When personal becomes relational

Investigating the role of personality traits in affective student-teacher relationships

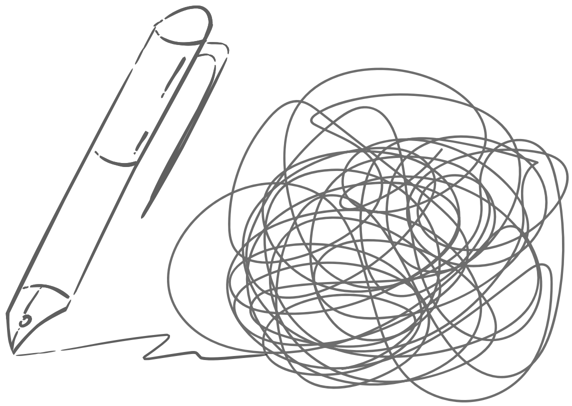
Du, Q.

Publication date
2025

Chapter 3

Mapping Young Minds

Validating a Brief Big Five Inventory for Chinese Children



Published as: Du, Q., Roorda, D. L., Koomen, H. M. Y., & Zee, M. (2025). Mapping young minds: Adapting and validating a brief Big Five Inventory for Chinese children in upper elementary school. *Social Development, 34*(4). https://doi.org/10.1111/sode.70022

Abstract

This paper describes the development and initial validation of a brief Big Five personality questionnaire for middle childhood children in China. Two independent samples, comprising 1,884 children and 2,245 children, were used to pilot and cross-validate the questionnaire. We first explored the underlying factor structure and composite reliability of the questionnaire, which included five hypothesized factors: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness. Additional evidence for validity was obtained by examining measurement invariance across gender, age, and grade level. Results revealed that the hypothesized five-factor model demonstrated satisfactory model fit. Each factor demonstrated sufficient to satisfactory composite reliability (ρ_c range = .71–.80). Regarding measurement invariance, partial scalar invariance was found across gender, and scalar invariance was found for age and grade level. Latent mean comparisons indicated that girls reported higher agreeableness and neuroticism but lower openness than boys. Age-related differences were only identified for conscientiousness, which decreased with age. Children in Grade 5 exhibited lower agreeableness, conscientiousness, extraversion, and openness than their peers in Grade 4 and Grade 6. Accordingly, this brief Big Five personality questionnaire can be seen as a valid and reliable measure that enriches the methodological repertoire for personality research and allows the investigation of personality traits among upper elementary school children.

Keywords: personality, measurement invariance, Big Five, middle childhood, China

Introduction

A wealth of empirical research has highlighted how personality traits can be powerful drivers of long-term success and well-being (John, 2021; Mäkikangas et al., 2013). Specifically, personality traits have been shown to impact various important life outcomes, such as academic success, work engagement, and social relationships (see Harris & Vazire, 2016; Kim et al., 2019; Poropat, 2009 for reviews). Accordingly, understanding those personality traits is essential, as they may fluctuate with life experiences, leading their impact to vary across different life stages (Shiner & DeYoung, 2013; Soto et al., 2008).

Early empirical research on personality traits has primarily focused on adulthood (Barbaranelli et al., 2003; Shiner & DeYoung, 2013). However, recent studies have highlighted that these traits can be relatively reliably measured as early as childhood, particularly in middle childhood (Barbaranelli et al., 2003; Brandt et al., 2020). Unfortunately, less is known about the role of personality traits in middle childhood compared to adulthood, as existing questionnaires are not specifically tailored for children and tend to be long and time-consuming to administer (Barbaranelli et al., 2003; Kupper et al., 2021). Moreover, most studies have been conducted in Western countries (Bleidorn et al., 2022; Shiner & Caspi, 2003), underscoring the need for validated questionnaires in Eastern contexts, such as China, to further understand the role of personality traits across cultures. In this study, we aimed to validate a personality questionnaire for children during middle childhood (aged from 9 to 14) in an Eastern context (i.e., China).

Personality traits are typically defined as the characteristic patterns of behaviors, feelings, and thoughts of individuals across situations and over a period of time (Shiner & Caspi, 2003). Decades of research have been dedicated to understanding and conceptualizing personality traits (John, 2021). Although there is an ongoing debate about the optimal number of factors that best describe the structure of personality, the five-factor structure has gained the most widespread support (John, 2021). Regardless of the investigative methods, languages, or informants used, five core traits have been shown to consistently reflect personality (Barbaranelli et al., 2003; Shiner & DeYoung, 2013). Specifically, these five traits are agreeableness, conscientiousness, extraversion, neuroticism, and openness. According to John (2021), agreeableness refers to how friendly, considerate, and easygoing individuals are. Conscientiousness refers to what degree individuals are responsible, organized, and rule-binding. Extraversion reflects the extent to which individuals like highly stimulating environments and enjoy social interactions. Neuroticism refers to how likely individuals are to be temperamental and experience negative emotions, such as stress and anxiety. Last, openness reflects the extent to which individuals are curious about different things and are willing to take on new experiences (John, 2021).

Although most research has been focused on adults (Shiner & Caspi, 2003), the Big Five model seems to be effective in describing children's personality traits as well (Barbaranelli et al., 2003; Brandt et al., 2020; Soto et al., 2008). These five traits consistently emerge in both free descriptions (e.g., parents' descriptions of their child) and behavioral ratings of children's personality traits (e.g., ratings by teachers), regardless of whether parents, teachers, or children themselves report on these personality traits (Halverson et al., 2003; Hendriks et al., 2008; Mervielde et al., 1995). Furthermore, although some traits in middle childhood may differ somewhat from those of adults due to children's still-developing behavioral repertoire (Shiner & Caspi, 2003), research has established connections between children's personality traits and important developmental outcomes such as school achievement and behavioral problems (Barbaranelli et al., 2003; Hendriks et al., 2008). This further supports the applicability of the Big Five model to the personality structure of children.

To understand personality traits during middle childhood, various instruments have been developed to measure the Big Five traits in this age group. Some commonly used measures include the Big Five Questionnaire for Children (BFQ-C; Barbaranelli et al., 2003), the Five-Factor Personality Inventory (FFPI; Hendriks et al., 2008), the Hierarchical Personality Inventory for Children (HiPIC; Mervielde & De Fruyt, 1999), and the NEO Personality Inventory–3 (NEO-PI-3) along with the Revised NEO Personality Inventory (NEO-PI-R; De Fruyt et al., 2009). In general, the reliabilities for these instruments were found to be satisfactory (De Fruyt et al., 2000; Hendriks et al., 2008; Rossier et al., 2007). Moreover, the construct and predictive validity have been supported by significant associations with key developmental outcomes, such as academic performance, social adjustment, and behavioral problems (e.g., Barbaranelli et al., 2003; Hendriks et al., 2008).

Despite advancements in psychological assessments for children, three primary challenges remain in developing effective questionnaires to measure personality traits in middle childhood. First, existing questionnaires tend to be lengthy, typically consisting of 60 items or more. For instance, the HiPIC includes 144 items and the

FFPI includes 100 items (Hendriks et al., 2008; Mervielde & De Fruyt, 1999). Such lengthy surveys can be burdensome for younger children, as they are more likely than adults to experience cognitive overload due to their limited attention span. Moreover, administering long questionnaires in large-scale studies is often impractical, as surveys typically assess multiple variables, leaving limited time for measuring personality traits (Kupper et al., 2021).

Second, existing personality questionnaires may include items that are challenging for middle-childhood children, as many of these questionnaires were validated with adolescents aged 12 and above. There is evidence that even adolescents may encounter difficulty with these questionnaires (De Fruyt et al., 2000; Ober et al., 2021). For instance, adolescents aged 12 to 17 reported that certain items in the NEO-PI-R, particularly those measuring openness, were difficult to understand (De Fruyt et al., 2000). Similarly, the Big Five Inventory (BFI-2) demonstrated suboptimal model fit in adolescent samples, indicating a need for refinement to suit younger age groups (Ober et al., 2021). These findings highlight the necessity of adapting or developing questionnaires to ensure they are age-appropriate and can accurately capture personality traits in middle childhood.

Third, most personality questionnaires have been developed and validated in Western countries, with far fewer questionnaires adapted for Eastern cultures (Shiner & DeYoung, 2013). Such an imbalance may obstruct the ongoing investigation of cultural differences in personality traits. For instance, a large-scale study across 24 countries found that although the trait openness was well-represented in many Western cultures, it was not consistently replicated in certain Eastern contexts, including China (De Fruyt et al., 2009). Additionally, although certain personality dimensions (e.g., interpersonal relatedness) appear to be culturally universal, their significance may vary across cultures (Lin & Church, 2004). These findings underscore the importance of rigorously validating personality measures from Western contexts to ensure their reliability and validity in Eastern settings (Shiner & Caspi, 2003).

The Present Study

In the present study, we aimed to develop and validate a brief personality questionnaire suitable for children to self-report on their own personality traits in an Eastern context (i.e., China). Three specific aims were addressed. First, we developed the questionnaire and investigated its factor structure as well as composite reliability using two independent samples. We first developed an initial version of the questionnaire and validated it with our pilot sample. Subsequently, based on the results in the pilot study, we improved the questionnaire and cross-validated it using our main sample. Based on previous studies (John, 2021; Soto & John, 2017), we hypothesized that the questionnaire would reveal the well-established five-factor structure of personality traits (i.e., agreeableness, conscientiousness, extraversion, neuroticism, and openness) and demonstrate adequate reliability for each factor.

Second, with our main sample, we investigated measurement variance and latent mean comparisons across gender, age, and grade level. Measurement invariance assesses whether the structure of our personality questionnaire is consistent across personal factors (Dong & Dumas, 2020; Putnick & Bornstein, 2016), ensuring meaningful comparisons of group means. Despite its importance, it remains underutilized in research on young children's personality traits. A systematic review suggested that among the 75 studies examining measurement invariance in personality measures, only eight studies focused on elementary or secondary school-aged children (Dong & Dumas, 2020). To address this gap, we tested three levels of measurement invariance (i.e., configural, metric, and scalar invariance) across gender, age, and grade level. Gender and age were examined concerning their potential influence on personality traits through socialization processes (e.g., societal expectations) and developmental changes (e.g., biological changes; Hendriks et al., 2008; Rossier et al., 2007). Grade level was included to account for differences in educational and social experiences. Based on previous studies (see Dong & Dumas, 2020 for a review), we expected to find at least metric measurement invariance across gender and age. The analysis across grade levels was exploratory because of the limited knowledge on this topic.

Third, if (partial) scalar measurement invariance was established, indicating that personality traits were interpreted similarly across groups, we proceeded to compare latent mean differences across gender, age, and grade level. Based on previous studies (e.g., Hendriks et al., 2008; Rossier et al., 2007), we hypothesized that girls would score higher than boys on all five traits. For age differences, we expected no significant changes, as minimal differences have been observed before age 15 (Bleidorn et al., 2022). Because of limited evidence, we approached grade level differences as exploratory.

Method

Sample

The pilot sample consisted of 1,884 children (44.9% girls, 54.0% boys). These children were from grade four (29.3%), grade five (26.4%), and grade six (44.4%). Children's mean age was 11.46 years (*SD* = 0.97; range = 9 to 14 years). Regarding ethnicity, 97.0% of the participating children identified themselves as belonging to the Han group (i.e., the ethnic majority group in China), 2.2% identified themselves as belonging to an ethnic minority group, and 1.0% did not provide this information.

Our main sample included 2,245 children (44.4% girls, 53.3% boys), from Grade 4 (34.7%), Grade 5 (36.9%), and Grade 6 (28.50%). Children were on average 11.44 years old (*SD* = 0.97; range = 9 to 14 years). Regarding ethnicity, most of children identified themselves as belonging to the Han ethnic group (87.5%), 2.4% identified themselves as belonging to an ethnic minority group, and 10.1% did not report their ethnic background.

Procedure

Data collection for the pilot sample took place from March to May 2023, and for the main sample from March to May 2024. Ethical approval was obtained from the Ethics Committee of the University of Amsterdam prior to the data collection. For both the pilot and the main samples, schools in Fujian, a southeastern province of China, were contacted through the first author's social network. Principals of these schools were contacted via telephone and were invited to participate in this study. Upon their agreement, parents received an information letter about the study. Children only refrained from participation in this study when parental consent was not provided.

During scheduled school visits, children completed a paper questionnaire covering demographic background (e.g., age, gender), their own personality traits, and other variables beyond the scope of this study. Before children filled in the questionnaire, the first author or trained research assistants provided brief instructions about the study. While children filled out the questionnaire, the first author or trained research assistants were available to answer questions from children. On average, it took about 20 minutes for children to complete the questionnaire.

Instrument Development (Pilot Sample)

To develop a brief personality questionnaire, we chose the shortened form of the BFI-2 (BFI-2-S; Soto & John, 2017) as the foundation for our questionnaire. The BFI-2-S is a brief measure with 30 items (Soto & John, 2017). This number of items allows for the inclusion of additional variables in the same survey without sacrificing measurement quality (Morizot, 2014). Furthermore, the BFI-2-S uses concrete descriptions (e.g., "Keeps things neat and tidy") instead of adjectives (e.g., "disorganized"), making it easier for children who typically have a more limited vocabulary than adults (Barbaranelli et al., 2003).

Despite its brevity and the use of concrete descriptions, the BFI-2-S may need adaptation to better suit children in middle childhood (Ober et al., 2021). Therefore, we adapted the BFI-2-S in three ways. First, we reformulated the items from the third to the first-person singular, making them less cognitively demanding for children (Hendriks et al., 2008). Second, complex items were simplified or replaced with alternatives from other personality questionnaires or newly developed items while retaining their original meaning. For example, "Is compassionate, has a soft heart" was replaced by "I empathize with others" from FFFI (Hendriks et al., 2008), and "Is original, comes up with new ideas" was simplified to "I come up with new ideas". Three additional items for Agreeableness, Neuroticism, and Openness were added to increase the chance that a reliable questionnaire could be found. Third, we adopted a response scale that has been found to be suitable for this age group (Koomen & Jellesma, 2015). Specifically, we replaced the original scale (1 = "*disagree strongly*" to 5 = "*strongly agree*" with a new 5-point scale (1 = "*NO, that is not true*" to 5 = "*YES, that is true*").

Items were initially developed in Dutch by the third and fourth authors. We applied a back-translation procedure to translate the items into Chinese. To ensure the items were clear and understandable for the target children, we invited a small group of children to review the questionnaire. Altogether, the initial version of the adapted BFI-2-S for the pilot study included 33 items.

Data Analysis

Analyses were performed in M*plus* version 8.8 (Muthén & Muthén, 1998–2017). As children were nested within classrooms, we calculated intraclass correlation coefficients (ICCs) to determine the necessity of controlling for nesting. In both samples, ICC values ranged from .007 to .060, indicating minimal clustering effects. Consequently, nesting effects were not accounted for. Robust maximum likelihood estimation (MLR) was used to account for the non-normality in the data.

We validated the questionnaire in three steps. First, in both the pilot study and the main study, Confirmatory Factor Analysis (CFA) was used to investigate the factor structure of the questionnaire (Aim 1). We evaluated the hypothesized five-factor model and two alternative models, one-factor model (with all items loading on a general personality factor) and hierarchical model (adding a second-order variable of general personality factor on top of the hypothesized five factors; cf., Musek, 2007) to establish the best-fitting factor structure. Subsequently, composite reliability of the factors was estimated (Brown, 1989).

Second, using the main sample, we conducted multigroup analyses to investigate measurement invariance across gender, age, and grade level (Aim 2). Regarding gender, we distinguished between boys (n = 1197) and girls (n = 996). For age, we specified three groups: 9–10-year-olds (n = 312), 11–12-year-olds (n = 1364), and 13–14-year-olds (n = 264). For grade level, three categories were included: Grade 4 (n = 778), grade 5 (n = 828), and grade 6 (n = 639), respectively. Examining measurement invariance proceeded in three steps (Yoon & Millsap, 2007). First, we built a model in which all factor loadings and intercepts were freely estimated across groups (configural invariance). Second, the factor loadings were constrained to be equal across groups (metric invariance). Third, the intercepts were also constrained to be equal across groups (scalar invariance; Yoon & Millsap, 2007). If measurement invariance was not established, we identify items that displayed non-invariance and removed the equality constraints on these items, which led to partial measurement invariance (Byrne et al., 1989).

Third, after establishing at least partial scalar measurement invariance (Byrne et al., 1989), we compared latent mean differences across groups to investigate gender, age, and grade-level differences in children's personality traits (Aim 3). We fixed the latent mean of the reference groups to zero and used Wald z-tests to indicate whether there were significant differences in the latent means across groups.

To evaluate model fit, we used chi-square tests (χ^2), the root mean square error of approximation (RMSEA), the Comparative fit index (CFI), and the Standardized Root Mean Square Residual (SRMR). Non-significant χ^2-values, RMSEA and SRMR-values below 0.08, and CFI-values above 0.90 are considered to indicate satisfactory model fit (Bentler, 1992; Hu & Bentler, 1999). To compare the fit indices of nested models, we employed Satorra–Bentler scaled chi-square difference test (TRd; Satorra & Bentler, 2010), with non-significant chi-squares indicating equivalent fit. We also considered changes in CFI, RMSEA and SRMR as alternative indices. Models were considered to have equal fit when at least three out of four criteria were met: non-significant $\Delta\chi^2$-values, ΔCFI values $\leq$ 0.010, ΔRMSEA values $\leq$ 0.015, and ΔSRMR values $\leq$ 0.030 for metric invariance and $\leq$ 0.010 for scalar invariance (Chen, 2007).

All relevant code has been made publicly available on the Open Science Framework and can be accessed at https://osf.io/svg6j/.

Results

Descriptive Statistics

Descriptive properties of the items are presented in Table 1. For most items (particularly those for Agreeableness, Conscientiousness, Extraversion, and Openness), mean scores were above the midpoint of the scale (M_{range} = 2.22 to 4.45), suggesting a tendency toward positive responses. Standard deviations indicated moderate variability overall, with some items showing higher level of dispersion (SD_{range} = 0.84 to 1.36). Regarding skewness, all items fell within an acceptable range of ± 2 (skewness$_{\text{range}}$ = –1.78 to 0.69). However, some items exceeded the threshold of ± 2 for kurtosis (kurtosis$_{\text{range}}$ = –1.06 to 3.37), indicating a heavy-tailed distribution of these items.

Factor Structure and Reliability (Pilot Sample)

Table 2 provides the fit statistics for the different factor models in the pilot sample. As can be seen in Table 2, the five-factor model fitted the data significantly better than the one-factor model (ΔTRd (10) = 669.248, $p < .001$) and the hierarchical model (ΔTRd (5) = 67.165, $p < .001$). Moreover, the five-factor model had slightly better RMSEA, CFI, and SRMR values compared to the one-factor model and the hierarchical model. Therefore, the five-factor model was chosen as the best-fitting model.

Table 1

Descriptive Statistics of the Items and Scales for the Final Models of the Pilot Sample and the Main Sample

		Pilot Sample					Main sample				
		M	*SD*	Skewness	Kurtosis	Factor loadings	*M*	*SD*	Skewness	Kurtosis	Factor loadings
Agreeableness (7 items)											
A1	I empathize with others.	3.57	1.05	−0.47	−0.11	0.34	3.66	1.00	−0.50	−0.10	0.44
A2_R	I am rude to others.	4.45	0.84	−1.78	3.31	0.52	4.48	0.81	−1.76	2.94	0.41
A3	I believe that most people can be trusted.	3.63	1.18	−0.60	−0.46	0.52	3.62	1.17	−0.56	−0.49	0.52
A4_R	I only think about myself.	4.37	0.92	−1.59	2.23	0.40	4.37	0.91	−1.62	2.59	0.46
A5	I get along well with others.	3.97	0.98	−0.96	0.74	0.61	4.13	0.96	−1.13	1.05	0.64
A6	I treat other people with respect.	4.30	0.87	−1.25	1.42	0.70	4.34	0.84	−1.30	1.71	0.63
A7	I tend to take the others into account.	3.60	1.04	−0.49	0.05	0.52	3.55	1.06	−0.38	−0.30	0.63
Conscientiousness (8 items)											
C1_R	I tend to be messy.	3.98	1.01	−0.77	−0.01	0.51	3.74	1.02	−0.48	−0.29	0.49
C2_R	I find it difficult to start new tasks.	4.00	1.05	−0.86	0.12	0.46	4.10	1.00	−0.99	0.50	0.53
C4_R	I don't do my best for everything.	3.75	1.14	−0.66	−0.32	0.43	3.79	1.11	−0.69	−0.26	0.44
C5	I like keeping my things in order.	3.96	1.14	−0.89	−0.07	0.49	3.98	1.13	−0.87	−0.09	0.50
C6	If I start something, I will work until I get it done.	3.91	0.97	−0.64	−0.06	0.67	3.88	0.99	−0.58	−0.21	0.69
C7	I do things carefully.						3.35	1.06	−0.19	−0.45	0.65
C8	I do things according to my plan.						3.66	1.10	−0.50	−0.45	0.59
C9	I do things efficiently.						3.12	1.06	−0.07	−0.42	0.67
Extraversion (5 items)											
E3	I am full of energy.	4.16	1.05	−1.21	0.85	0.62	4.13	1.01	−1.06	0.55	0.57
E4	I like having lots of people around me.	3.45	1.33	−0.39	−1.00	0.47	3.46	1.32	−0.40	−0.98	0.53
E6_R	I am less active than the other children.	4.00	1.17	−0.96	−0.08	0.42	4.07	1.11	−1.06	0.35	0.38
E7	I like making new friends.						3.99	1.19	−0.94	−0.17	0.68
E10	I like working with the others						3.65	1.18	−0.57	−0.53	0.70

		Pilot Sample					Main sample				
		M	*SD*	Skewness	Kurtosis	Factor loadings	*M*	*SD*	Skewness	Kurtosis	Factor loadings
Neuroticism (6 items)											
N1	I tend to be worried.	3.08	1.21	−0.10	−0.74	0.41	3.05	1.25	−0.07	−0.91	0.57
N2	I feel sad easily.	2.22	1.12	0.58	−0.39	0.65	2.36	1.23	0.56	−0.66	0.74
N3	I get upset easily.	2.60	1.25	0.41	−0.78	0.67	2.58	1.26	0.40	−0.82	0.72
N6	I can suddenly become angry.	2.32	1.36	0.69	−0.73	0.71	2.04	1.17	0.93	0.01	0.56
N7	I am emotional.	2.66	1.33	0.41	−0.92	0.66	2.68	1.28	0.37	−0.85	0.59
N8	I get nervous easily						2.90	1.32	0.08	−1.06	0.59
Openness (6 items)											
O2	I like thinking about difficult things.	3.36	1.16	−0.24	−0.63	0.63	3.25	1.15	−0.18	−0.59	0.57
O3	I come up with new ideas.	3.45	1.15	−0.30	−0.64	0.61	3.44	1.15	−0.28	−0.67	0.67
O6	I like to think things through carefully.	4.01	1.01	−0.83	0.14	0.63	3.91	1.00	−0.65	−0.07	0.64
O7	I am creative.	3.56	1.21	−0.43	−0.75	0.50	3.65	1.17	−0.54	−0.52	0.56
O8	I am curious about new things						4.11	1.01	−1.05	0.58	0.47
O9	I like trying new things						3.84	1.13	−0.67	−0.36	0.65

Note. Reversed items were noted with “R” in the first column.

Table 2

Goodness-of-Fit Statistics from the Confirmatory Factor Analyses for the Pilot Sample and the Main Sample

	χ^2 (*df*)	RMSEA (90% CIs)	CFI	SRMR	TRd (*df*)	ΔRMSEA	ΔCFI	ΔSRMR
Pilot Sample								
Five-factor model	4030.396 (485)**	.062 (.061, .064)	.693	.078	/	/	/	/
One-factor model	5038.495 (495)**	.070 (.068, .072)	.607	.077	669.248 (10)**	.008	−.086	−.001
Hierarchical model	4104.204 (490)**	.063 (.061, .064)	.687	.079	67.165 (5)**	.001	−.006	.001
Final model	2169.638 (339)**	.054 (.051, .056)	.811	.059	1860.758 (151)**	−.008	.118	−.019
Main Sample								
Five-factor model	4323.269 (654)**	.050 (.049, .051)	.803	.065	/	/	/	/
One-factor model	7658.372 (664)**	.068 (.067, .070)	.625	.083	2524.513 (10)**	.018	−.178	.018
Hierarchical model	4597.264 (659)**	.052 (.050, .053)	.789	.068	257.477 (5)**	.002	−.014	.003
Final model	2384.225 (451)**	.044 (.042, .045)	.878	.055	2139.949 (203)**	−.006	.075	−.01

Note. All models were compared with the five-factor model; ** p < .001.

Nevertheless, the absolute CFI value (0.693) indicated a suboptimal fit for the five-factor model. Further inspection of the factor loadings revealed that three items had negative factor loadings ("I can easily see mistakes of the others" for Agreeableness; "I am a quiet person" and "I like it when others tell me what to do" for Extraversion). Removing these items led to an improvement in model fit (χ^2 (395) = 3396.896, p < .001; RMSEA = .064, 90% CI [.052, .066]; CFI = .723; SRMR = .077). Moreover, modification indices suggested that two items, "I am confident" and "I can handle stress well", which were intended to load on Neuroticism, may be relevant for Extraversion or Conscientiousness. Removing these two items resulted in a further improvement of model fit (χ^2 (340) = 2598.385, p < .001; RMSEA = .059, 90% CI [.057, .062]; CFI = 0.767; SRMR = .063). Finally, two items for Openness seemed to be alike (i.e., "I like doing handcraft" and "I like drawing"), adding a residual covariance between these two items further improved model fit (χ^2 (339) = 2169.638, p < .001; RMSEA = .054, 90% CI [.051, .056]; CFI =.811; SRMR = .059). Table 1 displays the factor loadings for the final model of the pilot study.

Concerning composite reliability, the internal consistency for Neuroticism achieved an acceptable level (ρ_c = 0.76), whereas it was below the recommended threshold for Agreeableness, Conscientiousness, Extraversion, and Openness (ρ_c = .69, .69, .53 and 0.64, respectively; Brown, 1989). Additionally, the absolute value of the correlations between factors ranged from .22 to .91.

Factor Structure and Reliability (Main Sample)

As the model fit was not optimal (CFI < 0.90) and the reliability of the questionnaire in the pilot study was not adequate for most factors, we added new items to improve the questionnaire for the main study. Specifically, three items for Conscientiousness, three items for Extraversion, one item for Neuroticism, and two items for Openness were added.

Using the main sample, the results again confirmed that the five-factor model fitted the data better than the two competing models (See Table 2 for results of model comparisons). Specifically, the TRd difference tests indicated that the one-factor model (ΔTRd (10) = 2524.513, p < .001) and the hierarchical model (ΔTRd (5) = 257.477, p < .001). Additionally, the hypothesized model showed better fit indices than the alternative models, indicated by lower RMSEA and SRMR values and a higher CFI.

However, the absolute CFI value was again below the optimal threshold (CFI =.803). Upon inspection of the factor structure, six items had factor loadings below .40.

Three of these items were related to Extraversion items “I am talkative”, “In a group, I like deciding what we are going to do”, and “I keep my feelings and thoughts to myself”; and three items for Openness, “I like music”, “I like doing handcraft”, and “I like painting/drawing”, respectively. Removing these six items led to an increase in the model fit (Model 2.1; χ^2 (454) = 2942.812, p < .001; RMSEA = .049, 90% CI [.048, .051]; CFI = .843; SRMR = .058).

Moreover, modification indices suggested that one item for Conscientiousness would be more appropriately load on Agreeableness. The item “I tend to take the others into account”, which was designed to capture children’s sense of responsibility toward others, may instead reflect care and consideration for others, aligning more closely with Agreeableness. Loading this item on Agreeableness resulted in a better model fit (χ^2 (454) = 2766.596, p < .001; RMSEA = .048, 90% CI [.046, .049]; CFI = .854; SRMR = .056). Additionally, three residual covariances were added to account for potential item interrelations not fully explained by the factors. After these adjustments, the fit statistics for the final model were: χ^2 (451) = 2384.225, p < .001; RMSEA = .044, 90% CI [.042, .045]; CFI = .878; SRMR = .055.

The final model consisted of 32 items, distributed across Agreeableness (7 items), Conscientiousness (8 items), Extraversion (5 items), Neuroticism (6 items), and Openness (6 items). Factor loadings for the included items were all statistically significant, with loadings above .40 for all items except one item for Extraversion, which had a loading of .38 (see the last column of Table 1 for factor loadings with the cross-validation sample). The correlations between the factors were generally satisfactory ($|r|_{\text{range}}$ = .21 to 81). Composite reliabilities for Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness were acceptable to satisfactory (ρ_c = .74, .80, .71, .80, .77, respectively; Brown, 1989). Table 1 displays the full list of items included in the final model.

Measurement Invariance (Main Sample)

Table 3 presents the results of the multigroup analyses for gender, age, and grade level). For gender, constraining the factor loadings across groups (metric invariance), led to a significant TRd difference (ΔTRd (27) = 60.418, p < .001). However, as the changes in the other fit indices were below the threshold (see Table 3), we considered that the metric invariance model fitted the data equally well as the configural model. Constraining the intercepts across gender (scalar invariance) also led to a significant TRd difference (ΔTRd (27) = 204.208, p < .001) as well as a change in the CFI value slightly exceeding the optimal threshold (ΔCFI = –.011), as such scalar invariance could not be established. Inspection of the modification indices revealed that intercepts for two Agreeableness items, "I am rude to others" (reversed item) and "I treat other people with respect", could not be constrained across groups. The intercept for girls was higher than for boys on both items. The model with these two intercepts being freely estimated had an equal model fit as the metric invariance model, indicating partial scalar invariance.

In terms of age, constraining the factor loadings to be equal across age groups led to a significant change in TRd (ΔTRd (54) =72.279, $p < .05$), nevertheless, the changes in CFI, RMSEA, and SRMR were all below the thresholds, supporting metric invariance. Moreover, constraining the intercepts did not significantly change the TRd value (ΔTRd (54) = 65.065, *ns*), the changes in CFI, RMSEA, and SRMR were all below the thresholds. These findings suggest that the criteria for both metric and scalar invariance across age groups were met.

Regarding measurement invariance across grade level, although TRd tests were significant (ΔTRd (54) ≥ 93.279, *p*s < .05), the alternative fit indices (RMSEA, CFI, SRMR) generally supported both metric and scalar measurement invariance (see Table 3).

Latent Mean Comparisons (Main Sample)

As meaningful comparisons in mean levels could still be made when two or more intercepts per factor can be constrained across groups (Byrne et al., 1989), we proceeded with latent mean comparisons for gender, age, and grade level. Table 4 presents the results of these comparisons. Regarding gender differences, significant differences between boys and girls were found for three out of five traits. Specifically, girls scored higher on Agreeableness ($\Delta M = 0.30$, $Z = 4.97$, $p < .001$) and Neuroticism ($\Delta M = 0.21$, $Z = 4.45$, $p < .001$), and lower on Openness than boys ($\Delta M = -0.25$, $Z = -4.82$, $p < .001$). No gender differences were observed for Conscientiousness and Extraversion.

Table 3

Model Fit Statistics for Measurement Invariance Across Gender, Age, and Grade Level in the Main Sample

	χ^2 (*df*)	RMSEA	CFI	SRMR	TRd (*df*)	ΔRMSEA	ΔCFI	ΔSRMR
Gender								
Configural	2748.631 (902)**	.043	.882	.056				
Metric	2811.157 (929)**	.043	.879	.059	60.418 (27)**	.000	−.003	.003
Scalar	3018.405 (956)**	.044	.868	.062	204.208 (27)**	.005	−.011	.003
Partial scalar	2979.107 (954)**	.044	.870	.061	165.023 (25)**	.001	−.009	.002
Age								
Configural	3317.016 (1353)**	.047	.867	.062				
Metric	3391.062 (1407)**	.047	.866	.065	72.279 (54)*	.000	−.001	.003
Scalar	3467.140 (1461)**	.046	.865	.066	65.065 (54)	−.001	−.001	.001
Grade								
Configural	3335.821 (1353)**	.044	.879	.059				
Metric	3440.184 (1407)**	.044	.876	.064	102.231(54)**	.000	−.003	.005
Scalar	3540.543 (1461)**	.044	.874	.064	93.279 (54)**	.000	−.002	.000

Note: **$p < .001$; *$p < .05$

Table 4

Latent Mean Differences Across Gender, Age, and Grade Level in the Main Sample

	Agreeableness		Conscientiousness		Extraversion		Neuroticism		Openness	
	Δ*M* (*SE*)	Wald *z*-test	Δ*M* (*SE*)	Wald *z*-test	Δ*M* (*SE*)	Wald *z*-test	Δ*M* (*SE*)	Wald *z*-test	Δ*M* (*SE*)	Wald *z*-test
Gender										
Girls (vs. boys)	0.30 (0.06)	4.97**	0.09 (0.05)	1.73	−0.10 (0.05)	−1.94	0.21 (0.05)	4.45**	−0.25(0.05)	−4.82**
Age										
11-12 years (vs. 9-10 years)	−0.05 (0.07)	−0.69	−0.15 (0.07)	−2.26*	−0.09 (0.07)	−1.29	0.09 (0.07)	1.26	0.06 (0.07)	−0.83
13-14 years (vs. 9-10 years)	0.07 (0.10)	0.76	−0.06 (0.09)	−0.69	− 0.07(0.09)	−0.77	−0.04 (0.09)	−0.45	0.00 (0.09)	0.00
13-14 years (vs 11-12 years)	0.12 (0.08)	1.49	0.08 (0.08)	1.09	0.02 (0.08)	0.23	-0.12 (0.08)	−1.63	0.06 (0.08)	0.75
Grade level										
Grade 5 (vs. Grade 4)	−0.14 (0.06)	−2.33*	−0.25 (0.06)	−4.38**	−0.16 (0.06)	−2.78**	0.06 (0.06)	1.14	−0.12(0.06)	−1.99*
Grade 5 (vs. Grade 6)	−0.29 (0.06)	−4.58**	−0.14 (0.06)	−2.51*	−0.07 (0.06)	−1.09	0.03 (0.06)	0.55	−0.13 (0.06)	−2.15*
Grade 6 (vs. Grade 4)	0.16 (0.07)	2.43*	−0.08 (0.06)	−1.44	−0.09 (0.06)	−1.43	0.03 (0.06)	0.49	0.02 (0.06)	0.26

Note: For each comparison, the reference category is added between brackets; ** p < .001; * p < .05.

As shown in Table 4, we did not find latent mean differences across age, except for Conscientiousness. Specifically, children aged 11-12 scored lower on Conscientiousness than children aged 9-10 ($\Delta M = -0.15$; $Z = -2.26$, $p < .05$). Regarding grade level, children in grade five reported lower levels of Agreeableness ($\Delta M = -0.14$, $Z = -2.33$, $p < .05$), Conscientiousness ($\Delta M = -0.25$, $Z = -4.38$, $p < .001$), Extraversion ($\Delta M = -0.16$, $Z = -2.78$, $p < .001$), and Openness ($\Delta M = -0.12$, $Z = -1.99$, $p < .05$) than children in grade four. Compared with children in grade six, children in grade five reported lower levels of Agreeableness ($\Delta M = -0.29$, $Z = -4.58$, $p < .001$), Conscientiousness ($\Delta M = -0.14$, $Z = -2.51$, $p < .05$) and Openness ($\Delta M = -0.13$, $Z = -2.15$, $p < .05$). In terms of the group differences between grade four and six, children in grade six scored higher on Agreeableness than children from grade four ($\Delta M = 0.16$, $Z = 2.41$, $p < .05$). No significant differences were found for Neuroticism in any group comparisons across grade level.

Discussion

In this study, we aimed to develop and validate a personality questionnaire among middle-childhood children (aged 9-14) in China. To achieve this, we first tested the factor structure of our questionnaire with a pilot sample. Subsequently, based on the results of the pilot study, the questionnaire was slightly revised by adding new items, and its psychometric properties were investigated using the main sample. Three specific research aims were addressed. First, we examined the factor structure and composite reliability of the developed questionnaire. Second, we assessed measurement invariance across gender, age, and grade level. Third, latent mean differences were compared across gender, age, and grade level. Although more research is needed, our study provides some preliminary evidence for the psychometric properties of this brief personality questionnaire in a Chinese context.

Factor Structure of the Personality Questionnaire

In this study, we found initial support for the psychometric quality of this newly developed personality measure. Both the pilot and main study demonstrated that a five-factor model provided the best fit to our data. Specifically, the five-factor model outperformed two alternative models, supporting its suitability over the one-factor model and confirming that the five traits adequately capture the underlying structure without requiring a higher-order general factor (cf., Musek, 2007). The correlations between the factors further indicated that the five factors were sufficiently independent from each other to warrant a five-factor solution. These factors corresponded with the

Big Five Model (i.e., agreeableness, conscientiousness, extraversion, neuroticism, openness; John, 2021). As such, our findings align with research among adults and adolescents (Ober et al., 2021; Soto & John, 2017) and suggest that the Big Five Model is also applicable to self-reports of children in middle childhood, providing further support for the universal value of this model.

In general, our final model displayed a relatively good model fit, except that the CFI value was close to but slightly below 0.90 (CFI = 0.878). Two reasons may account for this. First, CFA may be overly stringent for personality measures as it does not permit nontrivial cross-loadings (Barbaranelli et al., 2003; Morizot, 2014). This strictness may result in lower model fit indices, particularly in large samples where minor deviations may lead to a significant mismatch between the model and the data (Barbaranelli et al., 2003). Notably, the CFI value reported in our study is relatively comparable to those reported in prior research in middle childhood (e.g., Barbaranelli et al., 2003). Second, the broader conceptual scope of the questionnaire, developed based on the BFI framework, may contribute to the lower CFI. The BFI framework suggests a hierarchical personality structure with multiple facets contributing to each trait (Soto & John, 2017), potentially reducing item cohesion within each scale. Considering the above reasons, the final model still seems to have an acceptable fit.

Furthermore, standardized factor loadings confirmed that, overall, items effectively measured their intended constructs. In the final model of the main study, most items had factor loadings above the .40 threshold (Stevens, 2009), with only one exception. Specifically, the item “I am less active than the other children” had a factor loading of .38. This may be due to the addition of a residual covariance to account for its similarity with another item, i.e., “I am full of energy”. Despite this, the overall factor loadings exceeded our expectations, particularly in comparison with previous studies about personality measures for children (e.g., Barbaranelli et al., 2003; Charalampous & Kokkinos, 2014). Composite reliability for each trait was also acceptable to satisfactory, thereby supporting the reliability of this questionnaire.

It should be noted that the conceptualization of some traits (extraversion and openness in particular) may be influenced by the removal of certain items. With regard to extraversion, items related to “quietness” (e.g., “I am a quiet person”) or its opposite, “talkativeness” (e.g., “I am talkative”) had negative or low factor loadings, and were thus excluded. Aligning with cultural expectations of politeness in the Chinese context, quietness may be valued as a positive attribute of being respectful, attentive, and wellmannered (Li, 2003). Similarly, items related to “assertiveness” (e.g., “In a group, I like deciding what we are going to do”) were removed due to their negative or low factor loadings on extraversion. This may be due to the emphasis on academic performance in the Chinese culture, where children who have better academic performance are more likely to be assigned leadership roles in group activities (Zhu, 2021). Regarding openness, items related to aesthetic activities with low factor loadings were not retained (c.f., John, 2021). One possible explanation is that children’s engagement in these activities may be driven by parental expectations instead of their intrinsic interest, leading to the weaker associations between these items and openness (Kong, 2021). Despite these adjustments, the core aspects of extraversion and openness were preserved. Future research is recommended to refine our measure, for instance by incorporating culturally appropriate items for extraversion and openness.

Measurement Invariance

Our second aim was to examine measurement invariance of the developed questionnaire across different groups. Using our main dataset, we found evidence for scalar invariance across age and grade level. As such, these findings provide further support for the validity of our personality questionnaire. Regarding gender, partial scalar invariance was established, as the intercepts of two items for agreeableness, “I am rude to others” (reversed) and “I treat other people with respect”, needed to be freely estimated (Byrne et al., 1989). Girls rated themselves higher on these two items than boys. This may be caused by stronger social expectations for girls in the Chinese context to exhibit politeness and respect than boys. Therefore, the results with regard to gender differences in agreeableness need to be interpreted with somewhat more caution than the other results.

Overall, the questionnaire appeared to meet the criteria of (partial) scalar variance invariance, an essential prerequisite for conducting group mean comparisons (Dong & Dumas, 2020; Putnick & Bornstein, 2016). If children from different groups did not conceptualize these traits in similar ways, mean differences could be attributed to, for instance, item bias rather than true differences. This questionnaire, therefore, may provide a valuable tool for comparing children’s personality traits across gender, age, and grade levels.

Latent Mean Differences

Our third aim was to investigate latent group means. Results showed that girls reported higher levels of agreeableness and neuroticism, and a lower level of openness compared to boys. These findings align with prior research that gender-based personality differences may emerge already in middle childhood (Hendriks et al., 2008; Rossier et al., 2007). Notably, most studies exploring gender differences at this stage are based on Western samples (e.g., Hendriks et al., 2008; Rossier et al., 2007). By using a sample from an Eastern cultural context (i.e., China), our study demonstrates that gender differences could also be found in Chinese children. Therefore, it is important to account for gender when examining children's personality traits across cultural contexts.

Interestingly, unlike prior research (e.g., Hendriks et al., 2008; Rossier et al., 2007), we did not find significant gender differences for conscientiousness and extraversion. Such an inconsistency in findings may be explained by cultural expectations in the Chinese context. In Chinese society, qualities such as diligence, persistence, and active participation in groups (associated with conscientiousness and extraversion) are equally valued and encouraged in both boys and girls (Li, 2003). These shared expectations may minimize gender distinctions in these traits (Li, 2003).

Regarding age differences, our findings seem to align closely with prior meta-analytic research (Bleidorn et al., 2022), showing minimal mean differences in personality traits across age groups during middle childhood. Four out of five traits showed no significant differences, with only conscientiousness showing a slight decline with age. Future research could build on this by using longitudinal designs to track these subtle changes over several years, offering a more comprehensive understanding of personality development in childhood.

Last, grade-level differences were observed for four traits: Agreeableness, conscientiousness, extraversion, and openness, with children in grade five exhibiting lower levels of these traits and higher neuroticism compared to children from grades four and six. This is an interesting finding, as the role of grade level in personality traits has been less examined. Our findings seem to indicate that educational experiences may influence personality traits in middle childhood. In the Chinese context, the academic pressure associated with graduation exams in grade six could create a transitional period in grade five. As this grade bridges the early learning years and preparation for middle school, it may lack the clear structure and expectations of grades four and six, potentially contributing to the fluctuations in personality traits. Future research should further explore the role of grade level in personality development, particularly in contexts with high academic demands in middle childhood.

Limitations and Future Directions

Several limitations should be considered when interpreting the results of this study. First, it should be acknowledged that the data was collected in one specific region in China. As the region could be seen as quite representative in China in terms of economy, education, and culture, we would expect a good generalizability of the results to other parts of China. However, a more representative sample, drawn from various parts of China, would be preferred to replicate and strengthen our findings.

Second, in this study, we adopted an *etic* approach, which involves using universal constructs to develop the measure, ensuring consistency across different cultural contexts (Berry, 1989). However, to gain a more comprehensive understanding of personality traits within the Chinese context, an ideal approach would combine both *etic* and *emic* perspectives—where emic focuses on culturally specific constructs.

Conclusion

In sum, our findings provide preliminary evidence that this personality questionnaire can reliably and validly measure Chinese children's perceptions of their own personality in middle childhood. This is particularly important as personality traits tend to be predictive of various developmental outcomes (John, 2021; Poropat, 2009). Moreover, this self-reported questionnaire could provide researchers with access to information about children's internal states and thoughts, which are often challenging to capture through other informants such as parents or teachers (Barbaranelli et al., 2003; Hendriks et al., 2008). Additionally, the brevity of this questionnaire (32 items) makes it practical for use in both research and applied settings. It should be kept in mind, however, that this questionnaire was only tested in two Chinese samples in the same region of China. More research is needed to test its other psychometric qualities, such as predictive validity for meaningful outcomes (e.g., academic performance, developmental issues) and convergent validity with other personality measures.