Journal of Personality Assessment, *95*(1), 74–84, 2013

ISSN: 0022-3891 print / 1532-7752 online
DOI: 10.1080/00223891.2012.700466

Routledge
Taylor & Francis Group

Psychometric Properties of the Mini-IPIP in a Large, Nationally Representative Sample of Young Adults

RUTH E. BALDASARO,[1] MICHAEL J. SHANAHAN,[2] AND DANIEL J. BAUER[1]

[1]*Department of Psychology, University of North Carolina at Chapel Hill*
[2]*Department of Sociology, University of North Carolina at Chapel Hill*

Drawing on a large, nationally representative sample of young adults (the National Longitudinal Study of Adolescent Health; $N = 15{,}701$; M age $= 29.10$), we evaluated the psychometric properties of the Mini-IPIP, a 20-item inventory designed to concisely assess the 5 factors of personality. The results suggest that the Mini-IPIP has a 5-factor structure; most of the scales have acceptable reliability; all the scales have partial or full metric invariance; and the scales exhibit some degree of criterion validity. However, the absence of scalar invariance for many of the scales suggests caution when comparing personality scores among groups defined by sex or race and ethnicity. We offer practical considerations for researchers interested in using this inventory with this sample, and also suggestions for modification of the Mini-IPIP.

The Mini-IPIP, a 20-item short form based on the 50-item International Personality Item Pool five-factor model (IPIP–BF; Goldberg, 1999), was designed for circumstances in which the Big Five factors of personality need to be validly and reliably assessed with a small number of items (Donnellan, Oswald, Baird, & Lucas, 2006). On the one hand, there is a legitimate scientific interest in having a brief instrument with which to assess these constructs, reflecting their demonstrated utility beyond trait psychology. For example, the five factors have informed the study of diverse facets of health and well-being (Friedman, 2000; Lahey, 2009), including indicators of cardiovascular disease (T. W. Smith & MacKenzie, 2006), the progression of HIV-1 (Ironson & Hayward, 2008), risk-taking behaviors (Zuckerman & Kuhlman, 2000), antisocial behaviors (Roberts, Jackson, Burger, & Trautwein, 2009), and mortality (Kern & Friedman, 2008; Lahey, 2009). Yet patterns of morbidity, mortality, and health (broadly defined) are often studied in the framework of large, population-based data collection efforts, which are typically limited in the number of survey items that can be devoted to any one measure.

On the other hand, such an instrument must be long enough to obtain sound psychometric properties. The shortest Big Five instrument, the Ten-Item Personality Inventory (TIPI; see Gosling, Rentfrow, & Swann, 2003), yields scores that might not be adequately reliable. Moreover, the TIPI scales are moderately correlated (i.e., violating the assumption that facets of personality are orthogonal) and have been criticized for lack of construct breadth. Finally, the estimation of exploratory and latent factor models with TIPI is problematic with only two items per scale.

In addition to the IPIP–BF and TIPI, several other abbreviated personality scales have been developed, including the 60-item NEO Five-Factor Inventory (NEO–FFI; Costa & McCrae, 1992), the 44-item Big Five Inventory (BFI; John & Srivastava, 1999), the 40-item Mini-Markers (Saucier, 1994), and the FFM Rating Form (FFMRF; Mullins-Sweatt, Jamerson, Samuel, Olson, & Widiger, 2006). All of these scales have been shown to have convergent and divergent validity with full-length personality scales, acceptable reliability, and criterion validity (Costa & McCrae, 1992; John & Srivastava, 1999; Mullins-Sweatt et al., 2006; Saucier, 1994). However, the reliabilities of these abbreviated scales are typically lower than those of full personality scales (John & Srivastava; Mullins-Sweatt, 2006; Saucier, 1994). In addition, some evidence suggests that the five-factor structure for the NEO–FFI is not strongly supported (Egan, Deary, & Austin, 2000; Marsh et al., 2010; Rosellini & Brown, 2011). Overall, research on the psychometric properties of these inventories suggest that they can be used to assess the Big Five personality factors, but that using abbreviated scales involves accepting weaker psychometric properties. Indeed, one major problem with abbreviated scales is that they typically have weaker reliability and validity, which would be unacceptable for full-length forms (G. T. Smith, McCarthy, & Anderson, 2000).

Thus, in using any abbreviated personality scale, the researcher must balance two considerations: the desire to minimize the number of items needed to assess each personality domain, and the goal of maintaining content coverage and good psychometric properties. With these goals in mind, Donnellan and his colleagues (2006) designed the Mini-IPIP, which includes four items per domain, to provide a measure of the five factors that is both more concise than other abbreviated scales (e.g., NEO–FFI, IPIP–BF, BFI, Mini-Markers, and FFMRF) and a psychometric improvement on the shorter TIPI. Drawing on five samples of undergraduates and data that included all items from the IPIP, they selected 20 items from the 50-item IPIP–BF (based on exploratory factor models) and demonstrated that the resulting instrument produced scores with excellent content coverage, high test–retest correlations, and criterion validity (both short and long term).

Previous studies of the Mini-IPIP have found mixed support for a five-factor structure using exploratory factor analysis (EFA) and confirmatory factor analysis (CFA; Cooper, Smillie,

Received December 1, 2011; Revised April 12, 2012.

Address correspondence to Michael J. Shanahan, Department of Sociology, University of North Carolina at Chapel Hill, Hamilton Hall CB 3210, Chapel Hill, NC, 27599–3210; Email: mjshan@unc.edu

& Corr, 2010; Donnellan et al., 2006). Specifically, drawing on a sample of 296 college students, Donnellan et al. (2006) reported a five-factor CFA model with acceptable fit according to the root mean square error of approximation (RMSEA) fit index, but poor fit according to the comparative fit index (CFI) and chi-square test of exact fit. Despite the poor fit, the model generally had relatively high factor loadings (loadings > .6). In a larger college student sample ($N = 1{,}470$), Cooper et al. (2010) found similar model fit results with RMSEA and standardized root mean square residual (SRMR), indicating acceptable fit, but once again with a CFI indicating poor fit for a five-factor CFA model. Cooper et al. also fit a five-factor EFA model to the same data and found acceptable-to-good fit and a clear pattern of factor loadings (i.e., strong loadings on one factor and small cross-loadings).

In this article, we examine the reliability, factor structure, and criterion validity of the Mini-IPIP, drawing on the National Longitudinal Study of Adolescent Health (Add Health), a large, nationally representative sample of 15,701 young adults. EFAs and CFAs evaluated the hypothesized five-factor structure for the inventory. Single-factor CFAs then examined the unidimensionality of each scale and multiple-group CFAs tested measurement equivalence for the five scales. Measurement equivalence was examined across men and women and across blacks and whites. If demonstrated, such equivalence allows these groups to share the same scale and thus any group differences on scale scores could be attributed to personality trait differences and not measurement differences.

In addition to these factor structure analyses, we also examined reliability and criterion validity. Reliability was assessed using multiple reliability indexes and we compared the results to both previous reliability findings for the Mini-IPIP as well as other abbreviated personality measures. Because the Add Health study focuses on health-related constructs, we evaluated criterion validity using several mental and behavioral health variables that have been found in previous research to be associated with some of the Big Five personality domains.

Our analyses represent a step forward in several respects. First, to our knowledge, previous uses of the Mini-IPIP have relied exclusively on data collected from college students. Consistent with recent calls for more carefully specified samples of large size (Reynolds, 2010), the analyses presented here are based on a nationally representative sample. The broadly defined population that Add Health reflects thus allows for comparisons between its personality scale scores and those obtained with more narrowly defined samples. Its large size assures that the analyses are adequately powered.

Second, Donnellan et al. (2006) compared the 20-item Mini-IPIP with the 50-item IPIP–BF, raising the issue of whether the 20-item Mini-IPIP's reliability and validity might differ if not accompanied by the additional 30 items of the IPIP–BF. The smaller scale, when accompanied by the larger scale, might yield poorer reliability and validity, reflecting the respondent's limited patience and attention span with the longer form. This consideration, however, could be counteracted if the respondents need more items to "get in a suitable mindset" for responding to these types of queries. The data set used here includes only the 20-item Mini-IPIP and the results can thus be compared to those obtained by Donnellan et al. (2006). By evaluating the Mini-IPIP in a sample without the larger form items we also provide evidence for one of the methodological steps that G. T. Smith et al. (2000) recommended for assessing abbreviated scales, namely assessing the validity of the scale in the way it will typically be used.

Third, previous studies of the Mini-IPIP have examined internal consistency of the scales using Cronbach's coefficient alpha (Cooper et al., 2010; Donnellan et al., 2006). In addition to coefficient alpha, ρ, a reliability coefficient that is appropriate when scales are not tau equivalent (i.e., items do not load equally on the latent factor), will be used to examine scale reliability (Bollen, 1989; Raykov, 2001). Sijtsma (2009) demonstrated that Cronbach's coefficient alpha is not associated with the internal structure of a scale and provides no information regarding unidimensionality. Therefore, we follow the recommendation of Green and Yang (2009) and use CFA models to assess the structure of the scales for evidence of unidimensionality.

Fourth, no studies to date have examined measurement equivalence of the Mini-IPIP, although research questions often concern differences between groups defined by sex or race and ethnicity. Measurement equivalence (or invariance) can have several meanings, including equivalent factor loadings (metric or weak invariance), equivalent factor loadings and thresholds (scalar or strong invariance), or equivalent sample covariance matrices across groups (Vandenberg & Lance, 2000). A scale with metric invariance implies that the latent factor is measured in the same units across populations. Because the measurement units are the same, metric invariance allows researchers to examine relationships between personality scores and other variables using correlation or regression analysis. A scale with scalar invariance implies that the origin of the latent factor scale is the same across groups, which means that the scores can be compared in absolute levels across groups. Scalar invariance allows for the direct comparison of personality scores between groups (e.g., between males and females) because the scores within each group will be on the same scale. Therefore, any group differences are due to personality trait differences and not measurement differences (see Vandenberg & Lance, 2000, for more information on measurement invariance).

Both full and partial measurement invariance were examined for each scale for differences between males and females as well as between blacks and whites. Few studies have examined whether sex or race differences on personality scales are due to underlying trait differences or measurement differences (for some exceptions, see Mitchelson, Wicher, LeBreton, & Craig, 2009; Reise, Smith, & Furr, 2001; L. Smith, 2002). Yet, the credibility of group comparisons is compromised by a lack of measurement equivalence across demographic groups (e.g., Knight, Roosa, & Umana-Taylor, 2009). Because the Add Health study was designed to be a nationally representative sample, it is ideally suited to examine measurement equivalence for both sex and race.

Finally, we examine criterion validity in a large, representative sample that likely represents more diverse experiences than those observed in smaller, college-based samples. To date, one study suggests very good criterion-related validity of the Mini-IPIP (Donnellan et al., 2006). Data from Add Health allow us to examine several health-related criteria, including hostility, self-conception (mastery), perceived stress, alcohol abuse and dependency, and delinquency, that are thought to be associated with the Big Five personality domains. This study thus adds to the limited published data on the criterion validity of the Mini-IPIP.

METHOD

Sample

The Add Health study was based initially on a nationally representative sample of youth in Grades 7 through 12 in the United States. The National Quality Education Database provided the sampling frame with its list of all high schools in the United States ($N = 26{,}666$). From this frame, 80 schools were selected. The sample was stratified by region; suburban, urban, or rural setting; school type (public, private, parochial); ethnic mix; and size. Fifty-two of the 80 schools agreed to participate, and 28 replacements schools were selected based on the stratifying variables. Each of the 80 schools was paired with a middle school (based on its contribution to the high school student body). A total of 145 of the schools agreed to host a confidential in-school survey, which focused on adolescent health and friends. This first wave yielded 90,118 students from Grades 7 to 12 (in 1994).

From the school rosters, students were randomly selected for a 1.5-hour interview, conducted in the home. Approximately 200 students were recruited from schools in each school pair, regardless of size. This procedure resulted in a self-weighting sample. A total of 20,745 adolescents in Grades 7 through 12 (ages 11–19) were interviewed at home. This in-home wave of interviews with target child and parent was carried out between April and December 1995. This article draws on data collected in Wave IV. Of the eligible respondents who had participated in the first in-home interview, 92.5% were relocated and 80.3% were reinterviewed, resulting in 15,701 adult in-home interviews collected between January 2008 and February 2009. Survey data were collected using a 90-minute Computer Assisted Personal Interviewing/Computer Assisted Self Interviewing instrument. Complete Mini-IPIP data are available for all personality items from 15,471 respondents (about 98.5% of the entire sample). The 15,471 respondents with complete data on the Mini-IPIP items have a mean age of 29.10 ($SD = 1.75$); are 53.2% female; and 52% white, 22% black, and 26% other (Hispanic, Asian, Native American, other).

Materials

Mini-IPIP. A list of the Mini-IPIP items can be seen in Table 1. For each item, individuals were asked to respond to this question: How much do you agree with each statement about you as you generally are now, not as you wish to be in the future? Responses followed a 5-point Likert-type scale ranging from 1 (*strongly agree*) to 5 (*strongly disagree*) with a neutral midpoint at 3 (*neither agree nor disagree*). For the analyses in this study, item responses were coded as missing for any individuals who refused to respond or did not know the answer to an item. Table 1 contains descriptive statistics for the items. These statistics suggest that the items are univariate normal and without notable skewness or kurtosis (Curran, West, & Finch, 1996).

Perceived stress. Individuals completed the 4-item Cohen's Perceived Stress Scale using a 5-point scale (Cohen & Williamson, 1988; $M = 4.84$, $SD = 2.95$, $\alpha = .72$). Responses were summed to create a total perceived stress score. Previous research has found perceived stress to be negatively correlated with extraversion, conscientiousness, agreeableness, and openness, and positively correlated with neuroticism (Ebstrup, Eplov, Pisinger, & Jørgensen, 2011; Penley & Tomaka, 2002). We hypothesize that the Mini-IPIP scales have the same pattern of correlations with perceived stress.

Hostility. Individuals completed the 5-item, 5-point scale based on items found under the anger facet of the NEO PI–R ($M = 11.54$, $SD = 2.5$, $\alpha = .78$; Costa & McCrae, 1992). Items include "I get angry easily," "I am not easily bothered by things," "I rarely get irritated," "I lose my temper," and "I keep my cool." Responses were summed to create a total hostility score with higher scores indicating higher hostility. Previous research has found hostility to be negatively correlated with agreeableness and conscientiousness, and positively correlated with neuroticism (Donnellan et al., 2006). We hypothesize that in this sample the agreeableness and conscientiousness scales of the Mini-IPIP also correlate negatively with

TABLE 1.—Descriptive statistics for the Add Health Personality items associated with the Big Five personality domains.

Big Five Domain	Add Health Item	Text	*M*	*SD*	Skewness	Kurtosis
E	H4PE1	I am the life of the party	2.88	1.00	−.06	−.51
A	H4PE2	I sympathize with others' feelings.	1.81	.73	1.14	2.66
C	H4PE3	I get chores done right away	2.53	1.05	.31	−.75
N	H4PE4	I have frequent mood swings	3.28	1.10	−.34	−.76
I	H4PE5	I have a vivid imagination	2.36	.99	.47	−.40
E	H4PE9	I don't talk a lot	3.41	1.12	−.34	−.80
A	H4PE10	I'm not interested in other people's problems.	3.58	.95	−.70	.17
C	H4PE11	I often forget to put things back in their proper place	3.45	1.11	−.51	−.67
N	H4PE12	I am relaxed most of the time	2.38	.92	.78	.23
I	H4PE13	I am not interested in abstract ideas	3.36	.88	−.19	−.24
E	H4PE17	I talk to a lot of different people at parties	2.59	1.08	.44	−.76
A	H4PE18	I feel others' emotions	2.35	.85	.73	.44
C	H4PE19	I like order	2.10	.84	.81	.78
N	H4PE20	I get upset easily	3.42	.96	−.61	−.31
I	H4PE21	I have difficulty understanding abstract ideas	3.55	.83	−.54	.20
E	H4PE25	I keep in the background	3.28	.97	−.28	−.70
A	H4PE26	I am not really interested in others	3.82	.78	−.91	1.37
C	H4PE27	I make a mess of things	3.82	.82	−.90	1.03
N	H4PE28	I seldom feel blue	2.78	1.03	.31	−.89
I	H4PE29	I do not have a good imagination	3.94	.82	−.94	1.23

Note. $N = 15{,}548$–$15{,}676$. A = Agreeableness; C = Conscientiousness; E = Extraversion; N = Neuroticism; I = Intellect/Imagination.

hostility, whereas neuroticism would correlate positively with hostility.

Mastery. Individuals completed the 5-item mastery scale using a 5-point scale (Pearlin & Schooler, 1978; $M = 14.75$, $SD = 2.92$, $\alpha = .78$). Responses were summed to create a total mastery score with higher scores indicating higher mastery. Previous research has found environmental mastery to be positively correlated with agreeableness, extraversion, and conscientiousness, and negatively correlated with neuroticism (Schmutte & Ryff, 1997). We hypothesize that the Mini-IPIP scales have the same pattern of correlations with mastery.

Alcohol abuse. Individuals completed the five items based on the *Diagnostic and Statistical Manual of Mental* Disorders (4th ed. [*DSM–IV*]; American Psychiatric Association, 2000) alcohol abuse symptoms using a 3-point scale ($M = 5.83$, $SD = 2.12$, $\alpha = .70$). A total number of alcohol abuse symptoms was calculated.

Alcohol dependence. Individuals completed the seven items based on the *DSM–IV* alcohol dependence symptoms using a 2-point scale ($M = 1.89$, $SD = 1.62$, KR20 = .68). A total number of alcohol dependence symptoms was calculated. In previous research, extraversion was found to have a positive association with alcohol abuse and dependence symptoms, whereas agreeableness and conscientiousness were found to have negative associations with alcohol abuse and dependence symptoms (Flory, Lynam, Milich, Leukefeld, & Clayton, 2002). We hypothesize that the Mini-IPIP scales have the same pattern of correlations with alcohol abuse and dependence, although the correlations might be smaller because our sample includes fewer individuals with higher levels of substance use relative to the sample studied by Flory et al. (2002).

Deliquency. Individuals completed 12 items regarding delinquent behaviors using a 4-point scale ($M = 3.67$, $SD = 3.46$, $\alpha = .78$). Responses were summed to create a total delinquency score. In previous research extraversion was found to have a positive association with antisocial personality disorder symptoms, whereas agreeableness and conscientiousness were found to have negative associations with antisocial personality disorder symptoms (Ball, Tennen, Poling, Kranzler, & Rounsaville, 1997; Flory et al., 2002). We hypothesize that the Mini-IPIP scales have the same pattern of correlations with delinquency, although the size of the correlations might not be the same as in the previous studies, which focused on antisocial personality disorder symptoms, rather than the frequency of delinquent behaviors, which was used in this study.

Procedure

First, EFAs and CFAs were conducted with M*plus* (Version 6.0; Muthén & Muthén, 2010) to determine the appropriate factor structure for the items. Because the Big Five factors are hypothesized to have orthogonal factors (Costa & McCrae, 1995; Goldberg, 1993), an EFA was fit using varimax rotation extracting one to seven factors, followed by a five-factor CFA with uncorrelated factors. However, because previous studies have found factor analysis models with correlated factors to have better fit than uncorrelated factor models (e.g., Biesanz & West, 2004; Socha, Cooper, & McCord, 2010), we also performed an EFA using quartimin rotation (extracting one to seven factors) and fit a five-factor CFA model with correlated factors to determine if models with correlated factors would better fit the data.[1] The models were fit to the observed data using robust weighted least squares (WLSMV, weighted least squares with adjusted mean and variance) estimation. WLSMV was chosen because it performs well at this sample size (Flora & Curran, 2004). For the CFA models, the factor means and variances were constrained to be 0 and 1.00 to identify the model.

Second, based on the EFAs and CFAs, items from the inventory were assigned to a scale for one of the Big Five personality domains and the properties of the scales were examined. For each scale, reliability was assessed using Cronbach's alpha (Cronbach, 1951) and ρ (Bollen, 1989; Raykov, 2001). In addition, dimensionality was assessed using one-factor CFA models.

Third, for each scale measurement equivalence, models were fit using theta parameterization in M*plus* (Version 6.0; Muthén & Muthén, 2010). The scales were examined for measurement equivalence between male and female respondents as well as between black and white respondents for each personality domain. Following the advice of Millsap and Yun-Tein (2004), a baseline model with factor loadings, thresholds, and unique factor variances was freely estimated across groups, except the parameters necessary for model identification (for identification details see Appendix A in Millsap & Yun-Tein, 2004). This baseline model was fit separately for each personality scale. One baseline model was fit to compare male and female respondents, and another baseline model was fit to compare black and white respondents. All subsequent invariance models were compared to the appropriate baseline model to determine the level of invariance for each scale (e.g., the male–female full metric invariance model for conscientiousness was compared to the male–female baseline model for conscientiousness).

After the baseline model, a metric invariance model was fit with all factor loadings fixed to be equivalent across groups. A chi-square likelihood ratio test (LRT, $\Delta\chi^2$) was used to compare the full metric invariance model to the baseline model. Because of the large sample size, a p value of .01 was used as a cutoff criteria. If the LRT was not significant, the scale was determined to have metric invariance. If the LRT was significant, then the modification indexes for the metric invariance model were examined to select an item with a factor loading that could be freely estimated across groups to improve the model fit. Factor loadings were ordered based on modification index size and then the factor loading for the item with the largest modification index was freed one-by-one until a LRT showed no significant difference from the baseline model. Up to two factor loadings were freed. If, after freeing two factor loadings, the LRT showed a significant difference from the baseline model, the scale was considered to not have metric invariance. If a nonsignificant LRT was found, then testing for scalar invariance proceeded.

Based on the results from the metric invariance analysis, thresholds were constrained to be equal across groups for any items that were found to have metric invariance. An LRT was used to test whether the scalar invariance model with the additional constraints on the thresholds was significantly different from the baseline model. If the LRT was not significant, the scale

[1]There is disagreement regarding the hierarchical structure of personality factors (see Digman, 1997). However, the purpose of this study was not to determine whether personality has a hierarchical structure and therefore hierarchical personality models were not examined.

was determined to have full or partial scalar invariance. If the LRT was significant, then the modification indexes for the scalar invariance model were examined to select an item with thresholds that could be freely estimated across groups to improve the model fit. Items were selected to have freed thresholds by selecting the item with the largest sum of the modification indexes for that item's thresholds. Sets of item thresholds were then freed, one item threshold set at a time, until an LRT showed no significant difference from the baseline model. Up to two items' thresholds were freed and if, after freeing the thresholds, the LRT showed a significant difference from the baseline model, the scale was considered to not have scalar invariance. If a nonsignificant LRT was found, then the models were considered to have partial scalar invariance.

In addition to examining scale invariance, the criterion validity of the five personality scales was evaluated by correlating the personality scales with psychopathology-related outcomes. The psychopathology-related outcomes included perceived stress, hostility, mastery, alcohol abuse symptoms, alcohol dependency symptoms, and delinquency.

RESULTS

Exploratory Factor Analysis

The rotated factor pattern and factor correlation results of a five-factor EFA with quartimin rotation using WLSMV are reported in Table 2. The first eight eigenvalues were 4.08, 2.14, 1.99, 1.76, 1.43, 1.22, .88, and .81, with the first six eigenvalues above the Kaiser–Guttman rule for meaningful eigenvalues. The Kaiser–Guttman rule is often used to put an upper bound on the number of factors, but it tends to be liberal. Examination of the factor loadings for the six-factor EFA reveal that the first five factors are nearly identical to those found in the five-factor EFA and the last factor only has two items with loadings of |.30| or greater. The two items that loaded on the last factor were "I am not interested in abstract ideas" (.537) and "I have difficulty understanding abstract ideas" (.485). This suggests that the sixth factor might be accounting for some local dependence among the items, instead of representing a distinct personality factor. Given the similarity between the five- and six-factor solutions, only the five-factor solution is presented. Because the orthogonal (varimax) and oblique (quartimin) rotations resulted in nearly the same results, only the quartimin results will be presented to allow for comparisons with previous studies of the Mini-IPIP (Cooper et al., 2010; Donnellan et al., 2006). The factor loadings show a clear pattern of items with moderate to high loadings on one factor. All items had loadings of |.30| or greater and at least two of the four items on each factor had loadings of |.60| or greater. The weakest factor loading (−.32, without reverse coding) corresponded to the item "I seldom feel blue," which is similar to the CFA results reported by Donnellan et al. (2006). There was only one item, "I have difficulty understanding abstract ideas" from the openness scale, that cross-loaded on the neuroticism factor (+.33, without reverse coding). Overall, these results provide support for a five-factor structure for the Mini-IPIP inventory.

Confirmatory Factor Analysis

Because the CFA models with correlated and uncorrelated factors had nearly identical results, only the correlated factor

TABLE 2.—Rotated factor pattern loadings and factor correlations from five-factor exploratory factor analysis of the Mini-IPIP using weighted least squares with adjusted mean and variance with quartimin rotation.

Big Five Domain	Add Health Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Conscientiousness	H4PE3	**.64**	.06	−.11	−.04	.10
	H4PE11[a]	**.65**	−.04	.06	−.02	−.09
	H4PE19	**.56**	−.03	.08	.04	.11
	H4PE27[a]	**.64**	−.02	.06	.02	−.20
Extraversion	H4PE1	.04	**.66**	−.17	.10	.09
	H4PE9[a]	−.06	**.55**	.28	−.06	−.01
	H4PE17	.00	**.68**	.02	.08	−.03
	H4PE25[a]	−.01	**.68**	.15	−.03	−.14
Agreeableness	H4PE2	.16	.03	**.54**	.10	.22
	H4PE10[a]	−.03	.00	**.78**	−.04	−.01
	H4PE18	.11	.05	**.53**	.15	.23
	H4PE26[a]	.05	.14	**.69**	.02	−.11
Intellect/ Imagination	H4PE5	−.03	.09	−.09	**.74**	.14
	H4PE13[a]	−.11	−.14	.28	**.49**	−.18
	H4PE21[a]	−.02	−.10	.16	**.49**	**−.33**
	H4PE29[a]	.08	.08	.01	**.71**	−.04
Neuroticism	H4PE4	−.02	−.04	.02	.04	**.74**
	H4PE12[a]	−.07	−.11	.19	−.09	**.44**
	H4PE20	−.08	−.03	−.01	−.04	**.70**
	H4PE28[a]	.00	−.08	.02	.01	**.32**
		Latent Variable Correlations				
Factor		1	2	3	4	5
1. Conscientiousness		1.00				
2. Extraversion		.14	1.00			
3. Agreeableness		.19	.24	1.00		
4. Intellect/Imagination		.08	.23	.28	1.00	
5. Neuroticism		−.13	−.07	−.17	−.12	1.00

Note. Factor loadings >.30 are shown in bold.
[a]Indicates items that were reverse scored.

results are presented in Table 3. Because of the large number of observations ($N = 15{,}685$), the chi-square significance test is sensitive to model misfit (Bentler & Bonett, 1980; Hu & Bentler, 1999) and therefore it is unsurprising that the null hypothesis of perfect fit was rejected, $\chi^2(160, N = 15{,}685) = 16456.19$, $p < .05$. Other fit indexes provide mixed support for the model. Some indicate poor fit (CFI = .87; Tucker–Lewis Index [TLI] = .85), whereas other suggest adequate fit (RMSEA = .08). These model fit indexes are similar to those found in previous studies (Cooper et al., 2010; Donnellan et al., 2006). Because the EFA results suggested some local dependence, modification indexes from the CFA model were examined to see if allowing correlations among items would improve model fit. The modification indexes were very large (> 1,000) for several item correlations. These modification indexes suggest that a five-factor model might be appropriate for the Mini-IPIP, but that there could be local dependence among some of the items.

Mini-IPIP Scale Analysis

In addition to examining the factor structure of the full scale, the psychometric properties of the scales were also examined. Reliability was found to be acceptable according to Cronbach's alpha and ρ for each scale (Conscientiousness, $\alpha = .65$, $\rho = .72$; Extraversion, $\alpha = .71$, $\rho = .78$; Agreeableness, $\alpha = .70$, $\rho = .78$; Intellect/Imagination, $\alpha = .65$, $\rho = .75$; Neuroticism, $\alpha = .62$, $\rho = .68$). These statistics are surprisingly high given the number of items, but the Cronbach's alpha results are consistent with prior studies of the Mini-IPIP (Donnellan et al., 2006; Cooper et al., 2010).

Table 4 reports the model fit results for the one-factor CFA results for each personality scale. As with the five-factor CFA model, the large subsample sizes result in the chi-square test of fit consistently rejecting the null hypothesis of good fit. However, the other fit statistics in Table 4 provide evidence of good model fit for Conscientiousness, Extraversion, and Neuroticism scales according to the CFI (values > .95; Hu & Bentler, 1999). Conscientiousness and Neuroticism were found to have acceptable fit according to RMSEA (values < .08; Browne & Cudeck, 1993) and TLI (values > .95; Hu & Bentler). Based on these fit statistics, there is strong evidence that the Conscientiousness and Neuroticism scales are unidimensional and some evidence that the Extraversion scale is unidimensional. Agreeableness had acceptable fit according the CFI, but poor fit according to RMSEA and TLI. Therefore there is little support for the unidimensionality of the Agreeableness scale. Finally, all fit statistics for the Intellect/Imagination scale were poor, providing evidence that it might not be a unidimensional scale.

Measurement Equivalence

Table 5 reports the findings with respect to measurement equivalence between males and females. For both metric and scalar invariance, models that were fully invariant were fit first (i.e., all items invariant across groups), then partially invariant models were fit by examining the modification indexes from the fully invariant model. Given the large sample size, $p < .01$ was used as the cutoff value for the chi-square difference test. The results in Table 5 provide support for partial metric invariance. For the Conscientiousness scale, metric invariance was found for all the items except "I make a mess of things." For the Extraversion scale, metric invariance was found for all the items except "I talk to a lot of different people at parties." For the Agreeableness scale, metric invariance was found for all the items except "I feel others' emotions." For the Intellect/Imagination scale, metric invariance was found for all the items except "I have a vivid imagination" and "I do not have a good imagination." For the Neuroticism scale, metric invariance was found for all the items, except "I get upset easily."

Table 5 also shows no support for scalar invariance. This suggests that scores on these scales are not directly comparable for men and women. However, because evidence for metric invariance was found for all of the scales, scores on these scales can be used to examine relationships with other variables.

Table 6 reports results for measurement equivalence between blacks and whites. Both the Conscientiousness and Agreeableness scales were found to have full or partial metric

TABLE 3.—Mini-IPIP five-factor confirmatory factor analysis results for correlated factor model.

Big Five Domain	Add Health Item	C	E	A	O	N
Conscientiousness	H4PE3	.50				
	H4PE11[a]	.68				
	H4PE19	.53				
	H4PE27[a]	.78				
Extraversion	H4PE1		.51			
	H4PE9[a]		.66			
	H4PE17		.69			
	H4PE25[a]		.83			
Agreeableness	H4PE2			.61		
	H4PE10[a]			.68		
	H4PE18			.61		
	H4PE26[a]			.83		
Intellect/	H4PE5				.58	
Imagination	H4PE13[a]				.63	
	H4PE21[a]				.67	
	H4PE29[a]				.75	
Neuroticism	H4PE4					.69
	H4PE12[a]					.48
	H4PE20					.80
	H4PE28[a]					.35
		Latent Variable Correlations				
		1	2	3	4	5
1. Conscientiousness		1.00				
2. Extraversion		.17	1.00			
3. Agreeableness		.30	.44	1.00		
4. Intellect/Imagination		.15	.33	.44	1.00	
5. Neuroticism		−.24	−.18	−.15	−.28	1.00

Note. $N = 15{,}685$. Model fit results: $\chi^2(160) = 16456.12$, $p < .05$; comparative fit index (CFI) = .87; Tucker–Lewis Index (TLI) = .85; root mean square error of approximation (RMSEA) = .08. A = Agreeableness; C = Conscientiousness; E = Extraversion; N = Neuroticism; O = Openness to Experience.

[a]Indicates items that were reverse scored.

TABLE 4.—Confirmatory factor analysis results for a one-factor model for each personality scale.

Personality Domain	*N*	χ^2	*df*	*p*	CFI	TLI	RMSEA
Conscientiousness	15,657	145.19	2	<.05	.99	.98	.07
Extraversion	15,634	921.03	2	<.05	.97	.89	.17
Agreeableness	15,644	2322.07	2	<.05	.92	.77	.27
Intellect/Imagination	15,509	3370.67	2	<.05	.88	.62	.33
Neuroticism	15,652	201.40	2	<.05	.99	.97	.08

Note. CFI = comparative fit index; TLI = Tucker–Lewis Index; RMSEA = root mean square error of approximation.

TABLE 5.—Goodness-of-fit statistics related to tests for measurement equivalence across males and females.

Personality Domain	Model	χ^2	*df*	$\Delta\chi^2$	Δdf	CFI	TLI	RMSEA	*p*
Conscientiousness	Baseline	154.18	5	—	—	.992	.980	.062	—
	Full metric	148.84	8	22.72	3	.992	.988	.047	.000
	Partial metric 1	136.13	7	8.82	2	.993	.987	.049	.012
	Partial scalar 1	190.93	15	62.03	8	.990	.992	.039	.000
	Partial scalar 2	157.19	12	25.62	5	.992	.992	.039	.000
Extraversion	Baseline	843.49	5	—	—	.969	.925	.146	—
	Full metric	738.40	8	33.66	3	.973	.959	.108	.000
	Partial metric 1	722.37	7	6.98	2	.973	.955	.114	.030
	Partial scalar 1	2241.00	15	1533.21	8	.918	.934	.138	.000
	Partial scalar 2	1304.05	12	593.68	5	.952	.952	.117	.000
Agreeableness	Baseline	2049.83	5	—	—	.922	.812	.229	—
	Full metric	1733.45	8	15.81	3	.934	.901	.166	.001
	Partial metric 1	1793.62	7	2.86	2	.932	.883	.181	.240
	Partial scalar 1	2059.11	15	284.87	8	.922	.938	.132	.000
	Partial scalar 2	1896.01	12	60.07	5	.928	.928	.142	.000
Intellect/Imagination	Baseline	3299.78	5	—	—	.874	.698	.292	—
	Full metric	2875.55	8	44.52	3	.891	.836	.215	.000
	Partial metric 1	3059.32	7	28.59	2	.883	.800	.237	.000
	Partial metric 2	3190.39	6	1.32	1	.878	.757	.262	.251
	Partial scalar 2	3345.18	11	47.37	5	.873	.861	.198	.000
Neuroticism	Baseline	233.80	5	—	—	.986	.966	.076	—
	Full metric	209.13	8	22.54	3	.987	.981	.057	.000
	Partial metric 1	192.53	7	7.06	2	.988	.980	.058	.029
	Partial scalar 1	341.55	15	159.20	8	.980	.984	.053	.000
	Partial scalar 2	208.93	12	30.73	5	.988	.988	.046	.000

Note. Metric invariance refers to setting item factor loadings to be equal across groups. Scalar invariance refers to setting both the item factor loadings and thresholds to be equal across groups. Partial metric 1 (2) or Partial scalar 1 (2) means that 1 (or 2) of the items were not invariant and allowed to have their factor loading or their thresholds freely estimated for each group. CFI = comparative fit index; TLI = Tucker–Lewis Index; RMSEA = root mean square error of approximation.

invariance, and partial scalar invariance. The Conscientiousness scale was found to have full metric invariance, and the only items that were not scalar invariant were "I get chores done right away" and "I make a mess of things." The Agreeableness scale was found to have partial metric and scalar invariance, and the only items that were not invariant were "I am not interested in other people's problems" and "I am not really interested in others."

TABLE 6.—Goodness-of-fit statistics related to tests for measurement equivalence across blacks and whites.

Personality Domain	Model	χ^2	*df*	$\Delta\chi^2$	Δdf	CFI	TLI	RMSEA	*p*
Conscientiousness	Baseline	117.29	5	—	—	.992	.981	.062	—
	Full metric	106.39	8	12.18	3	.993	.990	.046	.007
	Partial metric 1	101.80	7	2.79	2	.993	.988	.048	.247
	Partial scalar 1	215.56	15	117.01	8	.986	.989	.048	.000
	Partial scalar 2	110.13	12	11.58	5	.993	.993	.037	.041
Extraversion	Baseline	920.14	5	—	—	.960	.903	.177	—
	Full metric	752.43	8	51.45	3	.967	.951	.126	.000
	Partial metric 1	764.44	7	22.23	2	.967	.943	.136	.000
	Partial metric 2	807.68	6	1.31	1	.965	.929	.151	.253
	Partial scalar 2	826.05	11	49.54	5	.964	.961	.112	.000
Agreeableness	Baseline	1404.69	5	—	—	.943	.863	.218	—
	Full metric	1236.16	8	50.55	3	.950	.925	.162	.000
	Partial metric 1	1301.68	7	29.63	2	.947	.910	.177	.000
	Partial metric 2	1318.33	6	5.89	1	.947	.893	.193	.015
	Partial scalar 2	1381.87	11	9.95	5	.944	.939	.146	.077
Intellect/Imagination	Baseline	2732.36	5	—	—	.885	.725	.306	—
	Full metric	2367.51	8	1.88	3	.901	.851	.225	.597
	Full scalar	2684.96	19	194.83	11	.888	.929	.155	.000
	Partial scalar 1	2587.98	16	69.40	8	.892	.919	.166	.000
	Partial scalar 2	2440.38	13	17.91	5	.898	.906	.179	.003
Neuroticism	Baseline	198.36	5	—	—	.987	.968	.081	—
	Full metric	570.32	8	321.43	3	.962	.943	.109	.000
	Partial Metric 1	183.57	7	7.57	2	.988	.979	.066	.023
	Partial Scalar 1	771.12	15	562.10	8	.948	.959	.093	.000
	Partial scalar 2	309.66	12	128.13	5	.980	.980	.065	.000

Note. Metric invariance refers to setting item factor loadings to be equal across groups. Scalar invariance refers to setting both the item factor loadings and thresholds to be equal across groups. Partial metric 1 (2) or Partial scalar 1 (2) means that 1 (or 2) of the items were not invariant and allowed to have their factor loading or their thresholds freely estimated for each group. CFI = comparative fit index; TLI = Tucker–Lewis Index; RMSEA = root mean square error of approximation.

The Extraversion, Intellect/Imagination, and Neuroticism scales were found to have either full or partial metric invariance, but not scalar invariance. The Intellect/Imagination scale was found to have full metric invariance. For the Extraversion scale, metric invariance was found for all the items, except "I am the life of the party" and "I talk to a lot of different people at parties." For the Neuroticism scale, metric invariance was found for all the items, except "I seldom feel blue." These results suggest that there is evidence for partial measurement equivalence for blacks and whites on the Conscientiousness and Agreeableness scales, which means that scores on these scales are directly comparable for blacks and whites. Extraversion, Intellect/Imagination, and Neuroticism scales were found to have either full or partial metric invariance, but no scalar invariance That is, scores on these scales are not directly comparable for blacks and whites; however, these scores can be used to examine relationships with other variables.

Overall, none of the scales had partial measurement equivalence for men and women or for blacks and whites. Thus, scores cannot be directly compared for the groups examined, but all the scales had at least partial metric invariance, which provides support for using scores on these scales to examine relationships with other variables.

Criterion Validity

Table 7 reports the correlations between the personality scale sum scores and the criterion variables. Given the large sample sizes, this study had high power to detect significant correlations. All of the personality scales were significantly correlated with perceived stress, hostility, and mastery ($p < .001$). The pattern of correlations with perceived stress, hostility, and mastery are similar to those reported by previous studies, although the magnitude of the correlations are typically lower in this case (Donnellan et al., 2006; Penley & Tomaka, 2002; Schmutte & Ryff, 1997). The largest correlations were between Neuroticism and hostility ($r = .69$), Neuroticism and perceived stress ($r = .46$), Neuroticism and mastery ($r = -.32$), Conscientiousness and mastery ($r = -.27$), and Intellect/Imagination and mastery ($r = .26$). There were also significant correlations between Conscientiousness and both alcohol abuse and alcohol dependence symptoms ($r = -.09$ and $r = -.11$). Agreeableness and Intellect/Imagination were correlated with delinquency ($r = -.09$ and $r = .10$). Neuroticism was significantly correlated with alcohol abuse ($r = .10$), alcohol dependence ($r = .16$), and delinquency ($r = .10$). Once again, this pattern of results is similar to the pattern observed in previous studies, although the magnitude of the correlations is often lower in this study (Ball et al., 1997; Flory et al., 2002).

Discussion

This article examines the psychometric properties of the Mini-IPIP in a nationally representative sample. Findings supported a five-factor structure based on both EFA and CFA. Similar to previous studies, EFA results revealed items loading on one primary factor and any cross-loadings were small (Cooper et al., 2010). The CFA results suggest poor to acceptable fit for a five-factor model, but high factor loadings were found for most items, with the exception of "I seldom feel blue," which also had a problematic loading in the Donnellan et al. (2006) study. Overall, the CFA results had similar model fit to those found in previous studies on the Mini-IPIP (Cooper et al., 2010; Donnellan et al., 2006). Relative to other personality scales, the EFA and CFA results for the Mini-IPIP show fewer cross-loadings $> |.30|$ than the NEO–FFI (Egan et al., 2000; Rosellini & Brown, 2011), more orthogonal personality domains relative to the BFI and the NEO–FFI (Gosling et al., 2003; Rosellini & Brown), and better CFA model fit (Hopwood & Donnellan, 2010; Marsh et al., 2010).

It is often argued that a CFA model is too restrictive for personality scales, and that model misfit could be a result of failing to specify item cross-loadings, item residual correlations, or minor factors (Marsh et al., 2010; Hopwood & Donnellan, 2010). Hopwood and Donnellan (2010) found that CFI values for CFA models that were fit to personality scales typically range from .61 to .79, TLI values range from .52 to .70, and RMSEA values range from .09 to .13. These values are lower than those reported here with the Mini-IPIP, suggesting that perhaps the Mini-IPIP does not have as many item cross-loadings, or item residual correlations, as the longer personality scales examined by Hopwood and Donnellan. The absence of these cross-loadings is somewhat supported by the EFA findings for the Mini-IPIP, which found only one cross-loading above .30. We thus suggest that the EFA and CFA results support the hypothesized five-factor structure of the Mini-IPIP inventory.

In addition to evaluating the overall inventory, the psychometric properties of the scales were also examined. One challenge for abbreviated scales is the need to balance measuring the full content of the scale and selecting items that appear to be from a unidimensional scale. The more narrow the content of the items, the more reliable and unidimensional a scale will appear. The reliability of the Mini-IPIP scales was found to be acceptable with both Cronbach's alpha and ρ. The coefficient alphas for each of the Mini-IPIP scales were found to be similar to those

TABLE 7.—Correlations between Mini-IPIP scales and psychopathology-related variables.

	Criteria					
Predictor	Perceived Stress	Hostility	Mastery	Alcohol Abuse	Alcohol Dependence	Delinquency
Conscientiousness	−.21***	−.16***	.27***	−.09***	−.11***	−.06
Extraversion	−.14***	−.09***	.20***	.03*	.03*	.00
Agreeableness	−.05***	−.12***	.24***	−.02	−.03*	−.09*
Intellect/Imagination	−.10***	−.13***	.26***	.04***	.03*	.10**
Neuroticism	.46***	.69***	−.32***	.10***	.16***	.10**
N	15,457	15,462	15,459	6,304	6,307	792

*$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 8.—Summary of Mini-IPIP scale results.

Scale	Confirmatory Factor Analysis	Measurement Invariance			
		Gender		Race	
		Assess Relationships With Other Variables?	Compare Group Scores?	Assess Relationships With Other Variables?	Compare Group Scores?
Conscientiousness	Unidimensional	Yes, partial metric	No[a]	Yes, partial metric	Yes, partial scalar
Extraversion	May be unidimensional	Yes, partial metric	No[a]	Yes, partial metric	No
Agreeableness	May be unidimensional	Yes, partial metric	No[a]	Yes, partial metric	Yes, partial scalar
Intellect/Imagination	Not unidimensional	Yes, partial metric	No[a]	Yes, full metric	No[a]
Neuroticism	Unidimensional	Yes, partial metric	No[a]	Yes, partial metric	No[a]

[a]Cheung and Rensvold (2002) ΔCFI criteria would support partial scalar invariance for these scales.

previously reported (Cooper et al., 2010; Donnellan et al., 2006). The results were similar to the reliability reported for the 30-item FFMRF (Mullins-Sweatt et al., 2006), but lower than the reliability found for slightly longer scales including the 40-item Mini-Markers (Saucier, 1994), 44-item BFI (John & Srivastava, 1999), or 60-item NEO–FFI (Costa & McCrae, 1992). These results suggest that part of the reason the reliability might be lower for the Mini-IPIP scales could reflect the difficulty of trying to measure broad content with only four items.

To examine the structure of the scales for unidimensionality, one-factor CFA models were fit to each of the scales. Model fit was good according to multiple indexes for the Conscientiousness and Neuroticism scales. Some evidence for acceptable model fit was found for the Extraversion and Agreeableness scales. The results support the unidimensional structure for Conscientiousness and Neuroticism scales, and weaker evidence for the unidimensionality of the Extraversion and Agreeableness scales. As previously noted, there is a trade-off between unidimensionality and broad content coverage. In this case, it might be that the weaker evidence for the unidimensionality of the Extraversion and Agreeableness scales reflects the items selected for these scales. For Extraversion, the items assess talkativeness or a preference for being the center of attention versus in the background. For Agreeableness, the items are about interest in others or concern for others' emotions. It is possible that the similarity of the item content produces some local dependence (correlated items) for these scales, which results in poorer model fit as assessed by the overall fit indexes.

Poor model fit was found for the Intellect/Imagination scale. The poor fit of the one-factor CFA model for Intellect/Imagination suggests that the Intellect/Imagination items might not be measuring a unidimensional construct. Intellect/Imagination has been the most controversial of the five personality domains with some arguing that Intellect/Imagination is an intellect or a culture factor (Goldberg, 1993). Intellect/Imagination also typically has the lowest reliability for abbreviated personality scales (Costa & McCrae, 1992; John & Srivastava, 1999; Mullins-Sweatt et al., 2006; Saucier, 1994). It has been suggested that selecting items for Intellect/Imagination is more difficult because the domain is not well defined and it might be harder for participants to answer because they are less clear than items for other domains (Mullins-Sweatt et al., 2006). In addition to these problems with defining and clarifying the domain, it is possible that both the lower reliability and the poor CFA model fit found for the Intellect/Imagination scale of the Mini-IPIP reflects the focus of the content on two facets, imagination and abstract ideas.

The results for the measurement invariance analyses indicate that all of the scales, except for Conscientiousness and Agreeableness when comparing blacks and whites, were not scalar invariant and, in turn, that scores for these groups cannot be directly compared. The measurement invariance results found some evidence of either partial or full metric invariance for all of the scales, which provides support for using scores on these scales to examine correlational relationships with other variables. Given that many of the scales only had partial metric invariance, we would suggest that researchers use scoring techniques that allow for differential item functioning, either using factor scores or item response theory (IRT) approaches.

Given the complexity of these results, Table 8 contains a summary of the results of the analyses for each scale. As seen in Table 8, researchers can be confident that the Conscientiousness and Neuroticism scales are each primarily measuring one construct. Regarding the invariance results, given the large sample size, it is possible that the chi-square difference tests are overpowered, which might result in the rejection of models that are not meaningfully different. Given this concern, the researcher could consider using an alternative method for evaluating invariance. Cheung and Rensvold (2002) recommended examining the change in CFI (ΔCFI) for the scalar invariance models, using the cutoff criteria of $\Delta CFI \leq .01$, as an additional method to test for scalar invariance. Using this criteria, all of the scales, except Extraversion when comparing men and women, have partial scalar invariance. In addition to examining other model fit criteria, one can also examine the results of the measurement invariance models for each scale to determine if the factor loadings and item thresholds are meaningful or trivial for the aims of any given research project.[2] If the differences in factor loadings are not substantively meaningful, one can accept the metric invariance of the scale and examine relationships with other variables. Similarly, if the item threshold differences are not substantively meaningful, one can accept the scalar invariance of the scale and directly compare groups on the scale scores. One way to examine the substantive meaning of these differences would be to generate scores assuming invariance (e.g., sum scores) and scores allowing for some parameter differences across groups (e.g., factor scores or IRT scores). A scatterplot and correlation of the two types of scores could be examined for any substantive differences in the scoring methods. Alternatively, a sensitivity analysis could be performed using the two types of scores in

[2]Researchers can examine tables of the measurement invariance results for each scale in an online appendix provided by the authors.

parallel analyses and the results could be examined for any differences in inferences or parameter estimates that would result in meaningful differences in the substantive conclusions drawn from the analyses.

The results examining the criterion validity of the Mini-IPIP showed a similar pattern of relationships between the Mini-IPIP scales and the health-related criteria. However, the results typically included correlations that were lower in magnitude than those found in previous studies. This pattern of correlations that are lower in magnitude might be the result of using a personality scale with lower reliability than the ones used in previous studies. As with many abbreviated scales, the reliability for the Mini-IPIP scales is lower than what is typically reported for longer personality scales. Alternatively, because the items selected for the Mini-IPIP were chosen to have the largest loading on the appropriate personality domain and lowest absolute average loading on all the other domains, the items on the Mini-IPIP might represent a narrower construct than the original scale. As discussed previously, several of the scales seem to focus on two facets of a given domain. This narrower construct could also explain the lower magnitude of the correlations between the Mini-IPIP scales and the health-related criteria.

In addition to encouraging researchers to examine what would be meaningful measurement differences, we also have several recommendations for future research. First, one limitation of this study is that the Mini-IPIP could not be compared to other personality scales. Ideally, we would examine validity by comparing the long and short forms of the IPIP inventory. Unfortunately, the long form was not given to the individuals in the Add Health study. Counterbalancing this weakness, however, this study focused on evaluating scale invariance and reliability using observations that are nationally representative, whereas previous studies have included the long form, but only examined the Mini-IPIP with university students. Second, this study examines some evidence of criterion validity by examining correlations between the scales and some variables that should be correlated with the scales, but more research on the criterion validity of the Mini-IPIP is needed. Finally, we recommend that future versions of the Mini-IPIP consider testing different items for the scales to improve their measurement invariance and, specifically, that the Intellect/Imagination scale should be revised to create a unidimensional scale.

ACKNOWLEDGMENTS

This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health Web site (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis. This article was supported by 1R21HD050261-01A2 from NICHD (Shanahan and Bauer).

REFERENCES

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Ball, S. A., Tennen, H., Poling, J. C., Kranzler, H. R., & Rounsaville, B. J. (1997). Personality, temperament, and character dimensions and the *DSM–IV* personality disorders in substance abusers. *Journal of Abnormal Psychology*, *106*, 545–553. doi:10.1037/0021-843X.106.4.545

Bentler, P., & Bonett, D. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606. doi:10.1037/0033-2909.88.3.588.

Biesanz, J., & West, S. (2004). Towards understanding assessments of the Big Five: Multitrait–multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer. *Journal of Personality*, *72*, 845–876. doi:10.1111/j.0022-3506.2004.00282.x

Bollen, K. A. (1989). *Structural equation models with latent variables*. New York, NY: Wiley.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255.

Cohen, S., & Williamson, G. 1988. Perceived stress in a probability sample of the United States. In S. Spacapan & S. Oskamp (Eds.), *The social psychology of health* (pp. 31–67). Newbury Park, CA: Sage.

Cooper, A., Smillie, L., & Corr, P. (2010). A confirmatory factor analysis of the Mini-IPIP five-factor model personality scale. *Personality & Individual Differences*, *48*, 688–691. doi:10.1016/j.paid.2010.01.004.

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

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. doi: 10.1007/BF02310555

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16–29. doi:10.1037/1082-989X.1.1.16

Digman, J. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, *73*, 1246–1256. doi:10.1037/0022-3514.73.6.1246

Donnellan, M., Oswald, F., Baird, B., & Lucas, R. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*, 192–203. doi:10.1037/1040-3590.18.2.192

Ebstrup, J. F., Eplov, L. F., Pisinger, C., & Jørgensen, T. (2011). Association between the Five Factor personality traits and perceived stress: Is the effect mediated by general self-efficacy? *Anxiety, Stress & Coping*, *24*, 407–419. doi: 10.1080/10615806.2010.540012

Egan, V., Deary, I., & Austin, E. (2000). The NEO–FFI: Emerging British norms and an item-level analysis suggest N, A and C are more reliable than O and E. *Personality and Individual Differences*, *29*, 907–920. doi: 10.1016/S0191-8869(99)00242-1.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491. doi:10.1037/1082-989X.9.4.466

Flory, K., Lynam, D., Milich, R., Leukefeld, C., & Clayton, R. (2002). The relations among personality, symptoms of alcohol and marijuana abuse, and symptoms of comorbid psychopathology: Results from a community sample. *Experimental and Clinical Psychopharmacology*, *10*, 425–434. doi:10.1037/1064-1297.10.4.425

Friedman, H. (2000). Long-term relations of personality and health: Dynamism, mechanisms, tropisms. *Journal of Personality*, *68*, 1089–1107. doi:10.1111/1467-6494.00127

Goldberg, L. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*(1), 26–34. doi:10.1037/0003-066X.48.1.26

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In

I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

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

Green, S. A., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, *71*(1), 121–135. doi:10.1007/s11336-008-9098-4

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*, 332–346. doi: 10.1177/1088868310361240

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. doi:10.1080/10705519909540118

Ironson, G., & Hayward, H. (2008). Do positive psychosocial factors predict disease progression in HIV-1? A review of the evidence. *Psychosomatic Medicine*, *70*, 546–554. doi:10.1097/PSY.0b013e318177216c

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

Kern, M., & Friedman, H. (2008). Do conscientious individuals live longer? A quantitative review. *Health Psychology*, *27*, 505–512. doi:10.1037/0278-6133.27.5.505

Knight, G. P., Roosa, M. W., & Umana-Taylor, A. J. (2009). *Studying ethnic minority and economically disadvantaged populations: Methodological challenges and best practices*. Washington, DC: American Psychological Association.

Lahey, B. (2009). Public health significance of neuroticism. *American Psychologist*, *64*, 241–256. doi:10.1037/a0015309

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

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*, 479–515. doi: 10.1207/S15327906MBR3903_4

Mitchelson, J., Wicher, E., LeBreton, J., & Craig, S. (2009). Gender and ethnicity differences on the Abridged Big Five Circumplex (AB5C) of personality traits: A differential item functioning analysis. *Educational and Psychological Measurement*, *69*, 613–635. doi:10.1177/0013164408323235

Mullins-Sweatt, S. N., Jamerson, J. E., Samuel, D. B., Olson, D. R., & Widiger, T. A. (2006). Psychometric properties of an abbreviated instrument of the Five-factor model. *Assessment*, *13*, 119–137. doi: 10.1177/1073191106286748

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Pearlin, L. I., & Schooler, C. (1978). The structure of coping. *Journal of Health and Social Behavior*, *19*(1), 2–21

Penley, J. A., & Tomaka, J. (2002). Associations among the Big Five, emotional responses, and coping with acute stress. *Personality and Individual Differences*, *32*, 1215–1228. doi:10.1016/S0191-8869(01)00087-3

Raykov, T. (2001). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*, 69–76. doi:10.1177/01466216010251005

Reise, S., Smith, L., & Furr, R. (2001). Invariance on the NEO PI–R neuroticism scale. *Multivariate Behavioral Research*, *36*(1), 83–110. doi:10.1207/S15327906MBR3601_04

Reynolds, C. (2010). Measurement and assessment: An editorial view. *Psychological Assessment*, *22*(1), 1–4. doi:10.1037/a0018811.

Roberts, B., Jackson, J., Burger, J., & Trautwein, U. (2009). Conscientiousness and externalizing psychopathology: Overlap, developmental patterns, and etiology of two related constructs. *Development and Psychopathology*, *21*, 871–888. doi:10.1017/S0954579409000479

Rosellini, A. J., & Brown, T. A. (2011). The NEO Five-Factor Inventory: Latent structure and relationships with dimensions of anxiety and depressive disorders in a large clinical sample. *Assessment*, *18*, 27–38.

Saucier, G. (1994). Mini-Markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment*, *63*, 506.

Schmutte, P. S., & Ryff, C. D. (1997). Personality and well-being: Reexamining methods and meanings. *Journal of Personality and Social Psychology*, *73*, 549–559. doi:10.1037/0022-3514.73.3.549

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *71*(1), 107–120. doi:10.1007/s11336-008-9101-0

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

Smith, L. (2002). On the usefulness of item bias analysis to personality psychology. *Personality and Social Psychology Bulletin*, *28*, 754–763. doi:10.1177/0146167202289005

Smith, T. W., & MacKenzie, J. (2006). Personality and risk of physical illness. *Annual Review of Clinical Psychology*, 2435–2467. doi:10.1146/annurev.clinpsy.2.022305.095257

Socha, A., Cooper, C., & McCord, D. (2010). Confirmatory factor analysis of the M5–50: An implementation of the International Personality Item Pool item set. *Psychological Assessment*, *22*(1), 43–49. doi:10.1037/a0017371

Vandenberg, R., & Lance, C. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–69. doi:10.1177/109442810031002

Zuckerman, M., & Kuhlman, D. (2000). Personality and risk-taking: Common biosocial factors. *Journal of Personality*, *68*, 999–1029. doi:10.1111/1467-6494.00124