JOURNAL OF PERSONALITY ASSESSMENT
2026, VOL. 108, NO. 1, 90–100
https://doi.org/10.1080/00223891.2025.2524373

OPEN ACCESS

Measuring Chinese Personality in 8 Minutes: A Short Measure of the Five-Factor Model of Personality

Michelle Yik[1], Felity H. C. Kwok[1] and Kim De Roover[2]

[1]Division of Social Science, Hong Kong University of Science and Technology; [2]Faculty of Psychology and Educational Sciences, KU Leuven–University of Leuven

ABSTRACT

Using data from three Chinese samples (*N*s = 611, 403, 299) collected using both monolingual and bilingual designs, we evaluated the psychometric properties and factor structure of the NEO Five-Factor Inventory 3 (FFI-3), the short form of the NEO Personality Inventory 3 (PI-3), for use in Chinese communities. Although the FFI-3 contains only a quarter of the 240 items of the PI-3, exploratory structural equation modeling revealed that it maintained the five-factor structure of the long form and achieved acceptable levels of internal consistency, cross-language validity, and test–retest reliability. The correlation coefficients between the short-form factors and the corresponding long-form factors were all above .86, indicating a strong association between the short and long versions of the scale. Taken together, our findings suggest that the FFI-3 is a viable tool for mapping personality in Chinese communities.

ARTICLE HISTORY
Received 29 July 2024
Revised 29 May 2025
Accepted 10 June 2025

Researchers investigating personality share the common goal of assessing participants' personalities as efficiently and effectively as possible. Although the likelihood of achieving this goal depends on multiple variables, including item readability, clarity of instructions, and participant interest in the topic, the length of personality measures plays a key role. Shorter versions place fewer time constraints on participants, allowing them to answer each item more carefully, which increases data quality. Researchers can then use the time saved to collect data on other variables pertinent to their research questions. In this study, we analyzed the structural and psychometric properties of the 60-item NEO Five Factor Inventory 3 (FFI-3; McCrae & Costa, 2010), a short version of the fourth generation of NEO scales, with three large Chinese samples. We also tested the equivalence between the English version and a translated Chinese version of the FFI-3.

Why the short form?

The NEO Personality Inventory 3 (PI-3; McCrae & Costa, 2010) measures the dimensions of the five-factor model of personality (FFM; McCrae & John, 1992; see Goldberg, 1981): neuroticism (N), extraversion (E), openness to experience (O), agreeableness (A), and conscientiousness (C). Each factor is measured by six facet scales, each tapping a specific aspect of the relevant factor. Using 240 items, the PI-3 thus assesses 30 traits in approximately 30 min. The FFI-3 is a short (60-item) version of the PI-3 that assesses the five factors without facet scales in 8 min.

Representing the fourth generation of NEO inventories, the PI-3 has shown high levels of reliability and validity, equaling or even exceeding those of its predecessors (McCrae & Costa, 2010), in most languages (De Fruyt et al., 2009; cf. Källmen et al., 2016; Quy, 2011). Developed for a wide range of respondents, including those as young as 12 years old and those with reading levels as low as Grade 5, the PI-3 is currently the most effective NEO measure (McCrae & Costa, 2010). However, soon after its publication in 2005 (McCrae et al., 2005), researchers began calling for shorter measures of the five personality factors (McCrae & Costa, 2007). The FFI-3 was developed to answer this call.

Short scales offer many advantages to researchers. As the number of items increases, respondents are more likely to experience fatigue or boredom, leading to lower response rates and lower quality data (Krosnick, 1991, 1999; Tourangeau et al., 2000; Ziegler et al., 2014). A short version minimizes the risk of experiencing fatigue or boredom and thereby encourages participants to cooperate in completing the questionnaire. With the growing popularity of experience sampling studies in which participants are regularly asked to report their emotions, perceptions, behaviors, or psychological states, short scales are typically used to minimize respondent burden, which can affect attrition rates and data quality (Krosnick, 1991).

However, short scales are not without limitations. Unlike full personality scales, which cover a wide range of attributes within each personality factor, short scales capture each factor at a global level, potentially reducing reliability and

CONTACT Michelle Yik Michelle.Yik@ust.hk Division of Social Science, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

validity (Paunonen & Ashton, 2001; see also Kemper et al., 2019). The global scores obtained for the five factors may be disadvantageous in certain contexts, encouraging researchers to use 30 facet scores to conduct an in-depth analysis of personality correlates (Rolstad et al., 2011; see also Credé et al., 2012). However, several short measures have had some success, showing adequate psychometric properties and providing useful information enabling researchers to map relationships between personality and other psychological correlates (Gosling et al., 2003; Rammstedt & John, 2005). In the present study, we conducted a psychometric evaluation of the FFI-3 with the main objective of providing a short personality scale for use in Chinese communities.

The FFI-3 in translation

Since the publication of the FFI-3 in 2007 (McCrae & Costa, 2007), only a handful of studies have focused on testing its psychometric properties. In North American samples, studies have reported acceptable internal consistency scores for all five factors (Perez, 2020), with alpha values ranging from .65 (O) to .87 (C) (see also Marjanovic et al., 2015; McCrae & Costa, 2007). However, mixed findings have been reported outside North America. Although acceptable alpha values have been found for the Arabic (Rabadi & Rabadi, 2021), Swedish (Axelsson et al., 2019), and Italian (Falgares et al., 2022) translations of the FFI-3, the levels of internal consistency of the Filipino (Reyes et al., 2019) and Indian (Kunnel John et al., 2019) translations are less encouraging.

Only three studies testing the five-factor structure of the FFI-3 have been published to date. Specifically, McCrae and Costa (2007) found support for the five-factor structure in their US normative sample, with all item loadings greater than .30 in absolute magnitude on their intended factors. Similar support was found in a study of an Arabic translation of the FFI-3 (Rabadi & Rabadi, 2021). However, a study of Hindi and Malayalam translations yielded mixed findings regarding the model's fit to the five-factor structure (Kunnel John et al., 2019).

The present study

The psychometric properties of various translations of the FFI-3 have been found to vary in samples outside of North America. Although the internal consistency and factor structure of the original (English) scale have largely been retained in translation into most Indo-European languages, E, O, and A have failed to reach acceptable levels of internal consistency in Hindi, Malayalam, and Filipino translations. Scholars have yet to pay attention to the adaptation of the FFI-3 to the Chinese language, the most widely spoken language in the world after English (Eberhard et al., 2021).[1] Moreover, most studies on translations of the FFI-3 have only tested the internal consistency of its scales, without examining the replicability of their five-factor structure. The present study is not only the first to test the psychometric properties of the Chinese FFI-3 but also the first to use exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009) to analyze its factor structure.

Many measurement tools, including the Big Five scales, have achieved well-defined exploratory factor analysis (EFA) structures, but have not been supported by confirmatory factor analysis (CFA; see Marsh et al., 2005). As noted by McCrae et al. (1996), "structures that are known to be reliable showed poor fits when evaluated by CFA techniques. We believe this points to serious problems with CFA itself when used to examine personality structure" (p. 568; see also McCrae & Costa, 1997). Marsh et al. (2010) argued that the requirement that each indicator loads on only one factor is too restrictive for personality research, as many indicators are likely to have secondary loadings on other factors (see McCrae et al., 1996). Compensating for the inappropriate imposition of zero loadings, ESEM offers a much more effective and less restrictive technique for mapping real-life personality data. Using ESEM, we also established the measurement invariance of the English and Chinese versions of the FFI-3.

Method

Participants and procedures

We reanalyzed the NEO data obtained from three studies involving Chinese participants who completed the FFI-3 or the PI-3 and scales measuring other variables of interest. All of the participants were bilingual undergraduate students at a university in Hong Kong.[2] They were recruited through email advertisements.

Dataset 1 came from Yik and Siu (2024; see also Yik & Siu, 2025), in which the English version of the FFI-3 was administered to 611 participants (329 women; $M_{\text{age}} = 20.31$, $SD_{\text{age}} = 1.80$). The participants completed the scale on surveyYIK, an app developed by the first author for use on Android and iOS devices. Datasets 2 and 3 came from Yik et al. (2023),[3] in which the participants completed the

[1] Based on a search of Google Scholar and the China National Knowledge Infrastructure database (capturing the literature published in simplified Chinese, including journal articles, theses/dissertations, proceedings, newspapers, and yearbooks), only one paper on the Chinese version of the PI-3 (long form) has been published to date (Yik et al., 2023). In that paper, the PI-3 was found to be a sound instrument for use in Chinese communities.

[2] The university's admissions procedures require enrolled students to demonstrate proficiency in both English and Chinese in standard language examinations. All of the participants were therefore considered to be fluent in both languages.

[3] Yik et al. (2023) used a standardized translation and back-translation procedure to prepare the Chinese language version of the PI-3 scales. Specifically, they recruited two bilinguals; the first bilingual translated the English items into Chinese and the second bilingual independently back translated the items into English. Discrepancies between the original and back-translated English versions were identified, discussed, and reconciled. Using multiple indicator growth modeling, they tested and attained the psychometric equivalence of the two language versions.

PI-3 scales twice online in a computer laboratory, two weeks apart. In Dataset 2, 403 participants (222 women; M_{age} = 20.28, SD_{age} = 1.36) completed the Chinese version of the scales in both sessions. In Dataset 3, 299 participants (146 women; M_{age} = 21.12, SD_{age} = 1.06) were randomly assigned to complete the English (or Chinese) version at Time 1 and the Chinese (or English) version at Time 2.

NEO scales

In all three studies from which the datasets were drawn, the participants completed the self-report version of the NEO scales. In Dataset 1, the participants completed the English version of the FFI-3, which includes 60 items. In Datasets 2 and 3, the participants completed the PI-3, which includes 240 items measuring the dimensions of the FFM, with 60 of these 240 items used to score the FFI-3 factors. The participants rated their agreement with each item on a 5-point scale ranging from 0 (*strongly disagree*) to 4 (*strongly agree*).

Results

First, we conducted ESEM to examine the psychometric properties of the English FFI-3. Second, we tested the measurement invariance of the 60 English FFI-3 items across the short and long forms using multigroup ESEM. Third, we examined the reliability and structural validity of the Chinese translation of the FFI-3. Finally, to test the psychometric equivalence of the Chinese translation, we tested the measurement invariance between the English and Chinese versions of the measure, using multiple indicator growth modeling to test two measurement moments.

Psychometric properties of the English FFI-3

Using Dataset 1 ($N = 611$), we first examined the internal consistency of the English FFI-3. The coefficient alpha values for the five factors were .795 for N, .746 for E, .689 for O, .680 for A, and .806 for C, with a median value of .746 (E).

Next, we examined the five-factor structure of the English FFI-3 by conducting ESEM (more specifically, exploratory factor analysis) in Mplus 8.8 (Muthén & Muthén, 2017). We used maximum likelihood with robust standard errors (MLR) estimation and oblique geomin rotation to extract the factors and evaluated the goodness of fit of the model using the comparative fit index (CFI), Tucker–Lewis index (TLI), and root mean square error of approximation (RMSEA) (Marsh et al., 2009). An acceptable fit is obtained when the CFI and TLI values are greater than .90 and the RMSEA value is below .09 (Marsh et al., 2004). The ESEM solution is presented in the left panel of Table 1 and the model fit statistics for the ESEM solution are presented in the upper part of Table 2. The ESEM solution yielded CFI and TLI values of .828 and .794, respectively, slightly below the respective thresholds for acceptable fit, but an RMSEA value of .038, indicating a good model fit.[4]

The factor solution showed that 47 (78%) of the 60 items had their highest loadings on the intended factors, with 36 of the loadings reaching .40 in absolute magnitude at $p < .001$. All N items except N36 (getting angry at how one is being treated) had their highest loadings on N, with N36 having its highest loading on A. All E items except E32 (feeling energetic), E42 (not enjoying chatting with others), and E47 (getting things done quickly) had their highest loadings on E, with E32 and E42 having their highest loadings on A and E47 having its highest loading on C. Eight of the O items had their highest loadings on O, with the remaining four items, O18 (controversial speakers mislead students), O23 (not sensitive to poetry), O28 (difficulty letting one's mind wander), and O33 (not sensitive to noticing the moods produced by the environment), having their highest loadings on A. Eight of the A items had their highest loadings on A, with the remaining four items, A04 (showing courtesy to everyone), A29 (forgiving those who insulted you), A34 (assuming the best about others), and A49 (being considerate), having their highest loadings on O. All C items except C45 (not being dependable) had their highest loadings on C, with C45 having its highest loading on N.

Of the 13 items whose highest loadings were not on their intended factors, 7 had their highest loadings on A (N36, E32, E42, O18, O23, O28, O33), many of which are related to social harmony. For instance, although N36, E42, and O18 are not intended to measure A per se, their content describes interpersonal interactions such as getting angry with others, not having pleasant chats with others, and misleading others, which could be viewed as behaviors that disrupt social harmony. Although E32, O28, and O33 do not explicitly describe interpersonal interactions, they could also be considered to connote social harmony. For example, the states of bursting with energy, letting the mind wander without control, and experiencing feelings produced by different environments are all likely to have interpersonal consequences if the associated emotions and thoughts are not handled appropriately. Such implications for social harmony may explain why these non-A items had high loadings on A.

[4] Our less-than-ideal CFI and TLI values may be due to the relatively weak correlations between the manifest variables, among which only 19 (1%) of the 1,770 possible correlations were above the absolute value of .40. Indeed, these 19 correlations were all between items belonging to the same FFI-3 factor, which should be more strongly correlated than the items belonging to different factors. In other words, among the 330 possible within-factor correlations (mean r = .20), only 19 (6%) were above the absolute value of .40, indicating relatively weak correlations. As such, these weak correlations align closely with the assumed uncorrelated manifest variables in the baseline model used to estimate these comparative indices. When we fitted the five-factor model to the data, the hypothesized model failed to significantly improve model fit from the baseline model, resulting in slightly lower values for CFI and TLI (see Lai & Green, 2016).

Table 1. ESEM factor structures for the English and Chinese FFI-3 Scales.

	English (Dataset 1; N=611)					Chinese (Dataset 2 at Time 1 and Dataset 3; N=702)				
Item	N	E	O	A	C	N	E	O	A	C
N01	**.49*****	−.12*	.00	.19	.08*	.21***	−.08	−.01	.01	−.03
N06	**.44*****	.04	.13	−.09	.10*	**.53*****	−.10**	.01	−.24***	.00
N11	**.52*****	.01	.17*	−.11	.01	**.58*****	.07	−.06	−.07	.00
N16	**.52*****	−.07	−.11*	.20	−.09*	**.60*****	−.12**	.06	.01	−.04
N21	**.54*****	−.06	.03	−.15	.06	**.60*****	−.04	−.03	.02	.01
N26	**.55*****	−.12**	.07	−.19	−.12**	**.53*****	−.06	−.03	−.10	−.08*
N31	**.61*****	−.05	−.07	.32*	.02	**.65*****	−.10*	.02	.01	−.05
N36	.24**	.11*	−.06	−.38***	−.09	.29***	.05	−.10*	−.23***	.05
N41	**.43*****	−.07	.02	−.26*	−.16**	**.49*****	.05	.00	−.04	−.19***
N46	**.58*****	.00	−.05	.23	.00	**.66*****	−.18***	.05	−.04	−.02
N51	**.46*****	−.02	−.03	−.27*	−.17**	**.56*****	.11**	−.12**	−.10*	−.12**
N56	.39***	−.18**	.17*	−.15	−.15**	**.48*****	−.11*	.07	−.14**	−.05
E02	.09	**.75*****	−.03	−.08	−.06	.14***	**.73*****	−.06	−.01	.02
E07	.01	.37***	.32**	.02	−.03	−.01	**.45*****	.14**	.08	−.08
E12	.00	**.52*****	−.14**	.25***	−.15**	−.06	**.40*****	−.06	.13**	−.05
E17	.04	**.68*****	.12*	.00	.02	.06	**.59*****	.12**	.07	.06
E22	−.05	.32***	.16**	−.20***	.09	.02	**.71*****	−.05	−.03	−.03
E27	−.08	**.61*****	−.23***	.09	.03	−.15***	**.57*****	−.06	.12**	−.01
E32	−.11	.11	.07	−.27***	.10	−.21***	**.45*****	.02	.00	.12**
E37	−.27**	**.50*****	.30**	−.08	.06	−.22***	**.61*****	.00	.08	.02
E42	.09	.26***	−.04	**.42*****	.02	.01	.39***	.11*	.21***	.02
E47	.15*	.15**	−.02	−.23***	.28***	.18***	.18***	−.05	−.15**	.39***
E52	−.13	**.48*****	.14**	−.26***	.20***	−.15***	**.64*****	.06	−.09*	.09**
E57	−.01	**.45*****	−.11	.15**	.07	−.10*	.35***	.00	.15**	.15**
O03	.05	.03	**.42*****	−.03	−.09	.14**	.05	.29***	−.18***	−.11*
O08	−.01	.05	**.45*****	.15**	.18**	−.03	.10*	**.43*****	.01	.18***
O13	−.05	−.03	**.55*****	.02	−.03	.08*	−.04	**.59*****	.11**	−.01
O18	.05	.08	.06	**.43*****	−.06	−.01	−.05	.29***	.16**	−.01
O23	−.09	.02	.21	.24***	−.13*	.00	−.11**	**.55*****	.12*	−.13**
O28	−.12	.05	.13	.37***	−.06	−.20***	.06	.29***	.04	−.13**
O33	.12	−.01	.11	.37***	.00	.09	.10*	.32***	.13*	.07
O38	.26***	.16**	.36***	.02	.09	.32***	.25***	.20***	.01	.12**
O43	−.10	−.01	**.45*****	.02	−.13*	.00	−.04	**.56*****	.00	−.13**
O48	−.03	−.10	.33**	.30***	−.08	−.02	−.06	**.60*****	−.03	−.08*
O53	−.08	.07	**.45*****	−.11	.21***	−.06	.05	**.52*****	−.12**	.20***
O58	−.14**	−.05	**.41*****	−.04	−.07	−.04	−.07	**.46*****	−.18***	.07
A04	.20**	.16**	.34***	.09	.13*	.19***	.12**	.16**	.28***	.24***
A09	.01	−.15**	.10	**.45*****	.05	.03	−.16***	−.08*	**.54*****	.01
A14	−.11	.07	.06	**.48*****	.04	−.06	.04	.02	**.64*****	.01
A19	.02	−.09	.01	.34***	.00	−.02	−.12**	−.08	.35***	−.04
A24	.22**	−.09	−.12*	.30***	−.23***	.28***	−.14**	−.25***	.16**	−.21***
A29	−.06	−.05	.26**	.02	.01	−.02	.15**	.18***	.29***	.02
A34	−.05	.12	.32**	.01	.12	.03	.21***	.05	.28***	−.05
A39	−.01	.28***	.07	.39***	−.20***	−.02	.15***	.01	**.54*****	−.09*
A44	.15	.07	.24***	**.49*****	.14**	.19***	.11**	.09*	.32***	.05
A49	.21***	.00	**.42*****	.22***	.19**	.15**	.11*	.12*	.36***	.11*
A54	.05	−.17**	.04	.37***	−.01	−.03	−.14**	−.01	.21***	−.03
A59	−.01	−.01	.01	**.43*****	−.01	−.01	−.13**	−.02	**.56*****	−.06
C05	.02	.00	−.03	−.03	**.47*****	.03	.01	.01	−.06	**.40*****
C10	−.12**	.06	−.07	−.06	**.58*****	−.07*	.03	−.04	.00	**.63*****
C15	−.16	.06	−.14	.27***	.35***	−.16**	−.05	−.08	.08	.15**
C20	.15**	.00	.13	.06	**.56*****	.15***	.02	.06	.18**	**.51*****
C25	−.11*	.10*	.03	−.16**	**.60*****	−.01	.00	−.06	−.05	**.70*****
C30	−.21	.03	−.30***	.15**	.36***	−.15**	−.09*	.00	.17**	.39***
C35	−.02	.07	.19**	.02	**.62*****	.07	.03	.01	.08*	**.62*****
C40	.12*	.07	.17*	.08	**.51*****	−.01	−.02	.10*	.15**	**.43*****
C45	−.26*	.07	−.14	.22**	.21***	−.21***	.00	.01	.25***	.33***
C50	−.08*	.00	−.03	−.03	**.66*****	−.13***	.10**	.05	−.07	**.66*****
C55	−.02	.00	−.18**	**.41*****	**.44*****	−.19***	−.07	−.01	.13*	**.54*****
C60	.06	.12*	.22***	.04	**.47*****	.02	.03	.21***	−.14**	**.53*****

Note. ESEM = exploratory structural equation modeling; FFI-3 = NEO Five-Factor Inventory 3; N = neuroticism; E = extraversion; O = openness to experience; A = agreeableness; C = conscientiousness. Items N01, N16, N31, N46, E12, E27, E42, E57, O18, O23, O28, O33, O48, A09, A14, A19, A24, A39, A44, A54, A59, C15, C30, C45, and C55 were reverse-coded prior to statistical analyses. The coefficient alpha values for the English FFI-3 factors were .795 for N, .746 for E, .689 for O, .680 for A, and .806 for C. The coefficient alpha values for the Chinese FFI-3 factors were .832 for N, .825 for E, .724 for O, .671 for A, and .809 for C. Loadings equal to or greater than |.40| are presented in bold. Loadings that were significant only on their intended factors at $p < .001$ are underlined.

$^{*}p < .05$.

$^{**}p < .01$.

$^{***}p < .001$.

Measurement invariance of the English FFI-3 as a stand-alone scale vs. as extracted from the PI-3

We examined the validity of the English FFI-3 by testing a series of measurement invariance models using the FFI-3 as a stand-alone scale (i.e., short form) and as an extraction of the PI-3 (i.e., long form), using the five-factor solution as a baseline model. We used data from Dataset 1 (N=611), in which the participants completed the English FFI-3 as a stand-alone

Table 2. Model fit results of ESEM for the English and Chinese FFI-3 Scales.

Model	χ^2	*df*	RMSEA	CFI	TLI
Overall models					
English (Dataset 1; $N=611$)	2,757.88	1,480	.038	.828	.794
Chinese (Dataset 2 at Time 1 and Dataset 3; $N=702$)	3,454.39	1,480	.044	.798	.759
Models of invariance for the English form (stand-alone form in Dataset 1; $N=611$ vs. extracted form in Dataset 3; $N=299$)					
1. Configural	5,393.78	3,080	.041	.796	.757
2. Metric (loadings)	5,728.28	3,355	.039	.791	.771
3. Strong (loadings, intercepts)	5,984.45	3,410	.041	.773	.756
3p. Partial strong (loadings, intercepts)[a]	5,684.15	3,289	.040	.785	.769
4. Latent mean (loadings, intercepts, factor means)[a]	5,701.63	3,294	.040	.784	.768
Models of language invariance (Dataset 3; $N=299$)					
1. Configural	10,051.16	6,475	.043	.735	.708
2. Metric (loadings)	10,529.07	6,750	.043	.720	.704
2p. Partial metric (loadings)[b]	10,435.22	6,736	.043	.726	.710
3. Strong (loadings, intercepts)[b,c]	11,062.71	6,793	.046	.684	.668
3p. Partial strong (loadings, intercepts)[b,c,d]	10,968.62	6,790	.045	.691	.675

Note. ESEM = exploratory structural equation modeling; FFI-3 = NEO Five-Factor Inventory 3; χ^2 = chi-square fit statistic; *df* = degrees of freedom; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index.

[a] Non-invariant intercepts of C15 (i.e., C155 on the PI-3) across forms.

[b] Non-invariant loadings of E22 and O48 across languages.

[c] Residual variances of E22 and O48 restricted to be above zero.

[d] Non-invariant intercepts of O53, E27, and O13 across languages.

Table 3. Two-week test–retest reliability of the Chinese FFI-3 Scales.

Scale	r_{cc} (Dataset 2; $N=403$)	r_{ce} (Dataset 3; $N=299$)
N: Neuroticism	.83	.77
E: Extraversion	.89	.82
O: Openness to Experience	.85	.71
A: Agreeableness	.80	.70
C: Conscientiousness	.86	.85
Mdn	.85	.77

Note. FFI-3 = NEO Five-Factor Inventory 3; r_{cc} = retest correlation for the Chinese version; r_{ce} = equivalence correlations between the Chinese and English versions. All correlations significant at $p < .001$.

scale, and Dataset 3 ($N=299$), in which the participants completed the full English PI-3 from which 60 FFI-3 items were extracted. We performed multigroup ESEM. The middle panel of Table 2 shows the fit indices for the series of invariance models we tested (Marsh et al., 2009; Meredith, 1993). We used changes in CFI and RMSEA values to compare the statistical properties of the nested models. In general, a more constrained model is supported when ΔCFI < .010 and ΔRMSEA < .015 (Chen, 2007; Cheung & Rensvold, 2002).

We began with a configural invariance model (Model 1) that had no invariance constraints across the two forms. This model showed suboptimal CFI and TLI values (CFI = .796, TLI = .757) but a good RMSEA value (.041), indicating the presence of configural invariance. We then tested a metric invariance model (Model 2) in which the factor loadings were held invariant across the two forms. Model 2 again showed suboptimal CFI and TLI values (CFI = .791, TLI = .771) but a good RMSEA value (.039). Model 2 showed an acceptable decrease in model fit relative to Model 1 (ΔCFI = −.005, ΔRMSEA = −.002), indicating the presence of metric invariance.

Based on the metric invariance model (Model 2), we tested a strong invariance model (Model 3) in which the item intercepts, along with the factor loadings, were held invariant across the two forms. Support for strong invariance would indicate that the intercepts are the same across forms, such that differences in the observed means of the 60 items can only be explained by differences in the latent means of the five factors. Similar to Model 2, Model 3 showed suboptimal CFI and TLI values (CFI = .773, TLI = .756) but a good RMSEA value (.041). However, Model 3 showed a significantly worse fit than Model 2 (ΔCFI = −.018, ΔRMSEA = .002). Based on the highest modification index of item C15 from the FFI-3 (i.e., C155 from the PI-3), we tested a partial strong invariance model (Model 3p) by allowing the intercepts of this item to vary across the two forms. Again, Model 3p showed suboptimal CFI and TLI values (CFI = .785, TLI = .769) but a good RMSEA value (.040). The model fit showed an acceptable decrease relative to that of Model 2 (ΔCFI = −.006, ΔRMSEA = −.002), indicating the presence of partial strong invariance. Building on the partial strong invariance model (Model 3p), we tested latent mean invariance (Model 4) by constraining the factor means to be invariant across the two forms. Again, Model 4 showed suboptimal CFI and TLI values (CFI = .784, TLI = .768) but a good RMSEA value (.040). The model fit showed an acceptable decrease relative to that of Model 3p (ΔCFI = −.001, ΔRMSEA = −.001), indicating the presence of latent mean invariance. We therefore considered the latent mean invariant model (Model 4) to be the best-fitting model, meaning that the factor means were invariant across the stand-alone FFI-3 and the FFI-3 scale extracted from the PI-3.

Psychometric properties of the Chinese FFI-3

Structural validity

We examined the internal consistency of the Chinese FFI-3 using Dataset 2 at Time 1 and Dataset 3 ($N=702$). The coefficient alpha values for the FFI-3 factors were .832 for N, .825 for E, .724 for O, .671 for A, and .809 for C, with a median coefficient alpha value of .809 (C). Compared with the Chinese PI-3, whose coefficient alpha values ranged from .840 (O) to .923 (N), with a median of .877 (E) (Yik et al., 2023), the Chinese FFI-3 showed slightly lower internal consistency, as the number of items used to define each factor was reduced.

Next, we examined the five-factor structure by conducting ESEM. The ESEM solution is presented in the right panel of Table 1 and the model fit statistics for the ESEM solution are presented in the upper panel of Table 2. The

Table 4. Correlations between Chinese FFI-3 and PI-3 factors (Dataset 2).

Factor	FFI-3 (N)	FFI-3 (E)	FFI-3 (O)	FFI-3 (A)	FFI-3 (C)	PI-3 (N)	PI-3 (E)	PI-3 (O)	PI-3 (A)	PI-3 (C)
Time 1										
FFI-3 (N)	--									
FFI-3 (E)	−.30**	--								
FFI-3 (O)	−.07	.17**	--							
FFI-3 (A)	−.10*	.15**	.05	--						
FFI-3 (C)	−.28**	.26**	.17**	.04	--					
PI-3 (N)	**.90****	−.30**	−.10*	−.15**	−.36**	--				
PI-3 (E)	−.33**	**.92****	.24**	.14**	.30**	−.34**	--			
PI-3 (O)	−.08	.24**	**.89****	.07	.09	−.11*	.33**	--		
PI-3 (A)	−.03	.17**	.03	**.87****	.02	−.07	.16**	.05	--	
PI-3 (C)	−.32**	.20**	.23**	.02	**.89****	−.43**	.25**	.16**	−.01	--
Time 2										
FFI-3 (N)	--									
FFI-3 (E)	−.29**	--								
FFI-3 (O)	.00	.15**	--							
FFI-3 (A)	−.11*	.15**	.06	--						
FFI-3 (C)	−.27**	.23**	.19**	.06	--					
PI-3 (N)	**.90****	−.30**	−.07	−.15**	−.39**	--				
PI-3 (E)	−.32**	**.92****	.23**	.14**	.29**	−.34**	--			
PI-3 (O)	−.01	.21**	**.90****	.09	.14**	−.08	.30**	--		
PI-3 (A)	−.05	.17**	.03	**.89****	.03	−.07	.14**	.04	--	
PI-3 (C)	−.33**	.20**	.23**	.05	**.90****	−.46**	.27**	.18**	.01	--

Note. $N=403$. FFI-3 = NEO Five-Factor Inventory 3; PI-3 = NEO Personality Inventory 3; N = neuroticism; E = extraversion; O = openness to experience; A = agreeableness; C = conscientiousness. The correlations between the FFI-3 factors and their corresponding PI-3 domains are presented in bold.
$^{*}p < .05$.
$^{**}p < .01$.

ESEM solution had CFI and TLI values of .798 and .759, respectively, which are lower than the respective thresholds for an acceptable fit, but an RMSEA value of .044, which is acceptable.[5] The factor solution showed that 56 (93%) of the 60 items had their highest loadings on the intended factors, with 39 of the loadings reaching .40 in absolute magnitude at $p < .001$. Only four items did not have their highest loadings on the intended factors: E47, whose highest loading was on C, while O38, A24, and C15 had their highest loadings on N. Compared with the 13 deviations found in the English factor solution, the 4 deviations in the Chinese factor solution provided stronger support for the five-factor model. E47 did not load on E in either solution, instead it loaded on C in both English and Chinese versions.

The deviations in the factor solutions may be due to numerous factors, such as test language/translations, item order, and how the 60 ratings were generated (stand-alone or extracted from the long form). The items are arranged in the same sequence in the Chinese and English versions. As explained in the previous section, we found support for the latent mean invariant model in the English FFI-3, implying that the factor means were invariant across the stand-alone FFI-3 and the FFI-3 extracted from the PI-3. Based on these findings, we suggest that the Chinese-language version gave our Chinese participants a slight advantage in self-assessment because it aligns slightly more closely with the theoretical five-factor model than the English version.

Test–retest reliability

Using Dataset 2 ($N=403$), we examined the two-week test–retest reliability of the Chinese FFI-3. The equivalence correlations (r_{cc}) are presented in the first column of Table 3. The five r_{cc} values ranged from .80 (A) to .89 (E), with a median of .85 (O). As anticipated, these values were lower than those for the Chinese PI-3 reported in Yik et al. (2023), which ranged from .89 (A) to .92 (C), with a median of .90 (O).

Correlations between FFI-3 factors and PI-3 domains

Using Dataset 2 ($N=403$), we examined the correlations between the Chinese FFI-3 factors (each measured by 12 items) and the PI-3 domains (each measured by 6 facets) at Time 1 and Time 2. The results are presented in Table 4. All of the correlations were very similar between the two time points. At Time 1, the correlation coefficients between the five FFI-3 factors ranged from −.30 (N and E) to .26 (E and C); at Time 2, the correlations ranged from −.29 (N and E) to .23 (E and C). Of special interest was the relationship between each FFI-3 factor and its corresponding PI-3 domain. Across the two time points, the correlations ranged from .87 (A) to .92 (E) and were all significant. In other words, a participant will receive extremely similar scores for the five personality factors whether they complete the long or the short form of the instrument.

Equivalence of the Chinese and English FFI-3

Measurement invariance

We examined the validity of the Chinese translation of the FFI-3 by testing a series of measurement invariance models across the English and Chinese versions of the scale, using the five-factor solution as a baseline model. We used data

[5] Similar to the results for the English FFI-3 in Dataset 1 ($N=611$), the less-than-ideal CFI and TLI values for the Chinese FFI-3 ($N=702$) reported here may be due to the fact that the baseline model of uncorrelated manifest variables had already achieved a model fit that left little room for improvement when the five-factor model was fitted to the data, resulting in slightly lower CFI and TLI values (see Lai & Green, 2016).

from Dataset 3 ($N=299$),[6] in which 149 participants (74 women) were randomly assigned to first complete the Chinese version followed by the English version and 150 participants (72 women) were randomly assigned to first complete the English version followed by the Chinese version. Given the repeated measures design, we performed multiple indicator growth modeling and correlated the uniqueness of identical items across the two language versions. The lower part of Table 2 shows the fit indices for the series of invariance models we tested (Marsh et al., 2009; Meredith, 1993).

We began with a configural invariance model (Model 1) that had no invariance constraints across the two languages. This model showed suboptimal CFI and TLI values (CFI = .735, TLI = .708) but a good RMSEA value (.043), indicating the presence of configural invariance. We then tested a metric invariance model (Model 2) in which the factor loadings were held invariant across languages. Model 2 again showed suboptimal CFI and TLI values (CFI = .720, TLI = .704) but a good RMSEA value (.043). Compared with Model 1, Model 2 showed a significantly worse fit, at least in terms of CFI (ΔCFI = −.015, ΔRMSEA = .000). Based on the two highest modification indices, for items E22 and O48, we tested a partial metric invariance model (Model 2p) by allowing all loadings of these two items to vary across languages. Model 2p again showed suboptimal CFI and TLI values (CFI = .726, TLI = .710) but a good RMSEA value (.043). The model fit showed an acceptable decrease relative to that of Model 1 (ΔCFI = −.009, ΔRMSEA = .000), indicating the presence of partial metric invariance (Chen, 2007).[7] For E22 and O48 whose loadings could vary across languages, their primary and cross-loadings showed important differences between the English and Chinese versions. E22 showed a high loading on E (.78, $p < .001$) for the Chinese version but a low loading on E (.23, $p < .001$) for the English version. The cross-loading of E22 on C was significant for the Chinese version (−.12, $p < .05$) but trivial for the English version (.07). Similarly, O48 showed a high loading (.65, $p < .001$) on O for the Chinese version but a low loading (.28, $p < .01$) on O for the English version. The cross-loading of O48 on A was trivial for the Chinese version (−.11) but significant for the English version (.21, $p < .05$).

Based on the partial measurement invariance model (Model 2p), we tested a strong invariance model (Model 3) in which the item intercepts, along with the factor loadings, were held invariant across the two languages. Support for strong invariance would indicate that the intercepts were the same across groups, such that differences in the observed means of the 60 items could only be explained by differences in the latent means of the five factors. Initially, the strong invariance model could not converge, due to the negative residual variances of items E22 and O48, which were the same two items that had non-invariant loadings in both languages. The negative residual variance and non-convergence may be due to the use of a relatively small sample for the complex model (with many factor loadings and factor covariances to be estimated). We then set the residual variances of these two items above 0 to enable the model to converge. Similar to Model 2p, Model 3 showed suboptimal CFI and TLI values (CFI = .684, TLI = .668) but a good RMSEA value (.046). However, Model 3 showed a significantly worse fit than Model 2p (ΔCFI = −.042, ΔRMSEA = .003). Based on the highest modification indices of items O53, E27, and O13, we tested a partial strong invariance model (Model 3p) by allowing the intercepts of these three items to be non-invariant across languages. Again, Model 3p showed suboptimal CFI and TLI values (CFI = .691, TLI = .675) but a good RMSEA value (.045). However, it also showed a significantly worse fit than Model 3 (ΔCFI = −.035, ΔRMSEA = .002). Although the change in RMSEA provided some indication of partial strong invariance, this was not confirmed by the other fit statistics. Based on the modification indices, we tested another less constrained partial strong invariance model by allowing one more intercept (O58) to vary. However, the model failed to converge due to a non-positive definite latent variable covariance matrix; therefore, the results for this model are not reported in Table 2.

Overall, the partial metric invariance model (Model 2p) best fitted our data. The standardized loadings of this invariance model are presented in Table 5. Based on the factor loadings obtained for the FFI-3, 58 of the 60 items in Chinese were equivalent to their original English versions. The only two exceptions were items E22 (like being where the action is) and O48 (having little interest in the universe or the human condition). Two explanations could account for this nonequivalence. First, the Chinese translations may convey different meanings than the original English items. For instance, the English E22 includes the phrase "where the action is," which refers to a place where something important or exciting is happening. Such a broad definition could include a party, a sports competition, a protest, an election, or even a crime scene. In contrast, the Chinese translation refers to a noisy and crowded place, which could have different connotations. While it could refer to a place "where the action is" (e.g., a party or a sports competition), it could also mean a crowded place where nothing important or exciting is happening, such as a busy shopping mall or

[6] We conducted Monte Carlo simulations using Mplus 8.8 to examine the power of the sample size to assess the measurement invariance of loadings, intercepts, and uniquenesses in the English and Chinese versions of the FFI-3 (Muthén & Muthén, 2017). Given the repeated measures design, we used multiple indicator growth modeling. Following Yik et al. (2023), we performed 5,000 replications using a sample size of 300 participants with the following population values: factor loading of .60 for both languages, cross-loading of .00 for both languages, intercept of 3.00 for both languages, uniquenesses of .50 for Chinese and .55 for English, correlated uniqueness of .40 for identical items across languages, factor means of .00 for Chinese and .05 for English, factor variance of 1.00 for Chinese and 1.10 for English, and factor correlations of .20 for Chinese and .25 for English. The results showed that parameter biases were less than 10%, standard error biases for factor loadings and intercepts were less than 4%, and coverage was between .93 and .96. Taken together, these results showed that the sample size of 299 participants had sufficient power to assess measurement invariance.

[7] We further tested for partial measurement invariance between the two languages by comparing Model 2p with a nested model of the invariant factor variance–covariance matrix using Mplus 8.8 (Muthén & Muthén, 2017). The results provided support for the invariance of the factor variance–covariance matrix across the two languages: ΔCFI = .001 and ΔRMSEA = .000.

Table 5. Standardized loadings for the English and Chinese FFI-3 scales in the final partial measurement invariance model.

	English					Chinese				
Item	N	E	O	A	C	N	E	O	A	C
N01	**.40*****	–.09	.01	.05	.04	.38***	–.08	.01	.04	.04
N06	**.45*****	–.09	.09	–.16**	–.04	**.46*****	–.08	.09	–.15**	–.03
N11	**.57*****	.11	–.02	–.04	.00	**.59*****	.10	–.02	–.04	.00
N16	**.45*****	–.08	.03	.05	–.01	**.56*****	–.09	.04	.06	–.01
N21	**.46*****	–.12**	.07	–.11	.03	**.56*****	–.13**	.08	–.12*	.03
N26	**.46*****	–.16**	.06	–.10	–.11*	**.50*****	–.15**	.06	–.09	–.11*
N31	**.54*****	–.15**	.10	.11*	–.02	**.60*****	–.15**	.10	.11*	–.02
N36	.25***	.02	–.11	–.32***	.02	.27***	.02	–.11	–.31***	.02
N41	**.45*****	.03	.08	–.15*	–.20**	**.50*****	.03	.08	–.15*	–.20**
N46	**.54*****	–.17**	.07	–.02	.00	**.60*****	–.17**	.07	–.02	.00
N51	**.54*****	.08	–.06	–.08	–.16**	**.61*****	.08	–.07	–.08	–.17**
N56	**.46*****	–.08	.03	–.25***	–.09	**.49*****	–.07	.03	–.23***	–.09
E02	.09*	**.70*****	–.11*	.02	–.04	.11*	**.76*****	–.12*	.02	–.04
E07	.11	**.48*****	.18**	.10	.02	.13	**.48*****	.19**	.10	.02
E12	–.03	**.42*****	–.05	.15**	–.09	–.04	**.47*****	–.06	.16**	–.10
E17	–.01	**.63*****	–.02	.08	–.02	–.02	**.63*****	–.03	.08	–.02
E22	.02	.23***	–.06	–.08	.07	.05	**.78*****	–.12	.02	–.12*
E27	–.17**	**.45*****	–.08	.14	–.06	–.20**	**.48*****	–.09	.15	–.07
E32	–.12*	.39***	–.11*	–.10	.17**	–.14*	**.41*****	–.12	–.10	.18***
E37	–.15**	**.61*****	.03	–.08	.16**	–.16**	**.58*****	.03	–.07	.16**
E42	–.02	.32***	.04	.23***	–.02	–.02	.35***	.04	.26***	–.02
E47	.11*	.16**	–.10	–.21***	.37***	.13*	.17**	–.10	–.21***	**.40*****
E52	–.13**	**.62*****	.02	–.15**	.09	–.15**	**.64*****	.02	–.16**	.09
E57	–.04	.30***	–.13*	.18**	.10	–.05	.29***	–.13*	.17**	.10
O03	.06	.14*	.36***	–.19**	–.11	.06	.13*	.35***	–.18**	–.10
O08	.01	.27***	.30***	.05	.17*	.01	.24***	.29**	.05	.16*
O13	.05	–.02	**.55*****	.06	.05	.06	–.02	**.61*****	.06	.06
O18	–.02	–.02	.21**	.23**	.01	–.02	–.02	.25***	.25***	.01
O23	.03	–.08	**.55*****	.11*	–.09*	.03	–.07	**.55*****	.10*	–.09*
O28	–.19***	–.02	.30***	.05	–.07	–.21***	–.02	.31***	.05	–.07
O33	–.02	.03	.22**	.15*	.08	–.03	.04	.26**	.17*	.09
O38	.30***	.23***	.09	.05	.08	.36***	.25***	.10	.05	.09
O43	.01	–.04	**.58*****	.03	–.07	.01	–.04	**.58*****	.03	–.07
O48	–.04	–.11	.28**	.21*	.06	–.06	–.05	**.65*****	–.11	.03
O53	–.07	.12*	**.43*****	–.14*	.26***	–.08	.11*	**.42*****	–.13*	.25***
O58	–.09	–.03	**.45*****	–.18**	.13*	–.11	–.03	**.48*****	–.19**	.14*
A04	.18***	.23***	.09	.18**	.22***	.23***	.27***	.11	.21**	.27***
A09	.04	–.12**	.02	**.55*****	.05	.04	–.12**	.02	**.52*****	.05
A14	–.03	.06	.02	**.67*****	.04	–.03	.06	.02	**.66*****	.05
A19	.03	–.07	.02	.38***	.04	.04	–.07	.02	**.42*****	.04
A24	.22***	–.19**	–.19**	.23**	–.30***	.24***	–.19**	–.20**	.23**	–.31***
A29	–.03	.18**	.14*	.24***	.05	–.03	.20***	.16*	.26***	.05
A34	.01	.36***	.08	.11	.05	.01	.35***	.08	.11	.05
A39	.03	.19***	.03	**.51*****	–.08	.04	.20***	.03	**.54*****	–.08
A44	.13*	.08	.07	.24***	.02	.15*	.09	.08	.26**	.02
A49	.15**	.16**	.10	.27***	.21**	.16**	.15**	.10	.25***	.20**
A54	–.08	–.13*	.06	.21**	–.01	–.10	–.14*	.07	.22***	–.01
A59	–.01	–.08	.06	**.50*****	–.07	–.01	–.08	.06	**.52*****	–.08
C05	.09	–.01	–.03	–.01	.36***	.10	–.01	–.04	–.01	**.41*****
C10	–.01	.01	–.08	.06	**.58*****	–.01	.01	–.10	.06	**.65*****
C15	–.15**	–.10*	–.09	.12	.14**	–.22**	–.13*	–.12	.15*	.19**
C20	.17**	.11*	–.01	.07	**.48*****	.20***	.12*	–.01	.08	**.54*****
C25	–.05	.04	–.05	–.10*	**.62*****	–.06	.04	–.05	–.11*	**.68*****
C30	–.17*	–.15**	–.02	.28***	**.46*****	–.18**	–.14**	–.02	.25***	**.44*****
C35	.08	.03	–.03	.06	**.59*****	.09	.03	–.04	.06	**.63*****
C40	.04	.10	.02	.10	**.45*****	.05	.10	.02	.10	**.45*****
C45	–.23***	.09	–.09	.28***	.26***	–.25***	.09	–.10	.26***	.26***
C50	–.08*	.06	–.04	–.03	**.71*****	–.09*	.06	–.04	–.03	**.74*****
C55	–.13*	–.11*	–.08	.22**	**.54*****	–.14*	–.11*	–.08	.21**	**.53*****
C60	.06	.12*	.08	–.11*	**.55*****	.07	.13*	.09	–.12*	**.59*****

Note. $N = 299$. ESEM = exploratory structural equation modeling; FFI-3 = NEO Five-Factor Inventory 3; N = neuroticism; E = extraversion; O = openness to experience; A = agreeableness; C = conscientiousness. Items N01, N16, N31, N46, E12, E27, E42, E57, O18, O23, O28, O33, O48, A09, A14, A19, A24, A39, A44, A54, A59, C15, C30, C45, and C55 were reverse-coded prior to statistical analyses. Loadings equal to or greater than |.40| are presented in bold. Loadings that were significant only on their intended factors at $p < .001$ are underlined.

$^{*}p < .05$.

$^{**}p < .01$.

$^{***}p < .001$.

street. Such differences in the meanings conveyed could have led to nonequivalence of factor loadings. Another possible explanation is that both E22 and O48 appear to measure hobbies or personal interests rather than personality traits per se. For instance, as discussed above, E22 could cover a variety of activities. While extraverts enjoy social activities, they may not necessarily like watching sports or election footage, or participating in protests. Similarly, O48 describes interest in the universe or the human condition, which is determined more by personal interest than by the extent to which one is open to new ideas. As such, depending on individual interests or hobbies, the ratings on E22 and O48 could vary across participants, regardless of their E and O personality traits.

Reliability estimates

We examined the test–retest reliability of the Chinese and English versions of the FFI-3 over the focal two-week interval using Dataset 3 ($N = 299$). The equivalence correlations (r_{ce}) for the cross-language data are presented in the second column of Table 3. The five factors had correlation coefficients ranging from .70 (A) to .85 (C), with a median of .77 (N).

Discussion

In this study, we tested whether the FFI-3 can be an efficient and effective measure of the FFM to be used in different research contexts in Chinese communities. Using only a quarter (60 items) of the item pool of the long-form PI-3, the short form can be completed in 8 min. How well does this short form perform in terms of psychometric properties? Although its factor structure yielded less-than-ideal values for the comparative indices, such as CFI and TLI, its good RMSEA value and internal consistency are still encouraging for those who are interested in using the short-form FFI-3 to map personality with Chinese samples. If the FFM is considered a unified framework for personality research, the FFI-3 provides a useful tool for establishing its heritability, outcome measures, and comparative findings in Chinese communities.

Using ESEM and multiple indicator growth modeling, we found that the FFI-3 achieved partial metric invariance across the English and Chinese versions. Based on the factor loadings of the FFI-3, 58 of the 60 items in Chinese were equivalent to their original English versions, providing substantial support for the cross-language generalizability of the factor structure of the short form (Meredith, 1993). The cross-language generalizability of the short form was also supported by its high cross-language test–retest reliability coefficients, which ranged from .70 (A) to .85 (C). One major concern with using the short form is that it has lower internal reliability than the long form (Smith et al., 2000). In this study, Cronbach's alpha values for the short form were all above .70, except for that of A (.67). Nevertheless, as a scale should be evaluated for its test–retest reliability and internal consistency (see McCrae et al., 2011), our finding that FFI-3 scores were stable across both languages and over time indicates that the FFI-3 is a valid measure of personality in Chinese populations.

When time is limited, researchers could use the short form to measure the FFM. This version not only saves time but also minimizes fatigue, thereby encouraging respondents to cooperate. Of interest to researchers is whether the FFM scores obtained using the two forms are comparable. For example, would a respondent be assessed as being high in N using both forms? To test the equivalence between the sum scores obtained from the long and short forms, we examined the correlations between them. The values ranged from .87 (A) to .92 (E), implying that the sum scores were highly comparable between the two forms. Of special note here is that the correlations in this study were obtained from the same dataset, and the results imply that the coefficients represented the upper bound estimates. Although our method of reanalyzing the same dataset controlled for the influence of noise related to individual differences, time-varying factors, or memory effects, future studies should validate the results using different datasets to test the equivalence between the two forms.

Our study is not without limitations. First, we only examined the self-report version of the FFI-3, the results of which should be cross-validated using its observer-report version (McCrae & Costa, 1987). For instance, to what extent do self-report ratings on the Chinese FFI-3 agree with observer-report ratings? Does agreement between self-report and observer-report ratings vary by test language? These questions have important implications for the convergent validity of the FFI-3 and warrant research attention. Answers to such questions will also advance our understanding of personality assessment in cross-cultural contexts (Yik, 2024; see also Götz & Yik, 2025).

Relatedly, our research focused on the internal properties of the English and Chinese FFI-3 but did not examine the nomological net of the five factors. Recent meta-analyses have reported that the Big Five factors are differentially related to well-being (e.g., physical and mental health; Beck & Jackson, 2022; Kang et al., 2023), everyday behaviors (e.g., smartphone usage, driving habits; Luo et al., 2023; Marengo et al., 2023), and other life outcomes (e.g., academic or work performance; Zell & Lesick, 2022). Future research should explore how each of the FFI-3 factors relates to such external correlates, thereby providing convergent and discriminant validity of the five factors. As personality essentially covers "a wide range of emotional, interpersonal, experiential, attitudinal, and motivational characteristics of the individual" (Costa & McCrae, 1992, p. 231), each of the FFI-3 factors has great potential to reveal the relationship between people's personality and their everyday behaviors. Testing the nomological net of the FFI-3 will further advance theoretical knowledge on the validity of the Big Five personality constructs.

A second limitation is that we included only undergraduate students in our sample, who are likely to have the highest reading levels in the population. Compared with the third-generation scales, the FFI-3 and the PI-3 were designed for better readability to accommodate a larger number of respondents with greater variation in age and reading levels (McCrae & Costa, 2010). It is thus critical to cross-validate the structural validity of the Chinese translated version in samples beyond the university, such as community samples.

A third limitation is that our ESEM results showed that several items of the FFI-3 had low loadings on their intended factors and had cross-loadings on other factors, especially for O and A. Ideally, a simple structural model without such cross-loadings would allow subscale scores to be used to assess particular personality traits. However, the cross-loadings observed in our study undermine confidence in the use of separate subscales. How should researchers or practitioners interpret the scores on these divergent items? Should items with cross-loadings still be included in subscale scores? These are practical questions that should be explored in future studies to understand the real-world consequences of using the subscales.

Finally, the short form is not intended for clinical assessments. Clinical decisions can have a significant impact on the treatment and well-being of individuals. Accordingly, there is a need for a comprehensive description of the FFM scores and their 30 facet scores. For illustration, the N factor describes an individual's sensitivity to negative events and their tendency to experience negative affect (McCrae & Costa, 2010). Its full spectrum is captured by six facets in the long form (which are missing in the short form): N1: Anxiety, N2: Angry Hostility, N3: Depression, N4: Self-Consciousness, N5: Impulsiveness, and N6: Vulnerability. Although the N factor score may be optimal in research on behavioral correlates of neuroticism, facet scores may better facilitate clinicians' assessments of patients. Scoring very high for N does not necessarily mean that an individual is anxious, hostile, depressed, and sensitive. To provide a finer-grained analysis, it is necessary to use facet scores.

Declaration of interest

No potential conflict of interest was reported by the author(s).

ORCID

Michelle Yik http://orcid.org/0000-0003-0104-3662
Felity H. C. Kwok http://orcid.org/0000-0003-2179-720X
Kim De Roover http://orcid.org/0000-0002-0299-0648

References

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

Axelsson, M., Jakobsson, J., & Carlson, E. (2019). Which nursing students are more ready for interprofessional learning? A cross-sectional study. *Nurse Education Today*, *79*, 117–123. https://doi.org/10.1016/j.nedt.2019.05.019

Beck, E. D., & Jackson, J. J. (2022). A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology*, *122*(3), 523–553. https://doi.org/10.1037/pspp0000386

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

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

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

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology*, *102*(4), 874–888. https://doi.org/10.1037/a0027403

De Fruyt, F., De Bolle, M., McCrae, R. R., Terracciano, A., Costa, P. T., Jr., & Collaborators of the Adolescent Personality Profiles of Cultures Project. (2009). Assessing the universal structure of personality in early adolescence: The NEO-PI-R and NEO-PI-3 in 24 cultures. *Assessment*, *16*(3), 301–311. https://doi.org/10.1177/1073191109333760

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2021). *Ethnologue: Languages of the world* (24th ed.). SIL International. https://www.ethnologue.com/

Falgares, G., Manna, G., Costanzo, G., De Santis, S., Kopala-Sibley, D. C., & Ingoglia, S. (2022). The predictive role of ideological, personality and psychopathological factors in homonegative attitudes in Italy. *Sexuality & Culture*, *26*(1), 339–353. https://doi.org/10.1007/s12119-021-09894-x

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. *2*, pp. 141–165). Sage Publications.

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

Götz, F. M., & Yik, M. (2025). Personality science around the world. *Personality Science*, *6*, 1–5. https://doi.org/10.1177/27000710241290623

Källmen, H., Wennnberg, P., Andreasson, P., & Bergman, H. (2016). Psychometric properties of the Swedish version of the personality inventory NEO-PI-3. *International Journal of Psychology and Psychoanalysis*, *2*(1), 1–3. https://doi.org/10.23937/2572-4037.1510011

Kang, W., Steffens, F., Pineda, S., Widuch, K., & Malvaso, A. (2023). Personality traits and dimensions of mental health. *Scientific Reports*, *13*(1), 7091. https://doi.org/10.1038/s41598-023-33996-1

Kemper, C. J., Trapp, S., Kathmann, N., Samuel, D. B., & Ziegler, M. (2019). Short versus long scales in clinical assessment: Exploring the trade-off between resources saved and psychometric quality lost using two measures of obsessive–compulsive symptoms. *Assessment*, *26*(5), 767–782. https://doi.org/10.1177/1073191118810057

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Krosnick, J. A. (1999). Maximizing questionnaire quality. In J. P. Robinson & L. S. Wrightsman (Eds.), *Measures of political attitudes* (pp. 35–57). Academic Press.

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

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research*, *51*(2-3), 220–239. https://doi.org/10.1080/00273171.2015.1134306

Luo, X., Ge, Y., & Qu, W. (2023). The association between the Big Five personality traits and driving behaviors: A systematic review and meta-analysis. *Accident Analysis and Prevention*, *183*, 106968. https://doi.org/10.1016/j.aap.2023.106968

Marengo, D., Elhai, J. D., & Montag, C. (2023). Predicting Big Five personality traits from smartphone data: A meta-analysis on the potential of digital phenotyping. *Journal of Personality*, *91*(6), 1410–1424. https://doi.org/10.1111/jopy.12817

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2015). The inter-item standard deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences*, *84*, 79–83. https://doi.org/10.1016/j.paid.2014.08.021

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics* (pp. 275–340). Routledge.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu & Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

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

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 439–476. https://doi.org/10.1080/10705510903008220

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

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist*, *52*(5), 509–516. https://doi.org/10.1037/0003-066X.52.5.509

McCrae, R. R., & Costa, P. T., Jr. (2007). Brief versions of the NEO-PI-3. *Journal of Individual Differences*, *28*(3), 116–128. https://doi.org/10.1027/1614-0001.28.3.116

McCrae, R. R., & Costa, P. T., Jr. (2010). *NEO Inventories for the NEO Personality Inventory-3 (NEO-PI-3), NEO Five-Factor Inventory-3 (NEO-FFI-3), NEO Personality Inventory-Revised (NEO PI-R): Professional manual*. Psychological Assessment Resources.

McCrae, R. R., Costa, P. T., Jr., & Martin, T. A. (2005). The NEO-PI-3: A more readable revised NEO Personality Inventory. *Journal of Personality Assessment*, *84*(3), 261–270. https://doi.org/10.1207/s15327752jpa8403_05

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28–50. https://doi.org/10.1177/1088868310366253

McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology*, *70*(3), 552–566. https://doi.org/10.1037/0022-3514.70.3.552

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (4th ed.). Muthén & Muthén.

Paunonen, S. V., & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, *81*(3), 524–539. https://doi.org/10.1037//0022-3514.81.3.524

Perez, R. M., Jr. (2020). *Examining follower perceptions of the relationships between public sector leadership behavior and personality traits* [Doctoral dissertation]. University of the Incarnate Word. The Athenaeum. https://athenaeum.uiw.edu/uiw_etds/372

Quy, G. S. (2011). *Adapting the NEO-PI-3 for a South African context: A pilot study using a South African student population* [Master's thesis]. University of the Witwatersrand. Electronic Theses and Dissertations. http://hdl.handle.net/10539/9752

Rabadi, R. I., & Rabadi, A. D. (2021). Validation of the psychometric properties of the NEO-FFI-3 in an Arabic context. *Psychology Research and Behavior Management*, *14*, 947–956. https://doi.org/10.2147/PRBM.S312829

Rammstedt, B., & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K) [Short version of the Big Five Inventory (BFI-K)]. *Diagnostica*, *51*(4), 195–206. https://doi.org/10.1026/0012-1924.51.4.195

Reyes, M. E., Davis, R. D., Miranda, M. I. D., Figueroa, A. D. R., Sim, K. M. U., & Sunga, M. A. M. (2019). Exploring the relationship between the MHI-38 and the NEO-FFI-3 among Filipinos. *Indian Journal of Health and Wellbeing*, *10*(1–3), 14–20.

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health*, *14*(8), 1101–1108. https://doi.org/10.1016/j.jval.2011.06.003

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

Tourangeau, R., Rips, L. J., & Rasinski, K. (Eds.). (2000). *The psychology of survey response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Yik, M. (2024). Hello, Neihou: Anchoring and adjustment in personality assessment. *Personality Science*, *5*, 1–9. https://doi.org/10.1177/27000710241287657

Yik, M., & Siu, N. Y. F. (2024). Extraverts suffer from social distancing: A 30-day diary study. *Personality and Individual Differences*, *218*, 112433. https://doi.org/10.1016/j.paid.2023.112433

Yik, M., & Siu, N. Y. F. (2025). Who thrives in a public health crisis? *Acta Psychologica*, *253*, 104636. https://doi.org/10.1016/j.actpsy.2024.104636

Yik, M., Sze, I. N. L., Kwok, F. H. C., & Lin, S. (2023). Mapping Chinese personality: An assessment of the psychometric properties of the NEO-PI-3 in monolingual and bilingual studies. *Assessment*, *30*(7), 2031–2049. https://doi.org/10.1177/10731911221126921

Zell, E., & Lesick, T. L. (2022). Big five personality traits and performance: A quantitative synthesis of 50+ meta-analyses. *Journal of Personality*, *90*(4), 559–573. https://doi.org/10.1111/jopy.12683

Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales – Five misunderstandings and ways to overcome them. *Journal of Individual Differences*, *35*(4), 185–189. https://doi.org/10.1027/1614-0001/a000148