Article

The Big Five Inventory–2 in China: A Comprehensive Psychometric Evaluation in Four Diverse Samples

Assessment
1–23

DOI: 10.1177/10731911211008245
journals.sagepub.com/home/asm
SAGE

Bo Zhang[1], Yi Ming Li[2], Jian Li[3], Jing Luo[4], Yonghao Ye[5], Lu Yin[6], Zhuosheng Chen[7], Christopher J. Soto[8], and Oliver P. John[9]

Abstract

The Big Five Inventory-2 (BFI-2) has received wide recognition since its publication because it strikes a good balance between content coverage and brevity. The current study translated the BFI-2 into Chinese, evaluated its psychometric properties in four diverse Chinese samples (college students, adult employees, adults treated for substance use, and adolescents), and compared its factor structure with those obtained from two U.S. samples. Across two studies, the Chinese BFI-2 demonstrated good reliability (Cronbach's α and test–retest reliability), structural validity, convergent/discriminant validity, and criterion-related validity at the domain level. At lower levels of analyses, some facets and negatively worded items functioned better among participants with higher than those with lower education levels. Implications, limitations, and future directions are discussed.

Keywords

Big Five, facets, personality assessment, validation, Chinese

Personality traits, broadly defined as "the relatively enduring patterns of thoughts, feelings, and behaviors that reflect the tendency to respond in certain ways under certain circumstances" (Roberts, 2009, p. 140), have been consistently shown to be related to various important life outcomes. For example, meta-analyses have revealed robust relationships between personality traits and physical and mental health (Strickhouser et al., 2017), mortality (Jokela et al., 2013), life satisfaction (Anglim et al., 2020), job performance (Judge et al., 2013), and academic success (Poropat, 2009). Moreover, such links between personality traits and life outcomes have been proven to be both replicable and generalizable across age and gender (Soto, 2020). The Organization for Economic Cooperation and Development (OECD) has incorporated personality as a core component of the social emotional skills assessment (Chernyshenko et al., 2018).

However, it is fair to state that most of these encouraging findings were obtained from western, educated, industrialized, rich, and democratic (WEIRD) populations, whereas most people in the world are not WEIRD (Rad et al., 2018). To what extent can these findings be generalized to non-WEIRD populations? One major challenge to addressing this question is the lack of psychometrically sound measures of personality for these populations. Most of the commonly used personality measures have been developed in WEIRD populations and only limited efforts have been devoted to validating them in non-WEIRD populations. There is thus an urgent need for more formal validation studies in non-WEIRD populations.

As the country with the world's largest population, China has a rich culture that is distinct from western ones. This makes it particularly important to test whether personality theories and research findings can be generalized to the Chinese cultural context. However, despite some important previous work (e.g., F. M. Cheung et al., 1996; Yang & Bond, 1990), the number of personality studies in Chinese populations seems disproportionately small compared with its population size and cultural richness. To facilitate

[1]Texas A&M University, College Station, TX, USA
[2]Beijing Institute of Education, Beijing, China
[3]Beijing Normal University, Beijing, China
[4]Northwestern University, Evanston, IL, USA
[5]Ministry of Justice, Beijing, China
[6]Beijing Tiantanghe Compulsory Isolation Detoxification Center, Beijing, China
[7]China University of Political Science and Law, Beijing, China
[8]Colby College, Waterville, ME, USA
[9]University of California, Berkeley, CA, USA

Corresponding Author:
Jian Li, Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, 19 Xinjiekou Outer Street, HaiDian District, Beijing 100875, China.
Email: jianli@bnu.edu.cn

Chinese personality studies, the present study aimed to validate the Chinese version of the Big Five Inventory–2 (BFI-2; Soto & John, 2017), a widely used Big Five personality measure that strikes a good balance between measurement breadth and fidelity, and examine its applicability in four diverse Chinese samples. Specifically, in Study 1, we validated our translation in a large college student sample and an adult employee sample. In Study 2, we further examined the applicability of the Chinese BFI-2 in two relatively less-studied samples: a sample of high school students who were in the transitional phase of adolescence, and a clinical sample of adults treated for substance abuse. Overall, we aim to provide researchers interested in studying personality in the Chinese context with a psychometrically sound measure that can be applied to different Chinese-speaking populations.

The Structure of Personality and Its Assessment

Since the early 1990s, the Big Five have become the most widely adopted framework to describe personality characteristics as decades of research converged to suggest the existence of five robust trait domains (John et al., 2008), most commonly labeled extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience. Moreover, researchers have increasingly recognized that personality traits are structured hierarchically, with each broad Big Five trait domain subsuming a number of more specific "facet" or "aspect" traits (Costa & McCrae, 1995; DeYoung et al., 2007; Zhang, Sun, Cao, et al., 2020).

The hierarchical nature of personality traits has motivated the development of personality inventories that assess the Big Five at different levels of abstraction. For example, many brief inventories only assess personality traits at the level of the broad Big Five: prominent examples include the Ten-Item Personality Inventory (Gosling et al., 2003), Big Five Mini-Markers (Saucier, 1994), BFI (John & Srivastava, 1999), NEO Five-Factor Inventory (Costa & McCrae, 1992), and IPIP Five-Factor Model scales (IPIP-FFM; Ehrhart et al., 2008). Other, typically longer, inventories assess more-specific traits beneath the Big Five. For example, the Big Five Aspect Scales (DeYoung et al., 2007) measures two aspects within each Big Five domain, the NEO Personality Inventory–Revised (NEO-PI-R; Costa & McCrae, 1992) and International Personality Item Pool-NEO Inventory (IPIP-NEO; Goldberg, 1999) assess six facets per domain, and the Abridged Big Five Circumplex scales for the IPIP (AB5C-IPIP; Goldberg, 1999) assesses nine facets per domain.

Despite the growing interest in lower level Big Five traits, researchers and practitioners face a dilemma when choosing among the available measures. On one hand, they want a scale that captures the prototypical facets of each Big Five domain because facet-level assessment has been shown to have incremental utility in predicting outcomes over and above domain-level scores (e.g., Judge et al., 2013). Inventories that only assess personality at the level of the broad Big Five ignore a substantial amount of meaningful personality information. On the other hand, researchers also need to take into consideration of the costs of more nuanced assessment. Facet-level measures like NEO-PI-R (240 items) and AB5C (456 items) are too long and too burdensome for many research and applied contexts. Thus, the ideal measure for many contexts would be one that strikes a good balance between content coverage and brevity such that it assesses the core facets of each Big Five domain while keeping to a manageable length.

The Big Five Inventory–2

Motivated by the need for a brief, hierarchically structured measure of personality traits, Soto and John (2017) developed the BFI-2 as a major revision of the widely used BFI (John et al., 1991). The original BFI used 44 short, easy-to-understand items to efficiently assess the Big Five domains in a conceptually and empirically coherent manner. The BFI-2 was developed to maintain the clarity, coherence, and efficiency of the original BFI, while also hierarchically assessing three prominent facet traits within each Big Five domain and controlling for individual differences in acquiescent response style (an individual's tendency to consistently agree or disagree with questionnaire items, regardless of their content). The BFI-2 therefore includes a total of 60 items assessing 15 four-item facets nested within 5 twelve-item domains with an equal number of positively and negatively keyed items on each facet/domain. Most people can complete these items in approximately 6 minutes. Thus, the BFI-2 is a hierarchically structured and relatively comprehensive, but still economical, measure of the Big Five personality traits.

Since its publication, the BFI-2 has become widely used in both research and applied contexts. As of February 18, 2021, the initial BFI-2 paper has been cited 690 times according to Google scholar. Validated and published translations are available in Danish (Vedel et al., 2020), Dutch (Denissen et al., 2019), German (Rammstedt et al., 2018), and Russian (Shchebetenko et al., 2020), Slovak (Halama et al., 2020), with preliminary translations available in many other languages. Therefore, the present study represents a timely effort to further expand the cultural coverage of the BFI-2 by developing and validating a Chinese version.

Personality Assessment in China

Some psychometric research has been conducted in China, either to develop and validate indigenous Chinese personality measures (F. M. Cheung et al., 1996; Wang & Cui, 2004), or to adapt personality scales developed in western cultures for the Chinese cultural context (e.g., Dai & Wu, 2005; Qian

et al., 2000; Yao & Liang, 2010; Zheng et al., 2008). Such validation work is important because it contributes to a shared conceptual and empirical framework for personality research around the globe. However, a closer examination of previous validation studies reveals limitations that may cast doubts on the construct validity of some translated scales. For example, Luo and Dai (2011) meta-analytically showed that the Chinese version of the BFI, NEO-FFI, and Big Five Mini Markers had substantially lower reliabilities than the original English versions. The difference was particularly salient for the Extraversion and Openness subscales of the BFI. Given that the BFI-2 is an updated version of the BFI, it is important to examine whether the BFI-2 will perform better than the BFI in Chinese populations. In addition, many previous translation and validation studies solely relied on convenient college student samples and reported incomplete information. For example, few studies examined the test–retest stability of Chinese personality measures, and even fewer examined criterion-related validity. Moreover, most imported scales only assess personality traits at the level of the broad Big Five, without differentiating among facet-level traits. Another concern is accessibility, given that many imported scales are commercially published and require licensing fees to use (e.g., the NEO-PI-R and NEO-FFI). There is thus a clear need to validate a personality scale in the Chinese cultural context that hierarchically assesses the Big Five domains and their constituent facets, possesses good psychometric properties, and is publicly available to facilitate cross-cultural scientific communication.

Overview of the Present Research

We conducted two studies to develop and validate the Chinese BFI-2, and to examine its applicability in four diverse samples. Specifically, Study 1 primarily focused on validating the translation in a college student sample and a sample of employed adults. These samples are demographically similar to those commonly used in personality research and in the initial BFI-2 scale development studies (Soto & John, 2017). Study 2 tested whether the psychometric properties of the Chinese BFI-2 could be generalized to two less commonly studied populations in personality research: a sample of secondary school students and a clinical sample being treated for substance abuse. We report how we determined our sample size, all data exclusions, and all measures used in the study. All data (except that from the clinical sample), scripts, and measures used in the study can be found on the Open Science Framework[1].

Study 1

Study 1 was conducted with four main goals. The first was to adapt the BFI-2 into the Chinese language and cultural context. The second goal was to test the quality of this translation by examining its reliability and structural validity, and compare the factor loading patterns with those obtained from two U.S. samples. The third goal was to examine the convergent and discriminant validity of the Chinese BFI-2. The fourth goal was to examine criterion-related validity. We aimed for at least 300 valid respondents in each sample so that ratios between the number of items and sample sizes were at least 1:5. This study was approved by the institutional review board at Beijing Normal University.

Method

Scale Translation. Six steps were followed to ensure the accuracy of translation. In Step 1, the first two authors independently translated the English language BFI-2 into Chinese. Nonredundant items from the original BFI were also translated as potential replacement items. In Step 2, the first three authors checked translations from Step 1 and discussed each item in-depth to select the most appropriate translation. When the two translations of a certain item were considered equally good, both were retained. In Step 3, the fourth author, who holds a PhD in personality psychology from a U.S. university and is fluent in both English and Mandarin Chinese, independently back-translated the candidate Chinese BFI-2 items into English. In Step 4, the first author checked for discrepancies between the original, English language BFI-2 and the back-translated BFI-2, and modified some Chinese BFI-2 item translations to resolve potential misunderstandings. In Step 5, two bilingual PhD candidates in psychology at a U.S. university further compared the Chinese translation from Step 4 with the original, English language BFI-2 and made some minor adjustments to make the Chinese translations easier to understand. Finally, in Step 6, the first author and the third author reviewed and finalized a set of 96 candidate Chinese BFI-2 items. This included the 60 item translations that the first two authors considered most promising, together with 36 potential backup items. Among the 36 backup items, about 2/3 of them were alternative translations that we considered usable but not as good as the first 60 translations and the remaining ones were from the BFI. They were included just in case that some items we considered the best did not perform well empirically. It was found that the first 60 items performed well, and that substituting alternative item translations would not consistently improve the BFI-2 scales' basic psychometric characteristics (in terms of item-total correlations and reliability). Therefore, the first 60 items were retained as the final Chinese BFI-2 (see the appendix for the full scale), and all analyses presented below were based on these 60 items.

Participants and Procedures

College Sample. As part of their final project, undergraduate students ($n = 96$) enrolled in the third author's course on

psychological testing at a Chinese university reached out to 10-20 of their friends and asked them to complete a personality survey. After their friends agreed, they sent the survey (through the Chinese survey platform www.wjx.cn) containing all measures reported below. To incentivize participants, they were offered the opportunity to know about their assessment results (e.g., personality scores, relative positions in the sample, and what these scores mean) from their friends if participants were interested. Students responsible for data collection were instructed to anonymize the surveys before they turned them in. In total, 1,290 college students responded. To ensure that our results were based on valid responses, we embedded four quality check items (e.g., Please choose "Agree") throughout the survey and used these items to screen out inattentive respondents.[2] After excluding respondents who missed more than 1 out of 4 quality check items, 1,194 participants were retained in the final sample. This included 470 males and 724 females spreading across China, with more than 200 cities and 50 college majors represented. These students ranged in age from 17 to 28, with most between the ages of 18 and 21 years (M_{age} = 19.46, SD_{age} = 1.46).

Adult Employee Sample. Forty-one human resources managers representing 41 different organizations disseminated the survey through the online platforms QQ and Wechat every workday morning for 5 days and encouraged employees to complete the survey. Employees were told that their responses would help their human resources department develop personnel selection tools in the future. All responses were anonymous. In total, 603 employees responded. After excluding respondents who missed more than 1 out of 4 quality control items, 486 participants were retained in the final sample. This included 195 males and 291 females representing a total of 41 organizations. The participants ranged in age from 18 to 80 years, with most in their 20s and 30s (M_{age} = 29.38, SD_{age} = 8.88).

U.S. comparison samples. To compare the psychometric properties of the Chinese BFI-2 with the original, English language BFI-2, we reanalyzed the internet (n = 1,000) and college student (n = 470) validation samples from Study 3 of Soto and John (2017). There were 500 males and 500 females in the internet sample, and 146 males and 313 females in the college student sample (11 participants in the latter sample did not report their gender). The internet sample was diverse in terms of age (M_{age} = 29.25, SD_{age} = 12.17), and most participants in the college sample were between the ages of 18 and 25 years (M_{age} = 21.68, SD_{age} = 3.26). Overall, the U.S. internet sample is demographically comparable to the Chinese employee sample, and the U.S. college sample is comparable to the Chinese college sample.

Measures

In addition to the BFI-2, participants in the Chinese student and employee samples also completed an alternative, forced-choice measure of the Big Five personality traits (Brown & Maydeu-Olivares, 2011), as well as measures of their parents' education level, subjective well-being (Diener et al., 1985), perceived stress (Cohen et al., 1983), depression (Norton, 2007), aggression (Buss & Perry, 1992), physical health (Schat et al., 2005), dark personality (Jonason, & Webster, 2010), childhood maltreatment experience (Straus et al., 1998), game addiction (Lemmens et al., 2009), job satisfaction, organizational status (homemade), counterproductive work behavior (Spector et al., 2010), organizational citizenship behavior (Spector et al., 2010), and job performance. Descriptive details (e.g., response options, reliability, sample items) about each scale can be found in the online supplementary Table S1. We also examined the structural validity of these criterion measures in each sample and model fit information can be found in the online supplementary Table S2.

Statistical Analyses

In addition to basic psychometric analyses (e.g., Cronbach's α, observed correlations), we used exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009) to examine the domain-level structure of the Chinese BFI-2 while accounting for the nontrivial cross-loadings consistently found in personality measures (Danner et al., 2021). We also conducted confirmatory factor analysis (CFA) to investigate the intended three-facet structure of each domain and cross-cultural measurement invariance.

The domain-level structural analyses aimed to examine whether the Big Five factors could be recovered and were performed to both item-level (individual items served as indicators of the Big Five domains) and facet-level data (facet scores served as indicators of the Big Five domains). At the item level, we first fitted an ordinary oblique ESEM model where all items were allowed to load on all five domain factors. These factors were rotated toward a 60 × 5 target matrix in which the positions of primary loadings were left unspecified and the positions of cross-loadings were set to zero. Because half of the BFI-2 items were negatively worded, we also fitted an ESEM model that included a negative wording factor alongside the five domain factors (ESEMNW). The rationale for this model is that many studies have found that negatively worded items may distort the structural validity of a measure (see Weijters & Baumgartner, 2012, for an excellent summary and discussion), especially in populations with lower reading ability (Gnambs & Schroeders, 2020) or from non-English-speaking countries (Zhang, Luo, et al., 2020). We therefore examined whether

explicitly modelling a negative wording factor could help recover the Big Five factors. In the ESEMNW, the previous target matrix was used for rotation. In addition, all negatively worded items were specified to load on a negative wording method factor. The negative wording factor was constrained to be orthogonal to the five domain factors, and the correlations among the five domain factors were freely estimated.[3] All items were treated as categorical and the estimator weighted least square mean and variance adjusted was used. In the facet-level ESEM, the observed facet scale scores were treated as continuous indicators and maximum likelihood estimation was used. Paralleling the item-level analysis, the factors were rotated toward a 15 × 5 target matrix with unspecified primarily loadings and secondary loadings set to zero. To comprehensively evaluate model fit, we report chi-square, comparative fit index (CFI), Tucker–Lewis index (TLI), and root mean square error of approximation (RMSEA) fit statistics. Tucker's congruence coefficient (TCC) was adopted to quantify the similarity between factor loading patterns across samples. TCC values between .85 and .94 indicate that the factors identified in two samples can be considered fairly similar, and values above .95 mean that the factors can be considered essentially identical (Lorenzo-Seva & Ten Berge, 2006).

The facet-level structural analyses aimed to confirm the three-facet structure of each domain. Specifically, we fitted the following seven models to each domain: (a) a single-factor model where all items loaded on a single domain factor (M1); (b) a single-factor plus acquiescence model where, in addition to the single domain factor, an orthogonal acquiescence factor was estimated with all loadings of positively worded items fixed to 1.00 and those of negatively worded items fixed to −1.00 (after reverse-coding the negatively worded items; M2); (c) a single-factor plus negative wording model where, in addition to the single domain factor, an orthogonal negative wording factor was estimated with all negatively worded items loading on it (M3); (d) a correlated two-factor model where positively worded items and negatively worded items loaded on two separate factors (M4); (e) a correlated three-facet model where the items of each facet scale loaded on their intended facet (M5); (f) a correlated three-facet model plus acquiescence where an additional orthogonal acquiescence factor was estimated (M6); (g) a correlated three-facet plus negative wording model where an additional orthogonal negative wording factor was estimated (M7). In evaluating these models, we focused on relative fit among different models because cutoffs for categorical factor analysis are still under debate (Xia & Yang, 2019).

We also examined whether the BFI-2 was invariant across the two languages by comparing the U.S. college student sample with the Chinese college student sample, and the U.S. internet sample with the Chinese working adult sample. The measurement invariance testing procedure for categorical data proposed by Wu and Estabrook (2016) was adopted because the traditional approach (establishing a baseline model and then imposing increasing restrictions on loadings and thresholds) is dependent on how the baseline threshold model is identified with regard to the scales of latent continuous variables. Imposing seemingly the same parameter constraint (e.g., constraining loadings to be the same across groups) on the same baseline model identified in different ways may in fact lead to different models and thus lead to very different conclusions, which is an important but rarely discussed issue before Wu and Estabrook (2016). The Wu–Estabrook approach circumvented the dependence on model identification strategies by testing the invariance of thresholds first and then factor loadings. It is beyond the scope of the current study to expand on the technical details of the approach. Interested readers are strongly recommended to first refer to Svetina et al. (2020) for a tutorial and then Wu and Estabrook (2016) for the core details. We followed the cutoffs specifically proposed by G. W. Cheung and Rensvold (2002) such that invariance would be indicated by $\Delta RMSEA \leq .01$ and $\Delta CFI \geq -.01$.

All criterion measures were scored by taking the average of all items after reverse coding. Aside from presenting bivariate correlations for criterion-related validity, we also regressed all criteria on either the Big Five factors or the 15 facets to study the unique contribution of each domain/facet via path models. Standardized regression coefficients can be found in the online supplementary Tables S9 to S13. ESEMs were performed in M*plus* 8.0 (Muthén & Muthén, 1998-2017) and all other analyses (e.g., descriptive statistics, reliability, criterion correlations, path analyses, CFA) were conducted in R 4.0.0 (R Core Team, 2020) primarily using psych 2.0.9 (Revelle, 2017) and lavaan 0.6-7 (Rosseel, 2012). Effect size for measurement noninvariance was calculated using the R package dmacs 0.1.0 (Dueber, 2019).

Results and Discussion

Descriptive Statistics. Means and standard deviations of the Big Five domains and their facets for the two Chinese samples and the two U.S. samples are shown in Table 1. As can be seen from the table, the mean personality profile of the Chinese college students was very similar to that of the U.S. college student sample (column-vector $r = .88$), and the mean profile of the Chinese adult employee sample was also similar to that of the U.S. internet sample ($r = .73$). The two Chinese samples also had very similar mean personality profiles ($r = .84$). One noteworthy cultural difference was that the standard deviations of all domains and facets in the two Chinese samples ($M_{SD} = 0.70$ and 0.66) were consistently smaller than their corresponding U.S. counterparts ($M_{SD} = 0.75$ and 0.88). The difference may be

Table 1. Means, Standard Deviations, and Cronbach's α of the Five Domains and Their Facets Across Samples.

	U.S. internet (n = 1,000)			U.S. college (n = 470)			CN college (n = 1,194)			CN employee (n = 486)			CN substance use (n = 765)			CN adolescent T1 (n = 315)			CN adolescent T2 (n = 315)			Test–retest
Domain and facet	*M*	*SD*	α	*M*	*SD*	α	*M*	*SD*	α	*M*	*SD*	α	*M*	*SD*	α	*M*	*SD*	α	*M*	*SD*	α	(adolescent)
Extraversion	3.23	0.80	0.88	3.28	0.71	0.87	3.19	0.66	0.87	3.24	0.60	0.85	3.34	0.57	0.72	3.19	0.70	0.83	3.20	0.70	0.83	0.80
Sociability	2.95	1.05	0.84	3.02	0.96	0.84	3.14	0.89	0.85	3.12	0.80	0.81	3.20	0.85	0.64	3.19	0.99	0.82	3.21	1.04	0.82	0.72
Assertiveness	3.28	0.93	0.77	3.28	0.84	0.76	3.05	0.73	0.70	3.15	0.66	0.67	3.29	0.68	0.44	2.93	0.76	0.56	2.96	0.75	0.56	0.72
Energy	3.47	0.89	0.74	3.53	0.75	0.70	3.39	0.75	0.75	3.44	0.72	0.73	3.56	0.68	0.47	3.45	0.79	0.61	3.44	0.77	0.61	0.65
Agreeableness	3.68	0.64	0.82	3.72	0.60	0.83	3.69	0.47	0.79	3.81	0.49	0.81	3.93	0.53	0.77	3.56	0.53	0.75	3.52	0.52	0.75	0.70
Compassion	3.84	0.78	0.62	3.85	0.75	0.70	3.75	0.61	0.66	3.85	0.60	0.65	4.07	0.62	0.51	3.76	0.66	0.59	3.68	0.68	0.59	0.65
Respectfulness	3.98	0.71	0.67	3.95	0.66	0.69	3.80	0.54	0.60	3.94	0.55	0.62	3.86	0.67	0.55	3.49	0.61	0.47	3.45	0.55	0.47	0.53
Trust	3.23	0.82	0.67	3.34	0.78	0.71	3.51	0.62	0.59	3.65	0.65	0.63	3.85	0.65	0.46	3.42	0.71	0.52	3.42	0.71	0.52	0.63
Conscientiousness	3.43	0.77	0.88	3.47	0.65	0.85	3.29	0.59	0.85	3.68	0.57	0.86	3.73	0.61	0.80	3.17	0.63	0.81	3.17	0.58	0.81	0.73
Organization	3.42	1.01	0.84	3.60	0.88	0.82	3.26	0.77	0.76	3.65	0.73	0.77	3.79	0.72	0.60	3.27	0.85	0.73	3.29	0.80	0.73	0.68
Productiveness	3.37	0.90	0.77	3.34	0.78	0.70	3.02	0.73	0.75	3.55	0.70	0.73	3.72	0.75	0.61	2.83	0.73	0.60	2.87	0.70	0.60	0.58
Responsibility	3.48	0.81	0.70	3.48	0.68	0.63	3.59	0.66	0.74	3.85	0.60	0.70	3.62	0.74	0.57	3.41	0.74	0.65	3.35	0.69	0.65	0.66
Neuroticism	3.07	0.87	0.90	2.91	0.77	0.89	2.96	0.67	0.87	2.72	0.61	0.86	2.54	0.64	0.80	3.15	0.72	0.82	3.10	0.67	0.82	0.76
Anxiety	3.43	0.93	0.76	3.43	0.84	0.77	3.31	0.75	0.73	3.01	0.70	0.68	2.88	0.75	0.51	3.42	0.79	0.59	3.29	0.71	0.59	0.64
Depression	2.85	1.02	0.82	2.57	0.92	0.81	2.86	0.77	0.74	2.63	0.66	0.67	2.40	0.75	0.59	3.03	0.85	0.63	2.98	0.76	0.63	0.64
Emotional volatility	2.93	1.05	0.83	2.74	0.95	0.83	2.70	0.87	0.85	2.52	0.81	0.83	2.35	0.82	0.68	3.01	1.02	0.80	3.04	0.98	0.80	0.73
Openness	3.92	0.65	0.83	3.64	0.64	0.85	3.57	0.59	0.84	3.52	0.57	0.83	3.31	0.59	0.75	3.28	0.62	0.75	3.30	0.65	0.75	0.74
Intellectual curiosity	4.10	0.70	0.66	3.82	0.72	0.72	3.53	0.68	0.65	3.51	0.64	0.60	3.27	0.63	0.28	3.37	0.71	0.39	3.35	0.65	0.39	0.51
Aesthetic sensitivity	3.80	0.92	0.75	3.57	0.90	0.82	3.67	0.85	0.82	3.42	0.86	0.80	3.13	0.84	0.63	3.13	1.00	0.73	3.15	1.03	0.73	0.72
Creative imagination	3.85	0.81	0.74	3.52	0.77	0.76	3.50	0.73	0.82	3.62	0.68	0.80	3.53	0.75	0.69	3.34	0.80	0.73	3.41	0.78	0.73	0.67

explained by the findings that Chinese respondents were more likely to endorse the middle option (endorsement $rate_{CN}$ = 27% and 28%; endorsement $rate_{U.S.}$ = 20% and 17%) and less likely to endorse the two extreme options (endorsement $rate_{CN}$ = 17% and 18%; endorsement $rate_{U.S.}$ = 25% and 34%) than American respondents.

Reliability. Cronbach's αs for the 5 domains and the 15 facets can be found in Table 1. At the domain level, the Chinese BFI-2 in general showed good internal consistency in the college sample (M = 0.84, 0.79-0.87) and the employee sample (M = 0.84, 0.81-0.86) comparable to that in the U.S. college sample (M = 0.86, 0.83-0.89) and internet sample (M = 0.86, 0.82-0.90). At the facet level, reflecting the brevity of the 4-item facet scales compared with the 12-item domain scales, reliabilities were relatively lower but most were still above 0.60 in both of the two Chinese samples (M = 0.73 and 0.71) and the two U.S. samples (M = 0.75 and 0.75). The shape of the reliability profiles was also very similar between demographically similar samples ($r_{college\text{-}college}$ = 0.65, $r_{internet\text{-}employee}$ = 0.69) and between the two Chinese samples (r = .94). In general, the scores derived from the Chinese BFI-2 had comparable reliabilities to those of the original U.S. samples.

Structural Validity

Domain-level structure. Standardized factor loadings for item-level ESEM and ESEMNW in each sample can be found in the online supplementary Tables S3 to S4. The facet-level loadings are presented in Table 2. Model fit, Tucker's congruence coefficients, and model-based interfactor correlations are displayed in Table 3 to 5.

Regarding item-level results, two patterns were clear. First, the ordinary ESEM displayed very similar and adequate fit in the two Chinese samples and the two U.S. samples. The majority of the items had moderate to large loadings on their intended primary factors. Factor loading patterns between the Chinese college sample and the U.S. college sample were similar, with congruence coefficients ranging from .88 to .95. The patterns were also similar between the Chinese employee sample and the U.S. internet sample, with congruence ranging from .87 to .94. The Chinese college sample and employee sample also displayed very similar loading patterns, with congruence ranging from .92 to .97. However, we noted that there were several substantial cross-loadings (absolute values greater than .30) in the employee sample. For example, four agreeableness items cross-loaded on neuroticism and three neuroticism items cross-loaded on conscientiousness. Interestingly, all seven of these cross-loading items were negatively worded.

Relatedly, a second clear pattern was that, compared with ESEM, ESEMNW substantially improved model fit in all samples (ΔCFI = .03, ΔTLI = .03, ΔRMSEA = −.01). Moreover, modeling the negative wording method factor reduced the size of cross-loadings and slightly increased primary loadings in the employee sample, thereby making the loading patterns between the U.S. internet sample and the Chinese employee sample more similar, with congruence now ranging from .89 to .94. The addition of the negative wording factor had almost no impact on primary loadings (an average decrease of .01) in the Chinese student sample. An inspection of loadings on the negative wording factor revealed that individual differences in responding to negative items exerted a stronger influence on loadings in the employee sample than in the college student sample (Mean loadings = .31 and .26). Model-based interfactor correlations were generally low in both models across samples (mean absolute rs of .04-.28), indicating good distinction among the five factors.

As for facet-level ESEM, it fitted the data very well in all four samples (CFI = .97-.99, TLI = .91-.98, RMSEA = .03-.06). Moreover, the Big Five factors were very well recovered in the two Chinese samples and the two U.S. samples. All facets loaded substantially on their intended factors. In the Chinese college student sample, there were only two cross-loadings greater than .30. In the employee sample, there were four cross-loadings just above .30. Loading patterns were almost identical between the Chinese employee sample and the U.S. internet sample (with congruence coefficients ranging from .96 to .98), and between the Chinese and U.S. college samples (with congruence ranging from .95 to .98). Model-based interfactor correlations were slightly higher than those found in item-level models (with mean absolute rs of .05-.39), but still demonstrated good distinction among the five factors.

Facet-Level Structure. We expected that the BFI-2 would show a robust multidimensional structure not only at domain level but also at the facet level within each domain. Results from a series of CFA models confirmed this hypothesis. For the sake of space, detailed model fit information for M1 to M7 can be found in the online supplementary Table S5 and standardized factor loadings for M5 to M7 can be found in online supplementary Table S6.

These models revealed several noteworthy findings. First, consistent with the BFI-2's multifacet design, all models that did not differentiate among facets (M1-M4) did not fit the data well across the five domains and the two samples, evidencing multidimensionality within each Big Five domain. Second, modeling the three facets within each domain substantially improved fit. For extraversion, the three models (M5-M7) that differentiated among facets displayed similarly good fit to the data in both the college student and adult employee samples, regardless of whether the models accounted for acquiescence or negative item wording effects. Factor loadings were moderate to large (Msoc = .77, Mast = .64, Meng = .70) and very similar

Table 2. Transposed Standardized Factor Loadings From Facet-Level ESEM.

Sample	Factor	SOC	AST	ENG	COM	REP	TRU	ORG	PRO	RES	ANX	DEP	EMO	INT	AES	CRE
U.S. internet	E	.86	.62	.69	.14	−.18	.07	−.05	.12	−.04	−.01	−.29	.20	−.05	−.07	.15
	A	.05	−.25	.20	.77	.74	.63	−.06	.00	.17	.02	−.02	−.05	−.01	.08	.01
	C	−.13	.19	.08	.06	.14	−.11	.77	.80	.67	.11	−.10	−.09	−.04	−.06	.01
	N	−.01	−.08	−.02	.19	−.06	−.23	.03	−.02	−.07	.86	.67	.82	.02	.06	−.08
	O	−.12	.19	.05	.07	.02	−.01	−.08	.03	−.04	.01	.02	−.04	.75	.67	.65
U.S. college	E	.90	.59	.70	.13	−.27	.12	−.08	.17	−.02	.00	−.25	.14	.04	−.08	.10
	A	.02	−.24	.20	.76	.66	.71	.00	−.04	.12	.06	−.11	−.07	.01	.12	−.03
	C	−.12	.19	.09	.07	.13	−.14	.70	.74	.69	.17	−.17	−.13	.09	−.08	−.06
	N	−.01	−.08	−.02	.20	−.11	−.21	.02	−.07	−.05	.85	.60	.76	.01	.12	−.15
	O	−.04	.19	.00	.11	.08	−.06	−.04	−.03	.06	−.08	.06	.01	.66	.66	.69
CN college	E	.85	.56	.61	.31	−.20	.07	−.07	.07	.05	−.12	−.16	.29	−.03	−.09	.23
	A	.09	−.16	.18	.61	.68	.63	−.11	.01	.35	.11	−.02	−.23	−.05	.10	−.05
	C	−.14	.23	.08	.11	.14	−.05	.80	.77	.49	.04	−.01	−.08	−.04	−.16	.12
	N	.01	−.06	−.14	.17	−.08	−.23	.01	−.06	.02	.84	.80	.60	−.04	.15	−.08
	O	−.07	.21	.12	.02	.07	−.05	−.05	−.04	.05	.00	.04	−.06	.77	.62	.59
CN employee	E	.71	.51	.63	.21	−.22	.01	−.07	.18	−.03	−.12	−.31	.31	.01	−.06	.29
	A	.02	−.32	.28	.69	.55	.65	−.10	.07	.30	.14	−.13	−.16	−.03	.15	−.04
	C	−.05	.28	.04	.12	.20	.01	.83	.75	.52	.00	.00	−.12	−.10	−.04	.09
	N	−.06	−.11	−.10	.21	−.13	−.30	.01	−.06	−.04	.82	.65	.66	−.06	.16	−.08
	O	.04	.22	.11	.03	.19	−.04	−.01	−.11	.09	.02	.04	−.11	.84	.59	.56
CN substance use	E	.73	.33	.54	.15	−.11	.08	−.01	.15	−.07	−.10	−.14	.17	−.01	−.05	.20
	A	.04	−.17	.20	.71	.63	.67	−.01	.12	.20	.09	.03	−.26	−.01	.07	−.01
	C	−.18	.29	.12	.16	.10	−.12	.72	.62	.34	−.03	.01	−.17	−.15	−.02	.23
	N	−.03	−.13	−.05	.16	−.12	−.18	−.06	−.06	−.25	.66	.81	.52	−.03	.09	.01
	O	−.07	.22	.12	.01	.06	.03	.05	−.02	.15	−.02	.02	−.01	.85	.60	.45
CN teenager T1	E	.90	.49	.67	.19	−.19	.10	−.06	−.04	.12	−.06	−.15	.16	.02	−.05	.18
	A	.05	−.15	.16	.70	.61	.64	−.04	−.03	.26	.14	.01	−.38	.01	.06	−.10
	C	−.15	.15	.08	.17	.07	−.06	.76	.72	.51	.10	−.16	.01	.15	−.13	.06
	N	.01	−.10	−.04	.12	−.19	−.15	.05	−.08	.00	.83	.66	.55	.03	.17	−.20
	O	−.02	.20	.02	−.06	.06	.03	.04	.00	.04	.00	.09	−.07	.52	.72	.48
CN teenager T2	E	.76	.52	.75	.24	−.14	.06	−.06	−.06	.10	−.11	−.21	.27	−.04	−.04	.22
	A	.03	−.23	.25	.60	.62	.62	−.01	−.01	.23	.12	−.01	−.27	.05	.06	−.10
	C	−.18	.28	.00	.17	.07	.01	.67	.60	.61	.09	−.20	.01	.08	−.09	.01
	N	−.07	−.04	−.02	.10	−.12	−.20	−.02	−.17	.05	.82	.65	.65	.04	.10	−.11
	O	.02	.15	.06	.00	.07	−.02	.04	.09	−.03	−.02	.12	−.07	.72	.64	.61

Note. ESEM = exploratory structural equation modeling; E = extraversion; A = agreeableness; C = conscientiousness; N = neuroticism; O = openness; SOC = sociability; AST = assertiveness; ENG = energy; COM = compassion; REP = respectful; TRU = trust; ORG = organization; PRO = productiveness; RES = responsibility; ANX = anxiety; DEP = depression; EMO = emotionality; Int = intellectual curiosity; AES = aesthetic sensitivity; CRE = creative imagination.

across the three models and the two samples. For agreeableness, a pure three-facet model did not fit the data well across the two samples (CFI = .87 and .89; TLI = .83 and .86; RMSEA = .12 and .13). Modeling an additional acquiescence factor (M6) improved fit substantially (ΔCFI = .03 and .05; ΔTLI = .03 and .05; ΔRMSEA = −.01 and −.02) in both samples. Modeling a negative wording factor had effects similar to modeling an acquiescence factor in the college sample (ΔCFI = .04; ΔTLI = .04; ΔRMSEA = −.02), but improved the fit in the adult employee sample even more effectively (ΔCFI = .06; ΔTLI = .08; ΔRMSEA = −.04). Despite fit differences across models, factor loadings were moderate to large (M_{com} = .66, M_{rep} = .62, M_{tru} = .57) and similar across models and samples. Similar patterns were also observed for conscientiousness and neuroticism such that either modeling an acquiescence or a negative wording factor improved model fit substantially compared with M5, with moderate to large item loadings on the facet factors. For openness, M5 to M7 fitted the data well according to their absolute values. Relatively speaking, M7 showed the best fit across the two samples (CFI = .98 and .98; TLI = .96 and .97; RMSEA = .08 and .07). Again, factor loadings were moderate to large (M_{int} = .58, M_{aes} = .77, M_{cre} = .77) and very similar across models and samples.

Table 3. Model Fit.

Model	Sample	Chi-square	*df*	CFI	TLI	RMSEA
ESEM item	U.S. internet	6458.41	1480	.878	.854	.058
	U.S. college	3903.16	1480	.862	.835	.059
	CN college	9039.88	1480	.845	.815	.065
	CN employee	3688.85	1480	.881	.858	.055
	CN substance use	3045.55	1480	.934	.921	.037
	CN teenager T1	2533.27	1480	.854	.826	.047
	CN teenager T2	2775.94	1480	.864	.837	.052
ESEMNW item	U.S. internet	5234.17	1450	.907	.887	.051
	U.S. college	3393.27	1450	.889	.865	.053
	CN college	7604.54	1450	.874	.846	.060
	CN employee	31.80	1450	.911	.892	.048
	CN substance use	2737.05	1450	.946	.934	.034
	CN teenager T1	2276.98	1450	.886	.861	.042
	CN teenager T2	2464.66	1450	.893	.870	.047
ESEM facet	U.S. internet	75.10	40	.994	.984	.030
	U.S. college	128.11	40	.967	.914	.068
	CN college	215.24	40	.970	.932	.061
	CN employee	68.73	40	.991	.976	.038
	CN substance use	97.07	40	.987	.966	.043
	CN teenager T1	54.25	40	.990	.974	.033
	CN teenager T2	64.36	40	.984	.958	.044

Note. ESEM = exploratory structural equation modeling; ESEMNW = ESEM with a negative wording factor; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; *df* = degrees of freedom.

Table 4. Tucker's Congruence Coefficients Among All Possible Pairs.

	ESEM item					ESEMNW item					ESEM facet				
Sample pair	E	A	C	N	O	E	A	C	N	O	E	A	C	N	O
U.S. internet college	.93	.91	.95	.96	.96	.93	.90	.95	.95	.96	.97	.96	.98	.97	.99
U.S. internet employee	.94	.90	.87	.89	.93	.94	.89	.93	.92	.94	.97	.96	.97	.98	.96
U.S. internet substance	.88	.85	.45	.78	.74	.87	.84	.78	.87	.87	.97	.97	.93	.95	.96
U.S. internet adolescent 1	.89	.81	.81	.91	.82	.91	.84	.90	.92	.90	.97	.95	.97	.97	.96
U.S. internet adolescent 2	.90	.81	.90	.89	.90	.90	.86	.90	.93	.91	.97	.96	.97	.98	.98
U.S. college–college	.92	.88	.91	.95	.95	.92	.89	.92	.95	.95	.96	.95	.95	.97	.98
U.S. college–employee	.93	.90	.83	.89	.93	.92	.90	.90	.92	.93	.96	.97	.94	.99	.96
U.S. college–substance	.87	.85	.44	.79	.74	.87	.84	.78	.87	.87	.97	.97	.88	.94	.95
U.S. college–adolescent 1	.88	.83	.81	.92	.81	.90	.85	.90	.92	.88	.97	.95	.97	.98	.96
U.S. college-adolescent 2	.89	.78	.90	.89	.90	.89	.86	.90	.93	.90	.95	.97	.97	.99	.97
College–employee	.97	.95	.93	.92	.97	.98	.95	.96	.96	.97	.98	.97	.99	.99	.99
College–substance	.92	.86	.85	.64	.91	.93	.87	.89	.88	.91	.96	.98	.97	.96	.98
College–adolescent 1	.93	.90	.88	.94	.87	.96	.93	.94	.95	.92	.98	.98	.97	.98	.96
College–adolescent 2	.95	.89	.93	.90	.94	.95	.93	.93	.94	.94	.98	.98	.95	.98	.98
Employee–substance	.92	.93	.73	.92	.79	.92	.90	.81	.91	.87	.97	.97	.96	.95	.98
Employee–adolescent 1	.91	.88	.94	.93	.86	.94	.90	.94	.94	.92	.94	.95	.94	.98	.94
Employee–adolescent 2	.94	.86	.89	.93	.93	.93	.91	.91	.93	.92	.96	.97	.94	.98	.96
Substance–adolescent 1	.94	.86	.73	.86	.85	.91	.84	.76	.89	.88	.96	.98	.90	.94	.94
Substance–adolescent 2	.94	.87	.59	.88	.81	.93	.90	.79	.89	.86	.96	.98	.89	.94	.96
Adolescent 1–adolescent 2	.97	.80	.90	.92	.94	.97	.93	.95	.94	.96	.98	.99	.98	.98	.97
Average	.92	.87	.81	.89	.88	.92	.89	.89	.92	.91	.97	.97	.95	.97	.97

Note. ESEM = exploratory structural equation modeling; ESEMNW = ESEM with a negative wording factor.

Table 5. Interfactor Correlations Among the Big Five Factors.

Model	Sample	AC	AE	AN	AO	CE	CN	CO	EN	EO	NO
Manifest	US1	.28	.14	−.29	.15	.22	−.30	−.02	−.34	.20	−.06
	US2	.28	.09	−.31	.26	.27	−.34	.14	−.36	.21	−.10
	CN1	.44	.27	−.39	.16	.30	−.38	.17	−.32	.33	−.07
	CN2	.51	.20	−.43	.33	.40	−.54	.35	−.44	.40	−.31
	CN3	.59	.35	−.57	.35	.46	−.64	.51	−.44	.45	−.40
	CN41	.39	.20	−.43	.08	.34	−.34	.32	−.34	.30	−.10
	CN42	.41	.18	−.45	.11	.40	−.41	.30	−.26	.31	−.09
ESEM item	US1	.18	.09	−.18	.14	.16	−.26	.05	−.25	.17	−.02
	US2	.21	.06	−.17	.20	.18	−.22	.15	−.27	.14	−.04
	CN1	.21	.05	−.22	.16	.18	−.33	.22	−.21	.19	−.06
	CN2	.21	.05	−.22	.16	.18	−.33	.22	−.21	.19	−.06
	CN3	.33	.21	−.32	.32	.27	.08	.11	−.07	.15	−.46
	CN41	.11	−.02	−.24	.09	.19	−.15	.22	−.18	.16	−.07
	CN42	.24	.09	−.15	.07	.24	−.20	.24	−.04	.16	−.05
ESEMNW item	US1	.16	.09	−.15	.14	.15	−.24	.04	−.26	.16	−.01
	US2	.19	.05	−.15	.19	.17	−.19	.14	−.26	.14	−.02
	CN1	.21	.04	−.20	.15	.17	−.32	.21	−.20	.18	−.05
	CN2	.21	.04	−.20	.15	.17	−.32	.21	−.20	.18	−.05
	CN3	.24	.17	−.38	.28	.12	−.42	.29	−.20	.23	−.29
	CN41	.15	.01	−.19	.03	.15	−.21	.30	−.22	.20	−.02
	CN42	.21	.06	−.20	.06	.23	−.22	.23	−.08	.17	−.04
ESEM-Facet	US1	.27	.09	−.27	.12	.21	−.29	.04	−.31	.21	−.06
	US2	.28	.04	−.26	.24	.23	−.30	.19	−.33	.19	−.09
	CN1	.37	.09	−.32	.16	.23	−.40	.28	−.28	.30	−.11
	CN2	.36	−.02	−.27	.27	.26	−.55	.45	−.32	.24	−.34
	CN3	.51	.14	−.56	.37	.31	−.58	.59	−.34	.38	−.48
	CN41	.35	.09	−.30	.05	.22	−.33	.40	−.36	.28	−.12
	CN42	.33	.02	−.35	.06	.27	−.39	.38	−.26	.28	−.13

Note. ESEM = exploratory structural equation modeling; ESEMNW = ESEM with a negative wording factor; US1 = U.S. internet sample; US2 = U.S. college sample; CN1 = Chinese college sample; CN2 = employee sample; CN3 = substance use sample; CN41 = adolescent sample at T1; CN42 = adolescent sample at T2.

In sum, these results indicate that the Chinese BFI-2 showed a clear Big Five structure, as well as a three-facet structure within each domain that closely paralleled the English language BFI-2. The results also show that accounting for individual differences in how participants responded to negatively worded items helps further clarify these structures.

Measurement Invariance. As the model with three correlated factors plus an orthogonal negative wording factor displayed the best fit to the data across samples and personality domains, we tested measurement invariance based on this model. In line with Steenkamp and Maydeu-Olivares (2020), factor loadings on the negative wording factor were freely estimated because we were only interested in the substantive personality facets. Model fit results are shown in Table 6.

Across the two pairs of samples (Chinese college vs. U.S. college samples, Chinese employee vs. U.S. internet samples), several patterns are noteworthy. First, all but one chi-square changes were significant, which is not surprising given our large sample size. Second, CFI and TLI indicated good fit of all models for the five personality factors. Most important, none of the changes in model fit exceeds the recommended cutoffs. We also calculated the *dmac* index (Nye & Drasgow, 2011) to quantify the degree of item-level measurement noninvariance. According to the empirically derived cutoff (Nye et al., 2019), *dmac* for most items was very small, with only one openness item ("Is curious about many different things") displayed a large effect size between the Chinese working adult sample and the U.S. internet sample. Details about these effect sizes can be found in the online supplementary Table S19. In sum, the BFI-2 is largely invariant across the two cultures.

Convergent and Discriminant Validity. Table 7 presents correlations for the test of convergent and discriminant validity with the Forced-Choice Five-Factor Model (FCFFM) scales. This table shows that the BFI-2 domain scores displayed moderate to strong convergence with the FCFFM

Table 6. Measurement Invariance.

Domain	Level of MI	Chi-square	*df*	CFI	TLI	RMSEA	ΔChi-square	*p*	ΔCFI	ΔTLI	ΔRMSEA
U.S. college students versus CN college students											
E	Configural	862.31	90	.961	.942	0.102					
	Threshold	963.70	114	.957	.950	0.095	64.144	<.001	−.004	.008	−.007
	Metric	1120.72	123	.949	.945	0.099	93.502	<.001	−.008	−.005	.004
A[a]	Configural	648.32	70	.906	.853	.100					
	Threshold	710.19	88	.899	.874	.092	44.562	<.001	−.007	.021	−.008
	Metric	624.41	96	.914	.902	.081	11.429	0.179	.015	.028	−.011
C	Configural	1291.81	90	.925	.890	.127					
	Threshold	1423.75	114	.918	.905	.118	73.519	<.001	−.007	.015	−.009
	Metric	1314.27	123	.925	.920	.108	20.186	.017	.007	.015	−.010
N	Configural	781.78	90	.964	.948	.096					
	Threshold	852.77	114	.962	.956	.088	45.862	0.05	−.002	.008	−.008
	Metric	887.73	123	.961	.958	.086	44.343	<.001	−.001	.002	−.002
O	Configural	556.80	90	.972	.959	.079					
	Threshold	626.32	114	.969	.964	.074	45.262	.005	−.003	.005	−.005
	Metric	673.56	123	.967	.964	.073	40.511	<.001	−.002	.000	−.001
U.S. internet sample versus CN working adults											
E	Configural	606.90	90	.966	.951	.088					
	Threshold	721.39	114	.960	.954	.085	82.139	<.001	−.006	.003	−.003
	Metric	775.46	123	.957	.954	.085	51.292	<.001	−.003	.000	.000
A	Configural	453.10	90	.965	.948	.074					
	Threshold	559.54	114	.956	.950	.073	84.463	<.001	−.009	.002	−.001
	Metric	633.98	123	.950	.946	.075	54.395	<.001	−.006	−.004	.002
C	Configural	917.32	90	.948	.924	.111					
	Threshold	1045.28	114	.941	.932	.105	101.481	<.001	−.007	.008	−.006
	Metric	1028.70	123	.943	.939	.100	40.599	<.001	.002	.007	−.005
N	Configural	716.43	90	.968	.952	.097					
	Threshold	823.73	114	.963	.957	.092	75.966	<.001	−.002	−.001	−.009
	Metric	894.77	123	.960	.957	.092	57.159	<.001	−.003	.000	.000
O	Configural	377.46	90	.976	.965	.066					
	Threshold	464.86	114	.971	.967	.064	71.856	<.001	−.005	.002	−.002
	Metric	501.93	123	.969	.967	.064	34.766	<.001	−.002	.000	.000

[a]As no one chose "strongly disagree" for the item "Is respectful, treats others with respect" in the U.S. college student sample, we had to exclude this item before testing measurement invariance for ordinal data.

scores ($r_E = .74$, $r_C = .73$, $r_A = .48$, $r_N = .76$, $r_O = .57$). The relatively low convergent validity for agreeableness and openness likely reflects differences in content coverage between the BFI-2 and the FCFFM. Specifically, most agreeableness items in the FCFFM focused on compassion for other people ($r = .49$ with the BFI-2 compassion facet), rather than respectfulness ($r = .29$) or trust ($r = .36$). Similarly, most openness items in the FCFFM focused on creativity ($r = .62$ with the BFI-2 creative imagination facet) and intellectual curiosity ($r = .53$), rather than aesthetic sensitivity ($r = .23$). As for discriminant validity, BFI-2 agreeableness (absolute $r = .19$-$.36$), conscientiousness (absolute $r = .20$-$.43$), neuroticism (absolute $r = .27$-$.28$), and openness (absolute $r = .01$-$.33$) showed low to moderate correlations with the other four domain scores in the FCFFM. BFI-2 extraversion had a relatively large correlation with FCFFM agreeableness ($r = .57$); however, this was still lower than its corresponding convergent validity ($r = .74$). Overall, these results support the convergent and discriminant validity of the Chinese BFI-2.

Criterion-Related Validity. Criterion-related validity coefficients for the BFI-2 are presented in Table 7 (college sample) and Table 8 (employee sample). Results for each criterion subscale are presented in the online supplementary Tables S7 to S8. We also present results from multiple regressions in which either the 5 domain scores or the 15 facet scores were used to predict criteria in the online supplementary Tables S10 to S11 so that interested readers can identify unique predictors for each criterion.

Generally, across the two samples, neuroticism, agreeableness, extraversion, and conscientiousness displayed significant associations with various psychological and behavioral outcomes, as well as mental and physical health

Table 7. Criterion-Related Validity (Chinese College Students).

	E				A				C				N				O			
Criterion	E	SOC	ASS	ENG	A	COM	REP	TRU	C	ORG	PRO	RES	N	ANX	DEP	EMO	O	INT	AES	CRE
FC-E	.74	.72	.54	.57	.30	.34	.10	.26	.20	.07	.22	.22	−.27	−.32	−.32	−.07	.28	.22	.08	.39
FC-A	.57	.46	.45	.51	.48	.49	.29	.36	.43	.29	.36	.41	−.27	−.25	−.27	−.17	.33	.30	.12	.38
FC-C	.15	−.03	.24	.20	.20	.14	.19	.14	.73	.70	.65	.40	−.28	−.18	−.25	−.26	.08	.09	−.06	.17
FC-N	−.15	−.05	−.11	−.24	−.36	−.14	−.37	−.37	−.32	−.26	−.25	−.27	.76	.57	.60	.73	−.01	−.04	.09	−.09
FC-O	.48	.30	.50	.43	.19	.23	.08	.15	.39	.29	.38	.29	−.28	−.30	−.27	−.14	.57	.53	.23	.62
Gender	−.05	−.01	−.08	−.04	−.01	−.02	.02	−.02	−.03	−.03	−.07	.03	.19	.16	.11	.20	−.01	−.13	.17	−.12
Edu-F	.12	.08	.10	.13	.03	.02	.02	.03	.04	.00	.06	.06	−.07	−.07	−.07	−.05	.17	.16	.12	.12
Edu-M	.12	.09	.11	.11	.02	.03	.02	.00	.03	.00	.03	.04	−.04	−.06	−.04	−.01	.21	.19	.15	.15
SWB	.27	.18	.21	.30	.22	.13	.17	.22	.24	.13	.26	.20	−.42	−.40	−.44	−.22	.09	.05	−.01	.18
PS	−.30	−.14	−.25	−.37	−.31	−.18	−.26	−.31	−.39	−.27	−.39	−.30	.68	.57	.62	.51	−.12	−.10	.07	−.26
DE	−.39	−.24	−.28	−.47	−.30	−.25	−.20	−.26	−.33	−.23	−.32	−.28	.50	.40	.55	.31	−.16	−.11	−.02	−.26
GA	−.16	−.07	−.16	−.18	−.23	−.18	−.22	−.15	−.38	−.31	−.34	−.28	.09	.01	.10	.10	−.06	−.01	−.08	−.04
AG	−.06	.01	.00	−.17	−.48	−.21	−.54	−.41	−.29	−.18	−.22	−.32	.48	.28	.37	.55	−.10	−.09	−.06	−.10
PH	−.15	−.08	−.06	−.24	−.21	−.14	−.14	−.22	−.19	−.13	−.18	−.15	.40	.34	.39	.27	.01	.00	.09	−.08
CM	.00	.03	−.02	−.03	−.08	−.07	−.07	−.05	−.12	−.09	−.12	−.09	.12	.08	.13	.09	.04	.04	.06	−.01
DT-MA	.08	.11	.14	−.04	−.39	−.26	−.30	−.37	−.24	−.12	−.20	−.29	.15	.07	.12	.16	.04	.07	.05	−.03
DT-PP	−.20	−.15	−.07	−.27	−.56	−.52	−.36	−.45	−.27	−.13	−.19	−.34	.18	.08	.19	.17	−.05	.02	−.02	−.11
DT-NA	.24	.20	.28	.13	−.06	.05	−.08	−.13	−.02	.00	−.05	.00	.11	.06	.04	.16	.10	.07	.05	.12

Note. $||r|| \geq .06$ is statistically significant at $\alpha = .05$. FC = forced choice; E = extraversion; A = agreeableness; C = conscientiousness; N = neuroticism; O = openness; Edu_F = father's highest degree of education; Edu_M = mother's highest degree of education; SWB = subjective well-being; PS = perceived stress; DE = depression; GA = game addiction; AG = aggression; PH = physical health (higher score means lower level of health); CM = childhood maltreatment; DT = dark triad; MA = Machiavellianism; PP = psychopathy; NA = narcissism.

Table 8. Criterion-related Validity (Adult Employee Sample).

	E				A				C				N				O			
Criterion	E	SOC	ASS	ENG	A	COM	REP	TRU	C	ORG	PRO	RES	N	ANX	DEP	EMO	O	INT	AES	CRE
Gender	−.01	.04	−.06	−.01	.03	.04	.01	.03	−.02	−.04	−.03	.03	.16	.15	.08	.16	−.01	−.09	.12	−.10
Age	.07	.02	.15	.00	.06	.13	−.02	.02	.27	.19	.28	.19	−.10	−.11	−.10	−.04	−.08	−.07	−.11	.00
Income	.19	.09	.29	.10	.05	.03	.06	.03	.24	.14	.24	.21	−.21	−.20	−.17	−.15	.18	.23	−.03	.27
JP	.15	.04	.18	.16	.10	.10	.10	.04	.21	.18	.16	.20	−.10	−.11	−.13	−.03	.19	.18	.08	.20
Status	.11	.06	.20	.03	−.09	−.04	−.07	−.10	.09	.09	.10	.03	−.03	−.09	−.04	.05	.05	.04	.00	.10
Jsat	.19	.18	.13	.15	.26	.13	.21	.29	.21	.13	.21	.17	−.34	−.27	−.32	−.26	.12	.15	−.03	.20
OCB	.34	.22	.31	.31	.42	.37	.33	.34	.40	.25	.41	.35	−.26	−.22	−.24	−.20	.32	.28	.17	.34
CWB	−.09	−.04	−.02	−.16	−.46	−.24	−.47	−.43	−.35	−.23	−.29	−.37	.36	.24	.24	.40	−.17	−.17	−.06	−.19
SWB	.24	.19	.19	.21	.15	.10	.07	.20	.20	.15	.23	.12	−.33	−.31	−.36	−.18	.10	.13	−.02	.15
PS	−.31	−.15	−.35	−.28	−.33	−.11	−.32	−.38	−.40	−.27	−.35	−.41	.66	.57	.54	.54	−.27	−.30	−.05	−.33
DE	−.42	−.23	−.32	−.50	−.39	−.22	−.36	−.38	−.40	−.26	−.34	−.41	.54	.41	.56	.40	−.29	−.31	−.05	−.37
AG	−.15	−.06	−.08	−.24	−.52	−.22	−.58	−.49	−.40	−.26	−.33	−.44	.56	.36	.40	.61	−.26	−.24	−.15	−.23
PH	−.15	−.06	−.12	−.19	−.19	−.08	−.23	−.17	−.17	−.10	−.13	−.22	.36	.33	.29	.29	−.15	−.18	−.02	−.18
DT-MA	.06	.07	.16	−.08	−.41	−.30	−.33	−.38	−.27	−.15	−.22	−.34	.17	.11	.07	.23	−.07	−.01	−.08	−.05
DT-PP	−.09	−.04	.02	−.18	−.58	−.48	−.45	−.49	−.32	−.17	−.24	−.43	.20	.07	.15	.27	−.21	−.14	−.20	−.16
DT-NA	.21	.13	.21	.18	−.05	.05	−.08	−.10	−.05	.00	−.05	−.09	.05	.07	−.03	.07	.15	.16	.09	.11

Note. $||r|| \geq .09$ is statistically significant at $\alpha = .05$. Jsat = job satisfaction; JP = job performance; OCB = organizational citizenship behavior; CWB = counterproductive work behavior; SWB = subjective well-being; PS = perceived stress; DE = depression; AG = aggression; PH = physical health; DT = dark triad; MA = Machiavellianism; PP = psychopathy; NA = narcissism.

measures, with both direction and magnitude as expected (details are presented in the supplement).

Regarding criterion measures that were sample-specific, in both the college student and the employee samples, all of the Big Five factors were significantly related to education/school related outcomes and job outcomes consistent with our expectation (see the supplement for details).

Regression coefficients largely supported these trait-criterion associations, although controlling for overlap between domain and facet traits reduced the strength of

some associations (see the online supplementary Tables S10-S11). In sum, the Chinese BFI-2 showed replicable criterion-related validities that were largely consistent with previous literature. These results indicate that the Chinese BFI-2 captures the Big Five traits' nomological network of trait-criterion associations.

Study 2

Study 1 provided strong evidence for the psychometric soundness of the Chinese BFI-2 in a college student sample and an adult employee sample. However, it is unclear whether these findings would be extended to less-studied populations, such as young adolescents and clinical populations, for whom personality has been shown to be relevant to important life outcomes (Chernyshenko et al., 2018; Kotov et al., 2010; Poropat, 2009). Therefore, Study 2 was conducted to examine the performance of the Chinese BFI-2 in a sample of secondary school students and a sample of adults receiving treatment for substance abuse. We aimed for at least 300 valid respondents in each sample so that ratios between the number of items and sample sizes were at least 1:5. The study design and data collection procedure were approved by the institutional review board at Beijing Normal University.

Participants and Procedures

Substance Abuse Sample. Seven hundred and sixty-five respondents from two substance abuse treatment institutions located in two Chinese cities were recruited. The clients in these institutions had a history of substance abuse that affected their functioning in everyday life. Staff members familiar with psychological testing administered the printed questionnaires to respondents in group sessions. Participants were informed that their responses were anonymous and would only be used for research purpose. There were 138 respondents who completed 6th grade, 397 respondents who completed 9th grade, 139 respondents who completed 12th grade, and 44 respondents who had some degree of college education, and 47 respondents who had a bachelor's degree or above. Respondents ranged in age from 18 to 88 years, with most in their 30s or 40s (M_{age} = 37.57, SD_{age} = 11.33). Due to privacy concerns, we did not collect information about respondents' specific substance problems, nor about gender (because the clients at both institutions were predominantly men). There were 383 respondents who missed more than 1 out of 8 quality control items.[4] However, as the test conducted in this sample was intended to evaluate the performance of the Chinese BFI-2 in the lower educated and at-risk populations (and thus are likely to have more problems in concentrating and complying with the test instruction), we reported results based on the full sample in the manuscript and results based on the "cleaned" sample can be found in the online supplementary Table S14 to S18.

Adolescent Sample. Three hundred and sixty-nine 10th-grade students were recruited from a high school in China. A psychology teacher administered printed questionnaires during class. Three months later, the same questionnaires were administered to the same students for estimating test–retest reliability. Students were told that their aggregated responses to the personality questionnaire would be used as teaching material and the psychology teachers would teach them how to interpret their own assessment results. After excluding respondents who missed more than 1 out of 3 quality control items on both occasions, 315 respondents were retained. Most participants were 15 or 16 years old (M = 16.20, SD = 0.61), and the sample included an approximately equal number of boys (n = 152) and girls (n = 163).

Measures

To replicate the criterion-related validity associations obtained in Study 1, we assessed subjective well-being, perceived stress, depression, aggression, physical health, dark triad, and childhood maltreatment in the substance abuse sample using the same measures administered in Study 1. In addition, we also measured parents' education, vocational interests (Rounds et al., 2016), Machiavellianism (Christie & Geis, 1970), counterproductive work behaviors toward individuals and organizations (Bennett & Robinson, 2000), and adverse childhood experiences (Finkelhor et al., 2015) in the substance abuse sample. In the adolescent sample, we measured deviant academic behaviors (homemade), bullying behaviors and experience of being bullied (Espelage & Holt, 2001), academic stress (Ang & Huan, 2006), procrastination (Özer et al., 2013), and academic achievement. More details can be found in the online supplementary Table S1.

Results and Discussion

Descriptive Statistics. Means and standard deviations of the Big Five and the facets for the substance abuse sample and adolescent sample are shown in Table 1. The mean personality profiles of the substance abuse sample and adolescent sample were moderately similar to those of the Chinese college student sample (r = .70 - .81) and the adult employee sample (r = .59-.95). The profiles were also similar to the mean profiles of the two U.S. samples (Mr = .64, range = .55-.79).

Reliability. As can be seen in Table 1, the BFI-2 displayed high reliability across the two samples at the domain level (α = .72-.80 in the substance abuse sample and .75-.83 in the adolescent sample; test–retest reliability = .70-.80 in

the adolescent sample). As in Study 1, facet-level reliabilities were on average lower (Cronbach's α: $M_{substance}$ = .55, $M_{adolescent}$ = .64; test–retest: M = .65). However, some facets showed particularly low reliability. For example, the intellectual curiosity facet of openness to experience had very low reliability across the two samples (α = .39 in the adolescent sample at both Time 1 and Time 2 and .28 in the substance abuse sample; test–retest reliability = .51 in the adolescent sample), as did the respectfulness and trust facets of agreeableness (α = .47-.52 in the adolescent sample, and .46-.55 in the substance abuse sample). The assertiveness and energy facets of extraversion also had relatively low reliability in the substance abuse sample (α = .44 and .47). These results suggest that some BFI-2 facet scales may be difficult for respondents with low levels of formal education to complete reliably.

Structural Validity

Domain-level structural validity. As shown in Table 2, ESEM of the 15 BFI-2 facet scales revealed a very clear five-factor structure in both samples with excellent model fit, moderate to large primary loadings, small cross-loadings, and good distinction among the five factors.

As for item-level ESEM, Table 3 shows that the BFI-2 showed reasonable fit to the data in the two samples. As shown in Table 4, the five factors were reasonably recovered in the adolescent sample, except that the conscientiousness factor at Time 1 was quite broad, with 11 items from other Big Five domains showing substantial cross-loadings onto this factor (in the online supplementary Table S3). However, the recovery of the Big Five factors in the substance abuse sample was less clear. Although the extraversion and openness factors were reasonably intact, the neuroticism, agreeableness, and conscientiousness factors were less clearly recovered. For each of these factors, some items did not show their expected primary loadings, and several items from other factors showed substantial cross-loadings. Notably, most of the problematic items were negatively worded.

As would be expected given this pattern of results, explicitly modeling a negative wording factor improved model fit substantially. Moreover, it reduced many problematic cross-loadings below .30 and increased factor congruence, presenting a clearer five-factor structure in both samples.

Facet-level structural validity. Successfully replicating the results of Study 1, supplementary Table S4 (available online) shows that the three-facet structure within each domain was generally supported in the adolescent sample. However, the pattern was more complicated in the substance use sample. The three-facet structure of extraversion, conscientiousness, and openness was supported because M6 and M7 showed similarly good fit compared with the remaining models. For agreeableness, modeling a single domain factor plus a negative wording factor (M3) fitted the data as well as M7; however, interfactor correlations among the three facet factors ranged from .85 to .90, thus weakening the support for a three-facet structure of agreeableness. For neuroticism, although M6 and M7 displayed excellent fit, the correlation between anxiety and depression approached 1.00, which did not support a well-differentiated three-facet structure.

Taken together, these results indicate that, in younger, less-educated, and at-risk populations, (a) the Chinese BFI-2's intended Big Five structure can be clearly recovered from its 15 facet scales, (b) this structure is more difficult to recover at the item level, but (c) explicitly modeling individual differences in how people respond to negatively worded items substantially clarifies the item-level structure. At the facet level, the current results suggest that (a) the BFI-2's intended hierarchical structure can be clearly recovered in adolescent samples, but that (b) some facets—especially within the agreeableness and neuroticism domains—may be difficult to distinguish in clinical samples with relatively low education level.

Criterion-Related Validity. Criterion-related validity coefficients in the two samples are shown in Table 9 and 10, illustrating several noteworthy patterns. First, the relationships between the Big Five factors and gender, parental education, subjective well-being, perceived stress, depression, aggression, physical health, dark triad, counterproductive work behaviors (despite the different measures used in Study 1 vs. Study 2), and childhood maltreatment (across two measures) were very similar to what we observed in Study 1 (average correlations among vectors of validity across samples were .87, min = .53, median =.95, max =.98). These findings show that trait-criterion relations proved quite replicable and generalizable across the four samples.

Study 2 also included a number of additional criteria beyond those included in Study 1. For example, in the substance abuse sample, the Big Five factors demonstrated associations with different types of vocational interests consistently with the findings suggested by previous meta-analytic reviews (Mount et al., 2005). Meanwhile, in the adolescent sample, the five personality factors exhibited significant associations with school-related behaviors, as well as academic stress (see the online supplement for detailed description).

Regression results are presented in the online supplementary Tables S12 and S13. These analyses largely support the pattern of trait-criterion associations summarized above, even while controlling for overlap between domain and facet traits. Overall, these correlation and regression results provide further evidence for the criterion validity of the Chinese BFI-2.

Table 9. Criterion-Related Validity (Substance Use Sample).

	E				A				C				N				O			
Criterion	E	SOC	ASS	ENG	A	COM	REP	TRU	C	ORG	PRO	RES	N	ANX	DEP	EMO	O	INT	AES	CRE
Age	−.12	−.13	.03	−.17	.15	.14	.16	.10	.14	.14	.12	.10	−.15	−.11	−.06	−.16	.06	.00	.08	.02
Edu-F	.13	.09	.14	.08	.03	.08	.02	−.03	−.05	.01	−.10	−.04	−.05	−.07	−.02	−.01	.13	.08	.10	.12
Edu-M	.14	.14	.08	.08	.05	.08	.02	.01	−.06	−.05	−.08	−.02	−.05	−.07	−.03	.00	.12	.09	.07	.11
SWB	.12	.04	.11	.10	.12	.08	.11	.08	.16	.10	.13	.13	−.23	−.14	−.21	−.16	.09	.06	.05	.09
PS	−.34	−.10	−.33	−.37	−.39	−.26	−.38	−.29	−.49	−.40	−.44	−.40	.63	.50	.53	.52	−.33	−.24	−.18	−.35
DE	−.35	−.13	−.31	−.40	−.36	−.25	−.35	−.30	−.44	−.35	−.40	−.36	.58	.41	.54	.47	−.30	−.28	−.16	−.29
AG	−.25	−.03	−.27	−.29	−.54	−.41	−.53	−.43	−.50	−.40	−.39	−.47	.60	.36	.49	.63	−.27	−.21	−.19	−.23
PH	−.26	−.13	−.18	−.29	−.21	−.15	−.20	−.19	−.21	−.13	−.20	−.22	.43	.39	.39	.29	−.15	−.17	−.06	−.12
DT-MA	.04	.13	−.03	−.08	−.43	−.33	−.36	−.38	−.35	−.25	−.27	−.39	.30	.14	.22	.37	−.17	−.12	−.17	−.10
DT-PP	−.29	−.13	−.20	−.37	−.52	−.47	−.40	−.41	−.43	−.34	−.36	−.38	.40	.20	.36	.41	−.22	−.15	−.15	−.23
DT-NA	.08	.06	.08	.05	−.11	−.05	−.15	−.15	−.12	−.10	−.13	−.14	.16	.10	.14	.18	.05	.03	.03	.08
MA	−.30	−.13	−.23	−.37	−.51	−.41	−.38	−.44	−.46	−.32	−.39	−.44	.42	.30	.37	.38	−.29	−.21	−.22	−.25
CM	−.06	.04	−.10	−.08	−.22	−.18	−.24	−.20	−.24	−.13	−.22	−.26	.28	.23	.24	.26	−.13	−.09	−.12	−.07
AE	−.05	.05	−.10	−.05	−.24	−.15	−.21	−.21	−.21	−.12	−.16	−.21	.28	.21	.21	.23	−.14	−.11	−.11	−.06
VI-R	.12	.02	.09	.18	.14	.13	.08	.12	.30	.21	.29	.24	−.14	−.10	−.15	−.15	.21	.15	.14	.20
VI-I	.16	.01	.17	.21	.14	.14	.09	.12	.28	.24	.24	.23	−.16	−.15	−.14	−.16	.33	.26	.27	.26
VI-A	.23	.07	.24	.23	.17	.17	.13	.11	.23	.20	.18	.18	−.12	−.16	−.07	−.12	.50	.34	.48	.33
VI-S	.24	.07	.25	.28	.28	.29	.18	.17	.31	.23	.24	.27	−.16	−.15	−.13	−.17	.40	.33	.34	.28
VI-E	.33	.16	.29	.31	.29	.29	.19	.19	.34	.26	.28	.29	−.23	−.20	−.19	−.19	.33	.27	.25	.28
VI-C	.18	.06	.13	.20	.23	.18	.17	.19	.32	.21	.30	.24	−.19	−.15	−.15	−.21	.29	.20	.24	.24
CWB-I	−.01	.10	−.05	−.07	−.40	−.35	−.38	−.32	−.37	−.27	−.34	−.34	.29	.16	.21	.34	−.18	−.15	−.16	−.11
CWB-O	−.18	−.05	−.17	−.20	−.42	−.39	−.35	−.33	−.49	−.36	−.45	−.41	.36	.25	.30	.36	−.23	−.17	−.18	−.19

Note. $||r|| \geq .07$ is statistically significant at $\alpha = .05$; Edu-F = education level of father; Edu-M = education level of mother; SWB = subjective well-being; PS = perceived stress; DE = depression; AG = aggression; PH = physical health; DT = dark triad; MA = Machiavellianism; PP = psychopathy; NA = narcissism; CM = childhood maltreatment; AE = adverse childhood experience; VI = vocational interest; R = realistic; I = investigative; A = artistic; S = social; E = enterprising; C = conventional; CWB-I = counterproductive work behavior-individual; CWB-O = counterproductive work behavior-organization.

Table 10. Criterion-Related Validity (Adolescent).

	E				A				C				N				O			
Criterion	E	SOC	AST	ENG	A	COM	REP	TRU	C	ORG	PRO	RES	N	ANX	DEP	EMO	O	INT	AES	CRE
Gender	.08	.06	.05	.09	−.04	−.01	−.04	−.03	−.06	−.10	−.08	.04	.14	.14	.12	.09	.08	−.03	.26	−.11
Edu_F	.04	.03	.02	.03	.06	.02	.12	.01	−.02	−.03	−.05	.05	−.01	.03	.02	−.07	.12	.07	.14	.05
Edu_M	.07	.04	.04	.11	.02	.00	.08	−.03	−.06	−.04	−.06	−.03	−.06	−.05	.02	−.11	.12	.06	.12	.06
AA	.06	.03	.06	.06	−.04	−.03	.09	−.13	−.03	−.07	−.04	.05	−.01	.00	.02	−.04	.10	.11	.12	−.01
DB-T1	.23	.27	.14	.12	−.16	−.15	−.15	−.08	−.24	−.19	−.24	−.15	.03	−.10	.01	.13	.00	−.01	−.08	.09
BU-T1	.00	.06	.00	−.07	−.24	−.17	−.22	−.20	−.19	−.11	−.15	−.21	.12	.03	.04	.19	−.08	−.05	−.17	.06
VB-T1	−.08	−.05	−.05	−.10	−.14	−.04	−.13	−.17	−.12	−.05	−.08	−.16	.12	.10	.07	.11	.03	.03	.04	.01
AS-T1	−.03	−.05	.02	−.03	.02	.10	−.05	−.02	.15	.15	.15	.07	.22	.28	.19	.09	.11	.16	.13	−.05
PR-T1	−.12	.02	−.14	−.20	−.27	−.23	−.18	−.24	−.51	−.38	−.50	−.35	.24	.18	.27	.15	−.06	−.10	.00	−.06
DB-T2	.21	.26	.12	.12	−.14	−.15	−.13	−.05	−.19	−.16	−.19	−.11	.03	−.11	.00	.14	−.03	−.04	−.09	.08
BU-T2	.02	.07	.01	−.03	−.24	−.15	−.22	−.22	−.18	−.09	−.15	−.19	.09	.05	−.02	.17	−.05	−.01	−.15	.08
VB-T2	−.13	−.10	−.07	−.15	−.13	−.06	−.11	−.13	−.06	−.02	.00	−.13	.05	.05	.02	.05	.06	.04	.04	.06
AS-T2	−.01	−.07	.04	.02	.00	.04	−.04	.00	.14	.14	.15	.05	.18	.23	.16	.07	.15	.19	.18	−.05
PR-T2	−.13	−.04	−.18	−.14	−.21	−.19	−.16	−.15	−.37	−.26	−.40	−.25	.22	.12	.23	.17	−.04	−.03	−.02	−.05

Note. $||r|| \geq .11$ is statistically significant at $\alpha = .05$. AA = academic achievement; DB = deviant behavior; BU = bully; VB = victim of bully; AS = academic stress; PR = procrastination.

General Discussion

The present study adapted the BFI-2 to the Chinese language and cultural context, and then carried out a comprehensive psychometric examination in four diverse samples. The inventory's intended five-factor structure was generally well recovered in these samples, with factor loading patterns similar to those found in two U.S. comparison samples. The three-facet structure of each domain was also

largely confirmed. The Chinese BFI-2 also demonstrated good reliability, especially at the domain level, satisfactory convergent validity and discriminant validity, and excellent criterion-related validity with a variety of important criteria related to health, work, education, and counseling. However, the results also suggested some cases in which researchers and practitioners should exercise caution when interpreting results from the Chinese BFI-2.

It is encouraging to find that the Big Five factors could be well recovered in all Chinese samples, and that the factor solutions were almost identical to two U.S. comparison samples when we used facet scores as indicators (average TCC among all possible sample pairs were .96 ~ .97). All facet indicators had high loadings on their intended primary factors and low cross-loadings on other factors, indicating that the Big Five structure is well captured by the Chinese BFI-2. The Chinese BFI-2 also largely overcame the reliability issue that was presented in many other translated personality measures in the Chinese context (Luo & Dai, 2011) in that the domain-level reliabilities of the Chinese BFI-2 were quite similar to those in other cultures. In addition, the Chinese BFI-2's domain scales showed good convergent validity with scores obtained from a forced-choice Big Five measure, the FCFFM. We chose a forced-choice measure for studying convergent validity because forced-choice measures are resistant to idiosyncratic response styles (Cao & Drasgow, 2019; Zhang, Sun, Drasgow, et al., 2020). Therefore, the estimates we obtained with the FCFFM should provide a stricter test of convergent validity as compared with other measures that use a Likert-type response scale. Moreover, the Chinese BFI-2's domain and facet scales showed significant associations with various important outcomes—including well-being, depression, stress, health, job satisfaction, organizational citizenship behaviors, counterproductive work behaviors, and job performance—and individual difference variables—including vocational interests and dark personality traits—in ways that were largely consistent with previous meta-analytic findings (e.g., Allen & Walter, 2018; Anglim et al., 2020; Barrick et al., 2005; Berry et al., 2007; Bogg & Roberts, 2004; Judge et al., 2002; Judge et al., 2013; Muris et al., 2017; Organ, & Ryan, 1995). Apart from these subjective measures, extraversion and conscientiousness were also found to be related to more objective outcomes like income and workplace status in ways consistent with previous findings (e.g., Lu et al., 2020; Ng & Feldman, 2010). In the adolescent sample, the five personality factors and their facet traits also showed consistent concurrent and prospective relationships with deviant behaviors, bullying behaviors, academic stress, and procrastination. In addition, across the college student sample, the working adult sample, and the adolescent sample, females consistently displayed slightly higher scores on neuroticism and aesthetic sensitivity than males, which is largely consistent with the gender-difference pattern reported in previous research (Soto & John, 2017). In sum, this evidence indicates that the Chinese BFI-2 is a psychometrically sound measure that is well-suited for both research and applied contexts calling for a reliable, valid, and efficient assessment of the Big Five personality traits and their constituent facets.

However, item-level analyses revealed some boundary conditions for administering the Chinese BFI-2, especially when facet-level interpretations are desired. The five factors were well recovered in the college student sample, the employee sample, and the adolescent sample, with factor solutions resembling those in two U.S. comparison samples. However, as predicted, there were more-pronounced individual differences in responding to negatively worded items among less-educated and at-risk samples, which negatively impacted structural validity. Especially in the sample of adults being treated for substance abuse, many negatively worded items had large cross-loadings and reduced primary loadings. Modelling a negative wording method factor substantially improved model fit and reduced cross-loadings, making factor solutions more comparable to those obtained in better-educated samples. However, some primary loadings remained low. A recent study examining the structural validity of a 15-item version of the BFI in 23 non-WEIRD countries observed similar issues (Laajaj et al., 2019). Based on these findings, we suspect that people with lower reading ability have a greater difficulty interpreting negatively worded items and therefore responding to those items as if they were positively worded (e.g., choose "strongly agree" for both "Has few artistic interests" and "Is fascinated by art, music, or literature"). Such inconsistent responses are likely to result in some additional covariance among negatively worded items, which requires modeling an additional wording factor to account for (Schmitt & Stuits, 1985). This is consistent with the finding reported by Gnambs and Schroeders (2020) where they showed that the strength of negative wording factor was stronger among respondents with lower reading ability.

In addition to the less-clear domain-level structure observed in the substance abuse sample, it also proved difficult to distinguish among some facet-level traits in this sample, especially within the agreeableness and neuroticism domains. Moreover, the trust and intellectual curiosity facets of agreeableness and openness showed relatively low reliability in Study 2, a finding similar to what was reported for Russian, German, Slovak, and Danish versions of BFI-2 (Halama et al., 2020; Rammstedt et al., 2018; Shchebetenko et al., 2020; Vedel et al., 2020). To address this issue, future studies could revise these scales by developing and testing alternative item translations. We recommend a multisample and multiculture approach to maximize the probability of identifying items that can be used across populations and thereby preserve—or even improve—the cross-cultural comparability the BFI-2.

Surprisingly, the substance use sample reported higher conscientiousness and lower neuroticism compared with the other normative adult samples, a pattern of which is contrary to previous meta-analytical findings (Kotov et al., 2010). In hindsight, we think testing environment may have contributed to the observed patterns. Specifically, even though participants in the substance use sample were informed that their responses were anonymous and would only be used for research purposes, they completed the survey in small groups with staff members present. This proctored paper-and-pencil testing environment might result in socially desirable responding and lead to higher observed scores on conscientiousness and lower score on neuroticism (Chuah et al., 2006; Ployhart et al., 2003). Therefore, these differences should be interpreted cautiously.

Taken together, these findings offer insight regarding when researchers and practitioners can be confident about administering and interpreting the Chinese BFI-2, and when they should exercise greater caution. Specifically, our findings indicate that the Chinese BFI-2 can be safely used with college students, as well as adults with high school or college-level education. When intended for populations with less education, researchers and practitioners can still be confident about the reliability and validity of the Chinese BFI-2 at the level of the broad Big Five domains. However, they should be more cautious about interpreting results regarding facet-level traits. We further note that accounting for individual differences in response style—such as by modeling a negative wording factor—can help improve the item-level structural validity and cross-sample comparability of the Chinese BFI-2.

Limitations and Future Directions

The present research had a number of important strengths, such as the use of multiple diverse samples, a comprehensive examination of the nomological network of the Big Five factors, and the use of both exact and conceptual replications of criterion-related validity. However, it also had some limitations. First, the adolescent sample in Study 2 was recruited from a single high school in a particular city. It is therefore unknown how well findings from this sample can be generalized to adolescents from different regions given that some regional differences in several psychological attributes have been reported (Talhelm et al., 2014). Moreover, as this high school is one of the top high schools in that city with rich educational resources and motivated students, there might be range restrictions in both personality factors and outcomes, attenuating criterion-related validities. Future studies can address this limitation by recruiting adolescent respondents from different regions and from some medium-level or low-level high schools. Second, the college student sample was collected using snowball sampling technique, which may result in some degree of nonindependence. Third, most of the data analyzed here were cross-sectional, with the BFI-2 and criterion measures administered concurrently. Thus, future longitudinal research can provide stronger tests of the Chinese BFI-2's test–retest reliability and predictive validity.

Beyond these methodological limitations, the present findings also suggest some promising directions for future research. For example, we noticed that the four Chinese samples consistently displayed smaller variance than the two U.S. samples for most of the Big Five domains and facets. This may reflect response style differences between Asian and American people. Previous research has consistently found that Asian respondents are more likely than Americans to choose middle response options on Likert rating scales, whereas Americans are more likely to choose extreme response options. Moreover, this pattern holds regardless of item content and respondents' true trait levels (Chen et al., 1995). Thus, the current finding highlights the importance of psychometric research to develop and test methods, such as recently proposed tree models, that can explicitly model individual and group-level differences in extreme responding and other response styles (e.g., Böckenholt, 2017; Plieninger & Heck, 2018; Sun et al., 2021).

Another potential approach for addressing response style would be to develop a forced choice version of the BFI-2. With the emergence of two statistical models designed to address problems regarding ipsativity (Brown & Maydeu-Olivares, 2011; Stark et al., 2005), forced choice measures have been shown to effectively resist response styles and faking in high-stakes situations (Cao & Drasgow, 2019). Given the excellent content coverage of the BFI-2, developing a forced choice version would further strengthen its applicability in cross-cultural and high-stakes settings.

Conclusion

The present study adapted the BFI-2 into the Chinese language and cultural context, and evaluated the psychometric properties of the Chinese BFI-2 in four diverse samples. We conclude that the Chinese BFI-2 provides researchers a psychometrically sound and efficient tool for assessing the Big Five personality domains and more-specific facet traits, especially among adults with a high school or higher levels of education. The Chinese BFI-2 can also be used to assess the Big Five domains in younger, less-educated, and at-risk respondents when domain-level tests are of interest. However, researchers and practitioners should be more cautious when interpreting item responses and facet scale scores in these populations.

Appendix

大五人格问卷第二版 (BFI-2)

下面是一些关于个人特征的描述，有些可能适用于你，有些可能不适用于你。比如，你是否同意“我是一个喜欢与他人待在一起的人”？请在下面每个句子前的横线上填入对应的数字以表明你同意或不同意这个描述。

1	2	3	4	5
非常不同意	不太同意	态度中立	比较同意	非常同意

我是一个. . .. . .的人

1. ___性格外向，喜欢交际
2. ___心肠柔软，有同情心
3. ___缺乏条理
4. ___从容， 善于处理压力
5. ___对艺术没有什么兴趣
6. ___性格坚定自信，敢于表达自己的观点
7. ___为人恭谦，尊重他人
8. ___比较懒
9. ___经历挫折后仍能保持积极心态
10. ___对许多不同的事物都感兴趣
11. ___很少觉得兴奋或者特别想要(做)什么
12. ___常常挑别人的毛病
13. ___可信赖的，可靠的
14. ___喜怒无常，情绪起伏较多
15. ___善于创造，能找到聪明的方法来做事
16. ___比较安静
17. ___对他人没有什么同情心
18. ___做事有计划有条理
19. ___容易紧张
20. ___着迷于艺术、音乐或文学
21. ___常常处于主导地位，像个领导一样
22. ___常与他人意见不和
23. ___很难开始行动起来去完成一项任务
24. ___觉得有安全感，对自己满意
25. ___不喜欢知识性或者哲学性强的讨论
26. ___不如别人有活力
27. ___宽宏大量
28. ___有时比较没有责任心
29. ___情绪稳定，不易生气
30. ___几乎没有什么创造性
31. ___有时会害羞，比较内向
32. ___乐于助人，待人无私
33. ___习惯让事物保持整洁有序
34. ___时常忧心忡忡，担心很多事情
35. ___重视艺术与审美
36. ___感觉自己很难对他人产生影响
37. ___有时对人比较粗鲁
38. ___有效率，做事有始有终
39. ___时常觉得悲伤
40. ___思想深刻
41. ___精力充沛
42. ___不相信别人，怀疑别人的意图
43. ___可靠的，总是值得他人信赖
44. ___能够控制自己的情绪
45. ___缺乏想象力
46. ___爱说话，健谈
47. ___有时对人冷淡，漠不关心
48. ___乱糟糟的，不爱收拾
49. ___很少觉得焦虑或者害怕
50. ___觉得诗歌、戏剧很无聊
51. ___更喜欢让别人来领头负责
52. ___待人谦逊礼让
53. ___有恒心，能坚持把事情做完
54. ___时常觉得郁郁寡欢
55. ___对抽象的概念和想法没什么兴趣
56. ___充满热情
57. ___把人往最好的方面想
58. ___有时候会做出一些不负责任的行为
59. ___情绪多变，容易愤怒
60. ___有创意，能想出新点子

请检查是否在每个句子前的横线上都填了相应的数字。

计分方式

大五人格维度及其下属子维度所对于应的条目如下所示。R表示此条目需要反向计分。

大五人格维度

外向性 (Extraversion): 1, 6, 11R, 16R, 21, 26R, 31R, 36R, 41, 46, 51R, 56
宜人性 (Agreeableness): 2, 7, 12R, 17R, 22R, 27, 32, 37R, 42R, 47R, 52, 57

尽责性 (Conscientiousness): 3R, 8R, 13, 18, 23R, 28R, 33, 38, 43, 48R, 53, 58R
负性情绪/神经质 (Negative Emotionality): 4R, 9R, 14, 19, 24R, 29R, 34, 39, 44R, 49R, 54, 59
开放性 (Open-Mindedness): 5R, 10, 15, 20, 25R, 30R, 35, 40, 45R, 50R, 55R, 60

大五人格子维度
社交 (Sociability): 1, 16R, 31R, 46
果断 (Assertiveness): 6, 21, 36R, 51R
活力 (Energy Level): 11R, 26R, 41, 56
同情 (Compassion): 2, 17R, 32, 47R
谦恭 (Respectfulness): 7, 22R, 37R, 52
信任 (Trust): 12R, 27, 42R, 57
条理 (Organization): 3R, 18, 33, 48R
效率 (Productiveness): 8R, 23R, 38, 53
负责 (Responsibility): 13, 28R, 43, 58R
焦虑 (Anxiety): 4R, 19, 34, 49R
抑郁 (Depression): 9R, 24R, 39, 54
易变 (Emotional Volatility): 14, 29R, 44R, 59
好奇 (Intellectual Curiosity): 10, 25R, 40, 55R
审美 (Aesthetic Sensitivity): 5R, 20, 35, 50R
想象 (Creative Imagination): 15, 30R, 45R, 60

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by National Key R&D Program of China (2020YFC2003000) awarded to Dr. Jian Li.

ORCID iD

Jian Li https://orcid.org/0000-0002-6521-5956

Supplemental Material

Supplemental material for this article is available online.

Notes

1. https://osf.io/bjfdc/?view_only=5942590d234d458aba06eaff75167955
2. We also plotted reliability and criterion-related validity from the cleaned sample and the full sample against each other for all the four samples. Please refer to the online supplementary material (Figures S1-S4) for details.
3. We also fit a random-intercept ESEM to control for potential acquiescent responding (Aichholzer, 2014). However, results were almost identical to these of the ordinary ESEM and the variance of the random intercept was very small across samples. These results are available on request from the corresponding author.
4. As the survey battery was long and the respondents had demonstrated some degree of noncompliance (e.g., do not follow instructions, give a substantial amount of careless responses) when responding to questionnaires (mainly health-related questionnaires) in the past according to the staff members, we decided to embed more quality control items.

References

Aichholzer, J. (2014). Random intercept EFA of personality scales. *Journal of Research in Personality*, 53(December), 1-4. https://doi.org/10.1016/j.jrp.2014.07.001

Ang, R. P., & Huan, V. S. (2006). Academic expectations stress inventory: Development, factor analysis, reliability, and validity. *Educational and Psychological Measurement*, *66*(3), 522-539. https://doi.org/10.1177/0013164405282461

Anglim, J., Horwood, S., Smillie, L. D., Marrero, R. J., & Wood, J. K. (2020). Predicting psychological and subjective well-being from personality: A meta-analysis. *Psychological Bulletin*, *146*(4), 279-323. https://doi.org/10.1037/bul0000226

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

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology*, *85*(3), 349-360. https://doi.org/10.1037/0021-9010.85.3.349

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology*, *92*(2), 410-424. https://doi.org/10.1037/0021-9010.92.2.410

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, *22*(1), 69-83. https://doi.org/10.1037/met0000106

Böckenholt, U. (2019). Contextual responses to affirmative and/or reversed-worded items. *Psychometrika*, *84*(4), 986-999. https://doi.org/10.1007/s11336-019-09680-7

Bogg, T., & Roberts, B. W. (2004). Conscientiousness and health-related behaviors: A meta-analysis of the leading behavioral contributors to mortality. *Psychological Bulletin*, *130*(6), 887-919.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3),460-502.https://doi.org/10.1177/0013164410375112

Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology*, *63*(3), 452-459.

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*, *104*(11), 1347-1368. https://doi.org/10.1037/apl0000414

Chen, C., Lee, S. Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, *6*(3), 170-175. https://doi.org/10.1111/j.1467-9280.1995.tb00327.x

Chernyshenko, O. S., Kankaraš, M., & Drasgow, F. (2018). *Social and emotional skills for student success and wellbeing*. OECD Publishing.

Cheung, F. M., Leung, K., Fan, R. M., Song, W. Z., Zhang, J. X., & Zhang, J. P. (1996). Development of the Chinese personality assessment inventory. *Journal of Cross-Cultural Psychology*, *27*(2), 181-199. https://doi.org/10.1177/0022022196272003

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Christie, R., & Geis, F. (1970). *Studies in Machiavellianism*. Academic Press. https://doi.org/10.1016/B978-0-12-174450-2.50006-3

Chuah, S. C., Drasgow, F., & Roberts, B. W. (2006). Personality assessment: Does the medium matter? No. *Journal of Research in Personality*, *40*(4), 359-376. https://doi.org/10.1016/j.jrp.2005.01.006

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, *24*(4), 385-396. https://doi.org/10.2307/2136404

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

Costa, P. T., Jr., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment*, *64*(1), 21-50. https://doi.org/10.1207/s15327752jpa6401_2

Dai, X., & Wu, Y. (2005). The application of the NEO-PI-R in a population aged from 16-20 years old. *Chinese Journal of Clinical Psychology*, *13*(4), 14-18.

Danner, D., Lechner, C. M., Soto, C. J., & John, O. P. (2021). Modelling the incremental value of personality facets: The domains-incremental facets-acquiescence bifactor show-model. *Assessment*, *35*(1), 67-84. https://doi.org/10.1002/per.2268

Denissen, J. J., Geenen, R., Soto, C. J., John, O. P., & Van Aken, M. A. (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*, *102*(3), 309-324. https://doi.org/10.1080/00223891.2018.1539004

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 Aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*(5), 880-896. https://doi.org/10.1037/0022-3514.93.5.880

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71-75. https://doi.org/10.1207/s15327752jpa4901_13

Dueber, D. (2019). *dmacs: Measurement nonequivalence effect size calculator* (R package version 0.1.0) [Computer software]. https://CRAN.R-project.org/package=dmacs

Ehrhart, K. H., Roesch, S. C., Ehrhart, M. G., & Kilian, B. (2008). A test of the factor structure equivalence of the 50-item IPIP Five-factor model measure across gender and ethnic groups. *Journal of Personality Assessment*, *90*(5), 507-516. https://doi.org/10.1080/00223890802248869

Espelage, D. L., & Holt, M. (2001). Bullying and victimization during early adolescence: Peer influences and psychosocial correlates. *Journal of Emotional Abuse*, *2*(2-3), 123-142. https://doi.org/10.1300/J135v02n02_08

Finkelhor, D., Shattuck, A., Turner, H., & Hamby, S. (2015). A revised inventory of adverse childhood experiences. *Child Abuse & Neglect*, *48*(October), 13-21. https://doi.org/10.1016/j.chiabu.2015.07.011

Gnambs, T., & Schroeders, U. (2020). Cognitive abilities explain wording effects in the Rosenberg Self-Esteem Scale. *Assessment*, *27*(2), 404-418. https://doi.org/10.1177/1073191117746503

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*(1), 7-28.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504-528. https://doi.org/10.1016/S0092-6566(03)00046-1

Halama, P., Kohút, M., Soto, C. J., & John, O. P. (2020). Slovak adaptation of the Big Five Inventory (BFI-2): Psychometric properties and initial validation. *Studia Psychologica*, *62*(1), 74-87. https://doi.org/10.31577/sp.2020.01.792

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *Big Five Inventory (BFI)* [Database record]. APA PsycTests.

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* (pp. 114-158). Guilford Press.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin, & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102-138). Guilford Press.

Jokela, M., Batty, G. D., Nyberg, S. T., Virtanen, M., Nabi, H., Singh-Manoux, A., & Kivimäki, M. (2013). Personality and all-cause mortality: Individual-participant meta-analysis of 3,947 deaths in 76,150 adults. *American Journal of Epidemiology*, *178*(5), 667-675. https://doi.org/10.1093/aje/kwt170

Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: A concise measure of the dark triad. *Psychological Assessment*, *22*(2), 420-432. https://doi.org/10.1037/a0019265

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology*, *87*(3), 530-541. https://doi.org/10.1037/0021-9010.87.3.530

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology*, *98*(6), 875-925. https://doi.org/10.1037/a0033901

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin*, *136*(5), 768-821. https://doi.org/10.1037/a0020327

Laajaj, R., Macours, K., Hernandez, D. A. P., Arias, O., Gosling, S. D., Potter, J., Rubio-Codina, M., & Vakis, R. (2019). Challenges to capture the big five personality traits in non-WEIRD populations. *Science Advances*, *5*(7), eaaw5226. https://doi.org/10.1126/sciadv.aaw5226

Lemmens, J. S., Valkenburg, P. M., & Peter, J. (2009). Development and validation of a game addiction scale for adolescents. *Media Psychology*, *12*(1), 77-95. https://doi.org/10.1080/15213260802669458

Lorenzo-Seva, U., & Ten Berge, J. M. (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, J. G., Nisbett, R. E., & Morris, M. W. (2020). Why East Asians but not South Asians are underrepresented in leadership positions in the United States. *Proceedings of the National Academy of Sciences*, *117*(9), 4590-4600. https://doi.org/10.1073/pnas.1918896117

Luo, J., & Dai, X. (2011). Meta-analysis of Big-five factor personality tests in China. *Chinese Journal of Clinical Psychology*, *19*(6), 740-752.

Mount, M. K., Barrick, M. R., Scullen, S. M., & Rounds, J. (2005). Higher-order dimensions of the big five personality traits and the big six vocational interest types. *Personnel Psychology*, *58*(2), 447-478. https://doi.org/10.1111/j.1744-6570.2005.00468.x

Muris, P., Merckelbach, H., Otgaar, H., & Meijer, E. (2017). The malevolent side of human nature: A meta-analysis and critical review of the literature on the dark triad (narcissism, Machiavellianism, and psychopathy). *Perspectives on Psychological Science*, *12*(2), 183-204. https://doi.org/10.1177/1745691616666070

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Ng, T. W., & Feldman, D. C. (2010). Human capital and objective indicators of career success: The mediating effects of cognitive ability and conscientiousness. *Journal of Occupational and Organizational Psychology*, *83*(1), 207-235. https://doi.org/10.1348/096317909X414584

Norton, P. J. (2007). Depression Anxiety and Stress Scales (DASS-21): Psychometric analysis across four racial groups. *Anxiety, Stress, & Coping*, *20*(3), 253-265. https://doi.org/10.1080/10615800701309279

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods*, *22*(3), 678-709. https://doi.org/10.1177/1094428118761122

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966-980. https://doi.org/10.1037/a0022955

Organ, D. W., & Ryan, K. (1995). A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. *Personnel Psychology*, *48*(4), 775-802. https://doi.org/10.1111/j.1744-6570.1995.tb01781.x

Özer, B. U., Saçkes, M., & Tuckman, B. W. (2013). Psychometric properties of the Tuckman Procrastination Scale in a Turkish sample. *Psychological Reports*, *113*(3), 874-884. https://doi.org/10.2466/03.20.PR0.113x28z7

Plieninger, H., & Heck, D. W. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivariate Behavioral Research*, *53*(5), 633-654. https://doi.org/10.1080/00273171.2018.1469966

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: Are personality, biodata, and situational judgment tests comparable? *Personnel Psychology*, *56*(3), 733-752. https://doi.org/10.1111/j.1744-6570.2003.tb00757.x

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*(2), 322-338. https://doi.org/10.1037/a0014996

Qian, M., Wu, G., Zhu, R., & Zhang, X. (2000). Development of the Revised Eysenck Personality Questionnaire Short Scale for Chinese (EPQ-RSC). *Acta Psychologica Sinica*, *32*(3), 317-323.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401-11405. https://doi.org/10.1073/pnas.1721165115

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*, *36*(1), 149-161. https://doi.org/10.1027/1015-5759/a000481

Revelle, W. R. (2017). *psych: Procedures for personality and psychological research*. Northwestern University.

Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. *Journal of Research in Personality*, *43*(2), 137-145.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Rounds, J., Ming, C. W. J., Cao, M., Song, C., & Lewis, P. (2016). *Development of an O* NET® Mini Interest Profiler (Mini-IP) for mobile devices: Psychometric characteristics*. Department of Labor O* NET Resource Center. https://www.onetcenter.org/reports/Mini-IP.html

Saucier, G. (1994). Mini-Markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment*, *63*(3), 506-516. https://doi.org/10.1207/s15327752jpa6303_8

Schat, A. C., Kelloway, E. K., & Desmarais, S. (2005). The Physical Health Questionnaire (PHQ): Construct validation of a self-report scale of somatic symptoms. *Journal of Occupational Health Psychology*, *10*(4), 363-381. https://doi.org/10.1037/1076-8998.10.4.363

Schmitt, N., & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*(4), 367-373. https://doi.org/10.1177/014662168500900405

Shchebetenko, S., Kalugin, A. Y., Mishkevich, A. M., Soto, C. J., & John, O. P. (2020). Measurement invariance and sex and age differences of the Big Five Inventory–2: Evidence from the Russian version. *Assessment*, *27*(3), 472-486. https://doi.org/10.1177/1073191119860901

Soto, C. J. (2020). Do links between personality and life outcomes generalize? Testing the robustness of trait–outcome associations across gender, age, ethnicity, and analytic approaches. *Social Psychological and Personality Science*, *12*(1), 118-130. https://doi.org/10.1177/1948550619900572

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

Spector, P. E., Bauer, J. A., & Fox, S. (2010). Measurement artifacts in the assessment of counterproductive work behavior and organizational citizenship behavior: Do we know what we think we know? *Journal of Applied Psychology*, *95*(4), 781-790. https://doi.org/10.1037/a0019477

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, *29*(3), 184-203. https://doi.org/10.1177/0146621604273988

Steenkamp, J. B. E., & Maydeu-Olivares, A. (2020). An updated paradigm for evaluating measurement invariance incorporating common method variance and its assessment. *Journal of the Academy of Marketing Science*, *49*(1), 5-29. https://doi.org/10.1007/s11747-020-00745-z

Straus, M. A., Hamby, S. L., Finkelhor, D., Moore, D. W., & Runyan, D. (1998). Identification of child maltreatment with the Parent-Child Conflict Tactics Scales: Development and psychometric data for a national sample of American parents. *Child Abuse & Neglect*, *22*(4), 249-270. https://doi.org/10.1016/S0145-2134(97)00174-9

Strickhouser, J. E., Zell, E., & Krizan, Z. (2017). Does personality predict health and well-being? A metasynthesis. *Health Psychology*, *36*(8), 797-810. https://doi.org/10.1037/hea0000475

Sun, T., Zhang, B., Cao, M., & Drasgow, F. (2021). Faking detection improved: Adopting a Likert item response process tree model. *Organizational Research Methods*. Advanced online publication. https://doi.org/10.1177/10944281211002904

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: an illustration using M*plus* and the lavaan/semTools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 111-130. https://doi.org/10.1080/10705511.2019.1602776

Talhelm, T., Zhang, X., Oishi, S., Shimin, C., Duan, D., Lan, X., & Kitayama, S. (2014). Large-scale psychological differences within China explained by rice versus wheat agriculture. *Science*, *344*(6184), 603-608. https://doi.org/10.1126/science.1246850

Vedel, A., Wellnitz, K. B., Ludeke, S., Soto, C. J., John, O. P., & Andersen, S. C. (2020). Development and validation of the Danish Big Five Inventory-2: Domain-and facet-level structure, construct validity, and reliability. *European Journal of Psychological Assessment*, *37*(1), 42-51. https://doi.org/10.1027/1015-5759/a000570

Wang, D., & Cui, H. (2004). Reliabilities and validities of the Chinese Personality Scale. *Acta Psychologica Sinica*, *36*(3), 347-358.

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, *49*(5), 737-747. https://doi.org/10.1509/jmr.11.0368

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81*(4), 1014-1045. https://doi.org/10.1007/s11336-016-9506-0

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409-428. https://doi.org/10.3758/s13428-018-1055-2

Yang, K. S., & Bond, M. H. (1990). Exploring implicit personality theories with indigenous or imported constructs: The Chinese case. *Journal of Personality and Social Psychology*, *58*(6), 1087-1095. https://doi.org/10.1037/0022-3514.58.6.1087

Yao, R., & Liang, L. (2010). Analysis of the application of simplified NEO-FFI to undergraduates. Chinese. *Chinese Journal of Clinical Psychology*, *18*(4), 457-459.

Zhang, B., Luo, J., Chen, Y., Roberts, B., & Drasgow, F. (2020). *The road less traveled: A cross-cultural study of the negative*

wording factor in multidimensional scales. PsyArXiv Preprints. https://doi.org/10.31234/osf.io/2psyq

Zhang, B., Sun, T., Cao, M., & Drasgow, F. (2020). Using bifactor models to examine the predictive validity of hierarchical constructs: Pros, cons, and solutions. *Organizational Research Methods*. Advanced online publication. https://doi.org/10.1177/1094428120915522

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2020). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. *Organizational Research Methods*, *23*(3), 569-590. https://doi.org/10.1177/1094428119836486

Zheng, L., Goldberg, L. R., Zheng, Y., Zhao, Y., Tang, Y., & Liu, L. (2008). Reliability and concurrent validation of the IPIP Big-Five factor markers in China: Consistencies in factor structure between Internet-obtained heterosexual and homosexual samples. *Personality and Individual Differences*, *45*(7), 649-654. https://doi.org/10.1016/j.paid.2008.07.009