agreeableness, conscientiousness, and their facets often found in studies of adult personality development (Caspi, Roberts, & Shiner, 2005). We also found a gradual, positive age trend in open-mindedness, especially the facets of intellectual curiosity and creative imagination. Although not predicted by the maturity principle, these trends match with some previous cross-sectional and longitudinal findings (e.g., Roberts, Walton, & Viechtbauer, 2006; Soto et al., 2011), and thus merit further investigation.

### *Limitations and Future Directions*

Despite a positive overall outcome of our research, it also had some important limitations that highlight promising directions for future research. For example, although the present research was the first to investigate MI of the BFI-2 across age and sex, the younger and older age groups that we compared only differed by about 12 years in their mean age. Further research is therefore needed to investigate whether MI of the BFI-2 also holds in different languages and cultures, as well as across more diverse age groups. Next, although the Russian BFI-2 has been validated in two different Russian samples (i.e., students and Internet users), they were both samples of convenience. A next step may be to validate the Russian BFI-2 using a sample that is more heterogeneous in terms of socioeconomic status and gender.

Despite the fact that the present Russian BFI-2 has solid measurement properties, we advise cautious use when studying nonconvenience, Russian-speaking samples, paying special attention to the psychometrics of the questionnaire among adolescents and noncollege-educated adults (cf. Rammstedt, Goldberg, & Borg, 2010; Soto et al., 2008). Another limitation is that the criterion validity of the Russian BFI-2 has been assessed solely by examining cross-sectional associations between self-report measures. Therefore, future research can examine whether the Russian BFI-2 is capable of prospectively predicting various behaviors and life outcomes, using informant-report or observer-report measures.

## Conclusion

Across several studies, we developed, refined, and validated a Russian adaptation of the BFI-2. The final Russian BFI-2 demonstrated good internal consistency and retest reliability, as well as good convergent, discriminant, and criterion validity in student and nonstudent samples. Of equal importance, we also found evidence for strict MI of the BFI-2 across sex and age groups. We therefore believe that the BFI-2 can effectively contribute to future personality research, both within and beyond the Russian language context.

## Acknowledgments

We gratefully acknowledge Robert Lee Bates (Perm, Russia), Kristina Dumas (Albuquerque, NM), Asya Kasimova (Leuven, Belgium), Timur Khaliulin (West Virginia University, WV), Elizabeth Kulazhenkova-Clerot (London, the United Kingdom), Elena Saly (London, the United Kingdom), Anna White (Westlake, OH), and Anastasia Anatoli Yakovlev (London, the United Kingdom) for their help in translating the BFI-2.

## Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Christopher J. Soto and Oliver P. John are copyright holders for the Big Five Inventory–2.

## Funding

The first author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2019-2020 (Grant No. 19-01-003) and within the framework of the Russian Academic Excellence Project “5-100.”

## ORCID iDs

Sergei Shchebetenko  <https://orcid.org/0000-0001-5790-9731>  
Arina M. Mishkevich  <https://orcid.org/0000-0001-6666-3454>

## Supplemental Material

Supplemental material for this article is available online.

## Notes

1. The original BFI-2 article (Soto & John 2017b) discuss why this measure adopts the labels of negative emotionality (rather than neuroticism) and open-mindedness (rather than openness to experience).
2. For an unusual case of adapting a Russian questionnaire to the English language, see Bishop, Jacks, and Tandy (1993).
3. The scales of emotional stability and intellect in the IPIP are conceptually similar to the domains of negative emotionality and open-mindedness according to the BFI-2 nomenclature.
4. Across all analyses in Studies 1 and 2 the revised version showed improved psychometric quality in terms of reliability, factorial, convergent, and discriminant validity (see supplementary materials available online for details).

## References

- Aichholzer, J. (2014). Random intercept EFA of personality scales. *Journal of Research in Personality, 53*, 1-4. doi:10.1016/j.jrp.2014.07.001
- Aichholzer, J., Danner, D., & Rammstedt, B. (2018). Facets of personality and “ideological asymmetries.” *Journal of Research in Personality, 77*, 90-100. doi:10.1016/j.jrp.2018.09.010
- Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 397-438. doi:10.1080/10705510903008204
- Bishop, D., Jacks, H., & Tandy, S. B. (1993). The Structure of Temperament Questionnaire (STQ): Results from a U.S. sample. *Personality and Individual Differences, 14*, 485-487. doi:10.1016/0191-8869(93)90318-W
- Bleidorn, W., Klimstra, T. A., Denissen, J. J. A., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2013). Personality maturation around the world: A cross-cultural examination of social-investment theory. *Psychological Science, 24*, 2530-2540. doi:10.1177/0956797613498396
- Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*, 306-307. doi:10.1207/s15327752jpa4803\_13
- Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology, 56*, 453-484. doi:10.1146/annurev.psych.55.090902.141913
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 233-255. doi:10.1207/S15328007SEM0902\_5
- Chiorri, C., Marsh, H. W., Ubbiali, A., & Donati, D. (2016). Testing the factor structure and measurement invariance across gender of the Big Five Inventory through exploratory structural equation modeling. *Journal of Personality Assessment, 98*, 88-99. doi:10.1080/00223891.2015.1035381
- Church, A. T., Alvarez, J. M., Mai, N. T. Q., French, B. F., Katigbak, M. S., & Ortiz, F. A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the Revised NEO Personality

- Inventory. *Journal of Personality and Social Psychology*, 101, 1068-1089. doi:10.1037/a0025290
- Clark, M. A., Lelchook, A. M., & Taylor, M. L. (2010). Beyond the Big Five: How narcissism, perfectionism, and dispositional affect relate to workaholism. *Personality and Individual Differences*, 48, 786-791. doi:10.1016/j.paid.2010.01.013
- Costa, P. T. Jr., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, 81, 322-331. doi:10.1037/0022-3514.81.2.322
- Danner, D., Rammstedt, B., Bluemke, M., Lechner, C., Berres, S., Knopf, T., . . . John, O. P. (2019). Das Big Five Inventar 2: Validierung eines Persönlichkeitsinventars zur Erfassung von 5 Persönlichkeitsdomänen und 15 Facetten [The Big Five Inventory 2: Validate a personality inventory to capture 5 personality domains and 15 facets]. *Diagnostica*. Advance online publication. doi:10.1026/0012-1924/a000218
- Denissen, J. J. A., Geenen, R., Soto, C. J., John, O. P., & van Aken, M. A. G. (2019). The Big Five Inventory-2: Replication of psychometric properties in a Dutch adaptation and first evidence for the discriminant predictive validity of the facet scales. *Journal of Personality Assessment*. Advance online publication. doi:10.1080/00223891.2018.1539004
- Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, 49, 71-75. doi:10.1207/s15327752jpa4901\_13
- Eysenck, H. J. (1947). *Dimensions of personality*. Oxford, England: Kegan Paul.
- Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences*, 75, 47-52. doi:10.1016/j.paid.2014.11.017
- Gebauer, J. E., Bleidorn, W., Gosling, S. D., Rentfrow, P. J., Lamb, M. E., & Potter, J. (2014). Cross-cultural variations in Big Five relationships with religiosity: A sociocultural motives perspective. *Journal of Personality and Social Psychology*, 107, 1064-1091. doi:10.1037/a0037683
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26-34. doi:10.1037/0003-066X.48.1.26
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40, 84-96. doi:10.1016/j.jrp.2005.08.007
- Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods*, 8, 88-101. doi:10.1037/1082-989X.8.1.88
- John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory—Versions 4a and 54*. Berkeley: Institute of Personality and Social Research, University of California.
- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big-five taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114-158). New York, NY: Guilford Press.
- Hanin, Y., Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. (1991). A cross-cultural study of personality: Russia and England. *Personality and Individual Differences*, 12, 265-271. doi:10.1016/0191-8869(91)90112-0
- Haslam, N., Whelan, J., & Bastian, B. (2009). Big Five traits mediate associations between values and subjective well-being. *Personality and Individual Differences*, 46, 40-42. doi:10.1016/j.paid.2008.09.001
- HBO Entertainment; co-executive producers, Martin, G.R.R., Gerardis, Vince, Vicinanza, Ralph, Casady, Guymon, & Strauss, Carolyn; producers, Huffam, Mark, & Doelger, Frank; executive producers, Benioff, David, & Weiss, D.B.; created by Benioff, D., & Weiss, D.B. Television 360; Grok! Television; Generator Entertainment; Bighead Littlehead. (2012). *Game of thrones. The complete first season*. New York, NY: HBO Home Entertainment.
- Kajonius, P. J., & Johnson, J. (2018). Sex differences in 30 facets of the five-factor model of personality in the large public ( $N = 320,128$ ). *Personality and Individual Differences*, 129, 126-130. doi:10.1016/j.paid.2018.03.026
- Kniazev, G. G., Mitrofanova, L. G., & Bocharov, V. A. (2010). Валидизация русскоязычной версии опросника Л. Голдберга «Маркеры факторов “Большой Пятерки” [Validization of the Russian version of the Goldberg’s “Big-Five Factor Markers” inventory]. *Psikhologicheskii Zhurnal*, 31, 100-110.
- Lang, F. R., John, D., Lüdtke, O., Schupp, J., & Wagner, G. G. (2011). Short assessment of the Big Five: Robust across survey methods except telephone interviewing. *Behavior Research Methods*, 43, 548-567. doi:10.3758/s13428-011-0066-z
- Madrid, H. P., & Patterson, M. G. (2016). Creativity at work as a joint function between openness to experience, need for cognition and organizational fairness. *Learning and Individual Differences*, 51, 409-416. doi:10.1016/j.lindif.2015.07.010
- Margolis, S., Schwitzgebel, E., Ozer, D. J., & Lyubomirsky, S. (2018). A new measure of life satisfaction: The Riverside Life Satisfaction Scale. *Journal of Personality Assessment*. Advance online publication. doi:10.1080/00223891.2018.1464457
- Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, 10, 85-110. doi:10.1146/annurev-clinpsy-032813-153700
- Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students’ evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, 16, 439-476. doi:10.1080/10705510903008220
- Marshall, T. C., Lefringhausen, K., & Ferenczi, N. (2015). The Big Five, self-esteem, and narcissism as predictors of the topics people write about in Facebook status updates. *Personality and Individual Differences*, 85, 35-40. doi:10.1016/j.paid.2015.04.039
- Martin, T. A., Costa, P. T., Oryol, V. E., Rukavishnikov, A. A., & Senin, I. G. (2002). Applications of the Russian NEO-PI-R. In R. R. McCrae & J. Allik (Eds.), *The five-factor model of personality across cultures* (pp. 261-277). Boston, MA: Springer. doi:10.1007/978-1-4615-0763-5\_13
- McCrae, R. R., Terracciano, A., & 78 Members of the Personality Profiles of Cultures Project. (2005). Universal features of personality traits from the observer’s perspective: Data from 50 cultures. *Journal of Personality and Social Psychology*, 88, 547-561. doi:10.1037/0022-3514.88.3.547

- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525-543. doi:10.1007/BF02294825
- Morin, A. J. S., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 116-139. doi:10.1080/10705511.2014.961800
- Nye, C. D., Allemand, M., Gosling, S. D., Potter, J., & Roberts, B. W. (2016). Personality trait differences between young and middle-aged adults: Measurement artifacts or actual trends? *Journal of Personality*, *84*, 473-492. doi:10.1111/jopy.12173
- Osin, E. N. (2012). Измерение позитивных и негативных эмоций: разработка русскоязычного аналога методики PANAS [Measurement of positive and negative emotions: Development of a Russian analogue of the PANAS measure]. *Psychology: Journal of the Higher School of Economics*, *9*(4), 91-110.
- Osin, E. N., & Leontiev, D. A. (2008). Аprobация русскоязычных версий двух шкал экспресс-оценки субъективного благополучия [Testing the Russian versions of two scales of subjective well-being]. Retrieved from <https://publications.hse.ru/mirror/pubs/share/folder/pjuun7fz60/direct/78753837>
- Pavot, W., & Diener, E. (2008). The Satisfaction With Life Scale and the emerging construct of life satisfaction. *Journal of Positive Psychology*, *3*, 137-152. doi:10.1080/17439760701756946
- Rammstedt, B., Danner, D., Soto, C. J., & John, O. P. (2018). Validation of the short and extra-short forms of the Big Five Inventory-2 (BFI-2) and their German adaptations. *European Journal of Psychological Assessment*. Advance online publication. doi:10.1027/1015-5759/a000481
- Rammstedt, B., & Farmer, R. F. (2013). The impact of acquiescence on the evaluation of personality structure. *Psychological Assessment*, *25*, 1137-1145. doi:10.1037/a0033323
- Rammstedt, B., Goldberg, L. R., & Borg, I. (2010). The measurement equivalence of Big-Five factor markers for persons with different levels of education. *Journal of Research in Personality*, *44*, 53-61. doi:10.1016/j.jrp.2009.10.005
- Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1-25. doi:10.1037/0033-2909.132.1.1
- Robins, R. W., Tracy, J. L., Trzesniewski, K., Potter, J., & Gosling, S. D. (2001). Personality correlates of self-esteem. *Journal of Research in Personality*, *35*, 463-482. doi:10.1006/jrpe.2001.2324
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Sadowski, C. J., & Cogburn, H. E. (1997). Need for cognition in the Big-Five factor structure. *Journal of Psychology*, *131*, 307-312. doi:10.1080/00223989709603517
- Samuel, D. B., South, S. C., & Griffin, S. A. (2015). Factorial invariance of the Five-Factor Model rating form across gender. *Assessment*, *22*, 65-75. doi:10.1177/1073191114536772
- Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, *94*, 168-182. doi:10.1037/0022-3514.94.1.168
- Shchebetenko, S. (2011). Психометрика русской версии Шкалы потребности в познании [Psychometrics of a Russian version of the Need for Cognition Scale]. *Vestnik Permskogo Universiteta: Filosofia. Psikhologiya. Soziologiya*, *6*, 87-100.
- Shchebetenko, S. (2014). "The best man in the world": Attitudes toward personality traits. *Psychology: Journal of the Higher School of Economics*, *11*(4), 129-148.
- Shchebetenko, S., Kalugin, A. Y., Mishkevich, A. M., Soto, C. J., John, O. P. (2019). "Measurement invariance and sex and age differences of the Big Five Inventory-2: Evidence from the Russian version," 2017-2018 [Data file and code book]. Retrieved from <https://data.mendeley.com/datasets/379m879wpf/1>
- Slobodskaya, H. R., Knyazev, G. G., Safronova, M. V., & Wilson, G. D. (2003). Development of a short form of the Gray-Wilson Personality Questionnaire: Its use in measuring personality and adjustment among Russian adolescents. *Personality and Individual Differences*, *35*, 1049-1059. doi:10.1016/S0191-8869(02)00317-3
- Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality*, *43*, 84-90. doi:10.1016/j.jrp.2008.10.002
- Soto, C. J., & John, O. P. (2017a). Short and extra-short forms of the Big Five Inventory-2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, *68*, 69-81. doi:10.1016/j.jrp.2017.02.004
- Soto, C. J., & John, O. P. (2017b). 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*, 117-143. doi:10.1037/pspp0000096
- Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2008). The developmental psychometrics of Big Five self-reports: Acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *Journal of Personality and Social Psychology*, *94*, 718-737. doi:10.1037/0022-3514.94.4.718
- Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology*, *100*, 330-348. doi:10.1037/a0021717
- Specht, J., Egloff, B., & Schmukle, S. C. (2011). Stability and change of personality across the life course: The impact of age and major life events on mean-level and rank-order stability of the Big Five. *Journal of Personality and Social Psychology*, *101*, 862-882. doi:10.1037/a0024950
- Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, *84*, 1041-1053. doi:10.1037/0022-3514.84.5.1041
- Suzuki, T., South, S. C., Samuel, D. B., Wright, A. G. C., Yalch, M. M., Hopwood, C. J., & Thomas, K. M. (2019). Measurement invariance of the DSM-5 Section III pathological personality trait model across sex. *Personality Disorders: Theory, Research, and Treatment*, *10*, 114-122. doi:10.1037/per0000291
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063-1070.