Article

Assessment
2015, Vol. 22(1) 65–75

DOI: 10.1177/1073191114536772
asm.sagepub.com
SAGE

# Factorial Invariance of the Five-Factor Model Rating Form Across Gender

**Douglas B. Samuel[1], Susan C. South[1], and Sarah A. Griffin[1]**

## Abstract

The Five-Factor Model Rating Form (FFMRF) provides a brief, one-page assessment of the Five-Factor Model. An important and unique aspect of the FFMRF is that it is the only brief measure that includes scales for the 30 facets proposed by Costa and McCrae. The current study builds on existing validity support for the FFMRF by evaluating its factorial invariance across gender within a sample of 699 undergraduate students. Consistent with other measures of the Five-Factor Model, men scored lower than women on the domains of neuroticism, extraversion, agreeableness, and conscientiousness but slightly higher on openness. The novel contribution of the current study is the use of exploratory structural equation modeling to determine that the FFMRF displayed a five-factor structure that demonstrated strong measurement invariance across gender. This factorial invariance adds important support for the validity of the FFMRF as a self-report measure as it indicates that the scores assess the same latent constructs in men and women. Although future work is needed to clarify some facet-level findings and evaluate for potential predictive biases, the present results add to the increasing body of research supporting the validity of the FFMRF as a self-report measure of personality.

## Keywords

Five-Factor Model, assessment, exploratory structural equation modeling, Five-Factor Model Rating Form, gender

Over the past two decades, an increasing consensus has emerged that personality functioning can be well described using an integrative model consisting of five broad dimensions (John, Naumann, & Soto, 2008). These five dimensions are conceptualized as bipolar and have been labeled using descriptors of the high or low pole, including extraversion versus introversion, agreeableness versus antagonism, conscientiousness versus undependability, neuroticism versus emotional stability, and openness versus closedness to experience. Although the Five-Factor Model (FFM) has been criticized (Block, 1995) and alternative models exist (Ashton, Lee, & Goldberg, 2004), the FFM has provided a framework that has helped to integrate findings from a variety of subdisciplines such as human development (Caspi, Roberts, & Shiner, 2005), health psychology (Deary, Weiss, & Batty, 2010), industrial/organizational psychology (Barrick, Mount, & Judge, 2001), and psychiatry (Kotov, Gamez, Schmidt, & Watson, 2010; Samuel & Widiger, 2008). In addition, the FFM has validity support including universality across cultures (McCrae, Terracciano, & Pro, 2005), and test–retest stability correlations of approximately .60 over periods of several years (Ferguson, 2010). The FFM has also evidenced meaningful empirical associations with a number of important life outcomes (Ozer & Benet-Martinez, 2006).

Considering this evidence, it is perhaps not surprising that there are a wide variety of instruments available to provide an assessment of the FFM (De Raad & Perugini, 2002). Of these instruments, the NEO Personality Inventory–Revised (NEO PI-R; Costa & McCrae, 1992) is the most widely used for research and clinical purposes. A specific advantage of the NEO PI-R is that it further subdivides each of the FFM domains into six facets that provide a more specific assessment of lower order components. For example, the domain of conscientiousness includes the facets of order, competence, dutifulness, achievement striving, self-discipline, and deliberation. These specific 30 NEO PI-R facets are not without critics. Some have criticized the fact that they were not developed in the lexical tradition (DeYoung, Quilty, & Peterson, 2007) and may not fully cover the lexical space (Goldberg, 1999). Nonetheless, the NEO PI-R facets have proven useful as a metric for differentiating near-neighbor constructs, such as the personality disorders (Reynolds & Clark, 2001). The facets have also been helpful for understanding differences and similarities across instruments, such as for social and clinical operationalizations of narcissism (Miller & Campbell, 2008) and measures assessing obsessive-compulsive personality disorder (Samuel &

[1]Purdue University, West Lafayette, IN, USA

**Corresponding Author:**
Douglas B. Samuel, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907, USA.
Email: dbsamuel@purdue.edu

Widiger, 2010b). Furthermore, the facets have demonstrated incremental validity beyond the domains for predicting specific behavioral outcomes (Paunonen & Ashton, 2001).

Beyond the NEO PI-R, there are a number of other inventories, many of which are decidedly briefer than the 240-item NEO PI-R that requires 20 to 30 minutes to complete. This length, no doubt, contributes to its validity support (Costa & McCrae, 2010) and provides greater psychometric precision than instruments with fewer items. But there are also situations for which this length might be prohibitive (e.g., longer testing batteries) and abbreviated measures are sought by clinicians and researchers. In this regard the ideal length for a personality inventory, like any tool, varies depending on the job one seeks to perform. The existence of multiple measures of various lengths with well-understood psychometric properties is crucial so that appropriate selections can be made in each particular situation.

Among the available briefer measures are the Big Five Inventory (John, Donahue, & Kentle, 1991), a 44-item measure that is used extensively within social-personality psychology research; the Ten-Item Personality Inventory (Gosling, Rentfrow, & Swann, 2003); the 20-item mini-IPIP (Donnellan, Oswald, Baird, & Lucas, 2006); the 60-item NEO Five Factor Inventory (Costa & McCrae, 1992); and the Five-Factor Model Rating Form (FFMRF; Mullins-Sweatt, Jamerson, Samuel, Olson, & Widiger, 2006). Although other measures also have lower order facet scales (e.g., Soto & John, 2009), the FFMRF is uniquely situated as an abbreviated measure in that it assesses all 30 facets on the NEO PI-R. The FFMRF accomplishes this via a single item for each of those facets.

The FFMRF is a single-page form that consists of 30 items that allow individuals to record personality ratings on a 5-point scale where 1 is *extremely low*, 2 is *low*, 3 is *neither high nor low*, 4 is *high*, and 5 is *extremely high*. Each item corresponds to a specific facet and includes the facet label as well as 2 to 4 unique adjective descriptors for both the high and low poles. For example, the facet of assertiveness appears within the domain of extraversion and is described on the high pole by the adjectives "dominant, forceful" and on the low pole by "unassuming, quiet, resigned." The FFMRF was first developed by Lynam and Widiger (2001) as a brief method of collecting personality descriptions from busy mental health professionals and continues to be used for this purpose (e.g., Lowe & Widiger, 2009; Mullins-Sweatt & Widiger, 2009, 2011; Samuel & Widiger, 2004, 2006, 2009, 2010a, 2011). In this capacity it has displayed reliability across independent raters (Lynam & Widiger, 2001), convergence across methods (Samuel & Widiger, 2010a), and temporal stability (Samuel & Widiger, 2011).

An emerging literature has also utilized the FFMRF as a self-report measure of the FFM (e.g., Eisenlohr-Moul, Walsh, Charnigo, Lynam, & Baer, 2012; Follingstad, Rogers, & Duvall, 2012; Howell, Dopko, Turowski, & Buro, 2011; Kaiser, Milich, Lynam, & Charnigo, 2012; Schenk, Ragatz, & Fremouw, 2012; Schuettler & Boals, 2011; Schwartz, Fremouw, Schenk, & Ragatz, 2012; Thomas et al., 2013). Mullins-Sweatt et al. (2006) provided the first empirical investigation of the FFMRF's psychometric properties and found that the domains obtained reasonable internal consistency, with a median alpha value of .69 across five samples. The FFMRF domain scores also converged reasonably well with the NEO PI-R with median values ranging from .57 (openness) to .68 (extraversion). The convergent correlations of the FFMRF facets, consisting of a single item, with the NEO PI-R were understandably lower ranging from .29 to .63. However, each FFMRF item's mean convergence with the target NEO PI-R facet was always larger than the respective mean discriminant validity correlation. More recently, Samuel, Mullins-Sweatt, and Widiger (2013) extended this work by evaluating the structural validity of the FFMRF using exploratory structural equation modeling, which combines aspects of exploratory and confirmatory factor analytic approaches for testing the fit of a five-factor model. In doing so, they demonstrated that the five-factor solution obtained at least acceptable fit according to all indices employed (e.g., CFI = .94; RMSEA = .033) and the specific facet loadings corresponded well with the prescribed FFM framework. Furthermore, the five factors from this solution obtained reasonable convergent and discriminant validity with the NEO PI-R domains.

This initial evidence has provided support for the psychometric properties of the FFMRF; however, additional empirical evaluation would be helpful to support its continued use as a self-report measure. One particularly important area of investigation is the degree to which the structure of the FFMRF is invariant across gender. There may well be important differences in personality across gender in terms of mean-level, structure, or the quality of items on the FFMRF. To date, the FFMRF has not been examined for measurement invariance across gender. This is a particularly important test for any measure of the FFM, as Feingold (1994) demonstrated important personality differences across men and women, but there has been only limited investigation of the extent to which the measures of these constructs demonstrate equivalent factor structures between genders.

Notably, mean differences in raw scale scores indicate men were generally higher on assertiveness whereas women were higher on extraversion, anxiety, trust, and tender-mindedness. Similar differences have been reported for the NEO PI-R (Costa & McCrae, 1992). It is important to note that gender differences alone do not connote bias (Anderson, Sankis, & Widiger, 2001); rather, they might reflect true differences in mean scores. Thus, it is important to examine for measurement equivalence, or invariance, which would indicate that the scores on the measure have the same meanings for men and women. Furnham, Guenole, Levine, and

Chamorro-Premuzic (2013) recently investigated measurement invariance within the NEO PI-R and noted that there was equivalence, suggesting that the relations of facets with the latent personality dimensions were the same across gender. In the current study, we provide the first investigation of the FFMRF's measurement invariance across gender within a large undergraduate sample.

## Method

### Participants

Our sample was drawn from a mass prescreening of the undergraduate psychology participant pool. From this total possible sample of 803 individuals, we excluded participants with invalid administrations based on providing the same answer to every item on at least one of the questionnaires. Participants included in the final analyses consisted of 699 individuals, 342 women (*M* age = 19.4 years, *SD* = 2.0), and 357 men (*M* age = 19.8 years, *SD* = 1.7). The total sample was predominantly White (73%) but included a number who identified as Asian/Pacific Islanders (15%), Black (4%), Hispanic/Latino (3%), international (3%), or other (2%). Most were in their first (52%) or second year (26%) of college and represented a wide variety of major areas of study. An explanation and rationale for the study was provided to all participants, and written informed consent was obtained.

### Statistical Analyses

As a first step, confirmatory factor analysis (CFA) using Mplus 7 (Muthén & Muthén, 1998-2010) was conducted for the full sample and separately by gender. Each of the 30 FFMRF items was fit to a latent factor corresponding to the posited higher order domain (e.g., the six items measuring the neuroticism facets were fit to load on a Neuroticism factor). We followed the findings from Samuel et al. (2013) in choosing which indicator to set to 1.0 in order to set the factor metric. We also allowed 6 cross-loadings based on the findings of Samuel et al. (2013) as well as a broader literature indicating that certain facets routinely load across domains (Costa & McCrae, 2010). Specifically, we allowed impulsivity (N5) to load on extraversion and conscientiousness; excitement seeking (E5) on openness; and angry hostility (N2), warmth (E1), and assertiveness (E3) on agreeableness. Results of the CFA were compared with exploratory structural equation modeling (ESEM), a relatively new procedure incorporated into the Mplus statistical package. ESEM is similar to CFA, in that the number of latent factors is specified a priori, but ESEM differs from traditional CFA by freely estimating the loadings of all indicators on all factors (see Asparouhov & Muthén, 2009). That is, the requirement of zero cross-loadings is relaxed and it is possible to use an EFA measurement model with factor loading matrix rotations. ESEM also provides overall model fit, standard SEM parameters, and standard errors of the rotated parameters. It was expected that the ESEM approach would be better suited to the FFMRF data because it allowed cross-loadings among facets that have been previously demonstrated. Likely for that reason, the fit indices reported for FFM instruments in prior studies also suggested ESEM would better fit the data (Marsh et al., 2010). All analyses were conducted in Mplus using the maximum likelihood robust (MLR) estimator (Muthén & Muthén, 1998-2010). We considered this decision carefully but ultimately treated these indicators as continuous because no items in our sample had skew greater than 1 and simulation studies have shown that maximum likelihood performs adequately (Dolan, 1994). For ESEM, we used the Mplus defaults of oblique geomin rotation and an epsilon value of .5.

Next, multigroup CFA methods were used to test for factorial invariance across gender. The test of increasingly strict factorial invariance is well understood with the use of continuous variables (Horn & McArdle, 1992; Meredith, 1993). Although the exact sequence and specification of models for testing measurement invariance in a CFA framework often varies by author (Vandenburg & Lance, 2000), there is generally an agreed-on pattern of steps across studies. We followed the procedures set forth by Meredith (1993) in testing a series of increasingly more invariant and restrictive models, with specific modifications for ESEM multiple group models as outlined in the Mplus manual (Muthén & Muthén, 1998-2010). The first type of invariance tested was configural invariance, or whether the same measured items serve as indicators for the same factors in both men and women (Horn & McArdle, 1992). The factor loadings, item intercepts, and residual variances are left free to vary. Next, we tested for metric invariance, or whether the FFMRF has the same structure and meaning for men and women. Metric invariance, similar to configural, sets an equivalent factor pattern but adds the condition of equivalent factor loadings for men and women; the intercepts and residual variances are freely estimated in both groups. In the strong invariance model, the intercepts (corresponding to the origin of the scale) and factor loadings are constrained across groups. Finally, in a test of strict invariance, the factor loadings, intercepts, and residual variances are all constrained to be equal across men and women. While not technically included in measurement invariance testing, it is also possible to examine the equivalence of factor variances, covariances, and means across groups.

To investigate the overall fit of the CFA/ESEM models, we evaluated the chi-square, comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square of approximation (RMSEA), and standardized root mean square residual (SRMR) The chi-square test indexes the overall fit of the model but can be very sensitive to sample size, such that statistically significant chi-square values are often found in larger samples (Hu & Bentler, 1993). The CFI and TLI compare the hypothesized model with a more restricted, baseline model; CFI ranges from 0 to 1, with 0 indicating poor fit and 1 indicating a perfect fit, while TLI can in some cases exceed 1. Although there are not strict, empirically derived cutoffs for interpreting these goodness-of-fit indices (Marsh, Hua, & Wen, 2004), generally CFI and TLI values more than .95 are desirable (Hu & Bentler, 1999), whereas those below .90 indicate model fit can be improved. The RMSEA is a measure of the error of approximation of the specified model covariance and mean structures to the covariance and mean structures in the population. General guidelines suggest that models with an RMSEA of .08 or below produce an adequate fit (Browne & Cudeck, 1993). SRMR estimates the residuals, or discrepancy, between the sample covariance and model-implied covariance. The SRMR ranges from .00 to 1.00, with a value of .08 or below indicating good fit (Hu & Bentler, 1999). Generally, CFI and TLI ≥.90 and RMSEA ≤.10 provide the more commonly employed (e.g., Lance, Butts, & Michels, 2006) thresholds for acceptable fit that we use for interpretation here (Hopwood & Donnellan, 2010).

When comparing the fit of nested models, we used the chi-square difference test. Because of the use of an MLR estimator, it was necessary to compute a Satorra-Bentler chi-square difference test in Mplus (see www.statmodel.com). When comparing the chi-square of a baseline model with the chi-square of a constrained (more invariant) model, a nonsignificant difference suggests support for the more invariant (and thus parsimonious) model. In addition, research has suggested that when comparing two nested models, a change in CFI of less than .01 would lend support to the more parsimonious model (Cheung & Rensvold, 2002). The Bayesian Information Criterion (BIC) can also be used to compare nested and nonnested models. BIC is an information-theoretic fit criterion, which balances both fit (in the sense of a small chi-square) and parsimony (in the sense of a large number of degrees of freedom) of the model. Models that are better in terms of fit and parsimony result in lower BIC values. If the BIC values between two models differ by 0 to 2, there is only weak evidence in favor of the model with the smaller BIC value; if they differ from 2 to 6, there is positive evidence (3:1-20:1 odds); if they differ by 6 to 10, there is strong evidence (20:1-150:1 odds), and if they differ by more than 10 there is very strong evidence (>150:1 odds), that the model with the smaller value is the better-fitting model (Raftery, 1995). For our series of nested models to test measurement invariance, we attempted to find the model that demonstrated good fit according to absolute and comparative fit indices (i.e., a low RMSEA and SRMR, high CFI and TLI, nonsignificant chi-square difference, change in CFI of less than .01, and low BIC).

## Results

Descriptive statistics for each domain and facet from the FFMRF are presented in Table 1 for both men and women separately, along with independent samples *t* tests and Cohen's *d* values indicating gender differences. Internal consistencies of domains in the combined sample were .72 (neuroticism), .77 (extraversion), .65 (openness), .66 (agreeableness), and .80 (conscientiousness). Consistent with prior results for the NEO PI-R, there were a number of significant differences, with most in the small to medium (.20-.50) range according to Cohen (1992). In most comparisons (4 of 5 domains; 22 of 30 facets) the females obtained a higher score than the males. The largest difference at the domain level was for agreeableness (.37; with women higher than men) and the largest facet discrepancies were for tendermindedness and order (.44; women higher). The only facets where men scored significantly higher than women were values (−.20) and actions (−.20) from openness.

### Confirmatory Factor Analysis Versus ESEM

The first step in the data analysis was to determine whether CFA or ESEM was a more appropriate method for conducting the factorial invariance testing. Thus, confirmatory factor analysis and ESEM were conducted for the full sample and separately by gender to determine which method provided a more acceptable fit to the FFMRF data. Table 2 presents the results of CFA for the five-factor solution with correlated factors. The five-factor solution fit poorly in the total sample as well as separately in both men and women according to chi-square, CFI, TLI, and SRMR. RMSEA was close to acceptable, which has been found previously in CFA analysis of FFM data where other fit statistics are uniformly poor (see Marsh et al., 2010). In contrast, ESEM yielded more acceptable fit statistics according to several indices. In both men and women separately, ESEM provided an acceptable fit to the data (with the exception of TLI for men, which was somewhat low at .88). Although ESEM does not provide eigenvalues, we ran the comparable exploratory factor analysis (ML) and the first seven eigenvalues were 5.728, 3.085, 2.467, 1.838, 1.521, 1.030, and .988. As expected given the nature of ESEM, the loadings of the indicators on the factors and the factor correlations were generally smaller using ESEM than when using CFA. Having determined that ESEM provided the better fit to the data, it was appropriate to proceed with factorial invariance models using ESEM.

### Factorial Invariance Models

We followed the procedures of Meredith (1993) for conducting sequential multigroup factor analytic comparisons,

**Table 1.** Sample Descriptive Statistics for FFMRF Domains and Facets by Gender.

| | Male ($n$ = 357) | | Female ($n$ = 342) | | | | |
|---|---|---|---|---|---|---|---|
| FFMRF domains | Mean | *SD* | Mean | *SD* | *t*(697) | Cohen's *d* | 95% CI |
| Neuroticism (n) | 2.34 | 0.68 | 2.47 | 0.65 | 2.60* | .20 | .05, .35 |
| Extraversion (e) | 3.40 | 0.73 | 3.60 | 0.61 | 4.05** | .30 | .16, .46 |
| Openness (o) | 3.40 | 0.62 | 3.37 | 0.53 | −0.60 | −.05 | −.19, .00 |
| Agreeableness (a) | 3.45 | 0.59 | 3.66 | 0.52 | 4.91** | .37 | .22, .52 |
| Conscientiousness (c) | 3.60 | 0.62 | 3.77 | 0.60 | 3.64** | .28 | .13, .42 |
| Anxiousness (n1) | 2.33 | 1.09 | 2.78 | 1.03 | 5.55** | .42 | .27, .57 |
| Angry Hostility (n2) | 1.87 | 0.99 | 1.94 | 1.00 | 0.82 | .07 | −.09, .21 |
| Depressiveness (n3) | 2.32 | 1.20 | 2.24 | 1.07 | −0.93 | −.07 | −.22, .08 |
| Self-Consciousness (n4) | 2.82 | 2.03 | 2.99 | 1.03 | 2.15* | .11 | .01, .31 |
| Impulsivity (n5) | 2.55 | 1.03 | 2.37 | 0.94 | −2.49* | −.18 | −.34, −.04 |
| Vulnerability (n6) | 2.15 | 0.93 | 2.53 | 0.91 | 5.45** | .41 | .26, .56 |
| Warmth (e1) | 3.50 | 1.09 | 3.92 | 1.00 | 5.24** | .40 | .25, .55 |
| Gregariousness (e2) | 3.28 | 1.16 | 3.72 | 1.00 | 5.40** | .41 | .26, .56 |
| Assertiveness (e3) | 3.03 | 0.94 | 3.15 | 0.93 | 1.72 | .13 | −.02, .28 |
| Activity (e4) | 3.53 | 1.07 | 3.68 | 0.92 | 1.97 | .15 | .00, .30 |
| Excitement Seeking (e5) | 3.37 | 1.02 | 3.28 | 0.86 | −1.29 | −.10 | −.25, .05 |
| Positive Emotions (e6) | 3.68 | 1.00 | 3.87 | 0.88 | 2.80* | .20 | .06, .36 |
| Fantasy (o1) | 3.28 | 1.13 | 3.25 | 1.06 | −0.38 | −.03 | −.18, .12 |
| Aesthetics (o2) | 3.37 | 0.89 | 3.52 | 0.80 | 2.28* | .18 | .02, .32 |
| Feelings (o3) | 3.83 | 0.91 | 4.02 | 0.76 | 3.02** | .23 | .08, .38 |
| Actions (o4) | 3.03 | 0.93 | 2.85 | 0.88 | −2.71* | −.20 | −.35, −.06 |
| Ideas (o5) | 3.45 | 0.96 | 3.38 | 0.89 | −0.98 | −.08 | −.22, .07 |
| Values (o6) | 3.42 | 1.08 | 3.21 | 1.02 | −2.60* | −.20 | −.35, −.05 |
| Trust (a1) | 3.00 | 1.07 | 3.32 | 0.99 | 4.16** | .31 | .17, .46 |
| Straightforwardness (a2) | 3.75 | 0.97 | 3.93 | 0.83 | 2.61* | .20 | .05, .35 |
| Altruism (a3) | 3.48 | 0.92 | 3.72 | 0.80 | 3.63** | .28 | .13, .42 |
| Compliance (a4) | 3.57 | 0.91 | 3.68 | 0.82 | 1.59 | .13 | −.03, .27 |
| Modesty (a5) | 3.53 | 0.95 | 3.51 | 0.88 | −0.34 | −.02 | −.17, .12 |
| Tender-mindedness (a6) | 3.37 | 0.94 | 3.78 | 0.92 | 5.88** | .44 | .29, .59 |
| Competence (c1) | 3.70 | 0.86 | 3.75 | 0.81 | 0.68 | .06 | −.10, .20 |
| Order (c2) | 3.35 | 1.02 | 3.77 | 0.90 | 5.78** | .44 | .29, .59 |
| Dutifulness (c3) | 3.71 | 0.88 | 3.85 | 0.86 | 2.16* | .16 | .01, .31 |
| Achievement (c4) | 3.61 | 0.88 | 3.79 | 0.81 | 2.71* | .21 | .06, .35 |
| Self-discipline (c5) | 3.64 | 0.86 | 3.76 | 0.83 | 1.92 | .14 | .00, .29 |
| Deliberation (c6) | 3.60 | 0.78 | 3.70 | 0.75 | 1.72 | .13 | −.02, .28 |

*Note*. FFMRF = Five-Factor Model Rating Form; CI = confidence interval.
*$p$ < .05. **$p$ < .01. (two tailed).

with some modifications to allow identification of the models using ESEM (Asparouhov & Muthén, 2009; Muthén & Muthén, 1998-2010). These procedures are designed to apply increasingly strict constraints on the measurement parameters of the model. As noted previously, the four levels of invariance tested, in order from least to most strict, were configural, metric, strong, and strict invariance. Each of these four forms of invariance was applied to the five-factor ESEM solution. The ultimate goal of invariance testing is to be able to set enough parameters equal to test mean differences in the factor (or factors) underlying the scale. The invariance models are shown in Table 3.

The first type of invariance tested, configural, constrains the fewest number of parameters equal across gender. In a typical configural invariance test using CFA, the factor structure (i.e., the pattern of fixed and free loadings) is held equal across gender while the factor loadings are freely estimated in each group. With ESEM, configural invariance specified the same items and number of factors for each group, and cross-loadings are allowed for all items on all factors. Factor means were set to 0 and factor variances to 1.0 in both men and women. If configural invariance is supported, the same latent variable (or variables) are present in men and women. As shown in Table 3, the configural

**Table 2.** Fit statistics for Confirmatory Factor Analysis Exploratory Structural Equation Modeling of FFMRF Items.

| | $\chi^2$ | *df* | RMSEA | CFI | TLI | SRMR | BIC |
|---|---|---|---|---|---|---|---|
| CFAs | | | | | | | |
| Full sample | 1,387.767 | 389 | .061 | .79 | .76 | .078 | 53,478 |
| Women (*N* = 342) | 884.48 | 389 | .061 | .78 | .76 | .090 | 25,454 |
| Men (*N* = 357) | 861.085 | 389 | .058 | .81 | .79 | .074 | 28,114 |
| ESEM | | | | | | | |
| Full sample | 594.276 | 295 | .038 | .94 | .91 | .027 | 53,175 |
| Women (*N* = 342) | 438.51 | 295 | .038 | .94 | .91 | .033 | 25,476 |
| Men (*N* = 357) | 499.547 | 295 | .044 | .92 | .88 | .034 | 28,244 |

*Note.* $\chi^2$ = adjusted chi-square fit statistic with robust standard errors; *df* = degrees of freedom; RMSEA = root mean square error of approximation; SRMR = standardized root mean residual; CFA = confirmatory factor analysis; ESEM = exploratory structural equation modeling; BIC = Bayesian information criterion.

**Table 3.** Fit Statistics for Factorial Invariance Models Across Gender.

| Gender invariance models | $\chi^2$ | *df* | SCF | RMSEA | CFI | TLI | SRMR | BIC | Model comparison | $\Delta\chi^2$ | $\Delta df$ | *p* | $\Delta$CFI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tests of measurement invariance across groups | | | | | | | | | | | | | |
| 1. Configural invariance | 937.766 | 590 | 1.05 | .041 | .93 | .89 | .03 | 53998 | | | | | |
| 2. Metric invariance | 1052.64 | 715 | 1.09 | .037 | .93 | .91 | .05 | 53342 | 2 vs. 1 | 127.212 | 125 | 0.4281 | .003 |
| 3. Strong invariance | 1111.24 | 740 | 1.10 | .038 | .92 | .91 | .05 | 53244 | 3 vs. 2 | 57.886 | 25 | 0.0002 | .007 |
| 4. Invariance of residual variances of measured variables | 1195.80 | 770 | 1.09 | .040 | .91 | .90 | .06 | 53140 | 4 vs. 3 | 85.009 | 30 | 0.0000 | 0.012 |
| 5. Factor variances and covariances | 1139.47 | 755 | 1.10 | .038 | .92 | .91 | .06 | 53182 | 5 vs. 3 | 27.179 | 15 | 0.0273 | 0.003 |
| 6. Factor means invariant | 1256.17 | 760 | 1.10 | .043 | .90 | .88 | .07 | 53276 | 6 vs. 5 | 145.064 | 5 | 0.0000 | 0.024 |

*Note.* $\chi^2$ = adjusted chi-square fit statistic with robust standard errors; *df* = degrees of freedom; SCF = scale correction factor; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index; SRMR = standardized root mean residual; BIC = Bayesian information criterion.

invariance model produced adequate fit statistics in terms of RMSEA, CFI, and SRMR, although TLI was slightly lower than accepted standards ($\chi^2$ = 937.766, RMSEA = .041, CFI = .93, TLI = .89, SRMR = .03). Balancing the preponderance of evidence, we concluded that configural invariance of the FFMRF items holds across men and women.

The next level of invariance is metric invariance. This type of invariance assumes configural invariance (i.e., equivalent factor structure across gender) and adds the additional constraint of invariant factor loadings over groups. A factor loading is the linear relationship between the item and the latent factor, and metric invariance tests whether the factor loading matrices are equivalent across the groups being tested (here, men and women). Factor means again were set to 0 in both groups while factor variances were set to 1.0 in women and free in men. The test of metric invariance resulted in fit that was as good, if not better, than the model testing configural invariance according to RMSEA, CFI, TLI, and SRMR ($\chi^2$ = 1052.64, RMSEA = .037, CFI = .93, TLI =.91, SRMR = .05). Additionally, the substantial decrease in BIC (656) was very strong evidence in support of the metric invariance model. Finally, the chi-square difference test was not significant ($\Delta\chi^2$ = 127.212, $\Delta df$ = 125, $p$ = .43) and the change in CFI was less than .01. As such, the hypothesis of invariant factor loadings across gender was *not* rejected.

However, because metric invariance does not constrain the origin of the scale, it is not possible to compare factor means across gender having only established that level of invariance. Thus, we next tested strong invariance, the hypothesis that intercepts (i.e., the origins) linking the observed items to the latent factors were constant across gender. This test of strong invariance presumes both configural and metric invariance. The test of strong invariance resulted in acceptable RMSEA, CFI, TLI, and SRMR ($\chi^2$ = 1111.24, RMSEA = .038, CFI = .92, TLI = .91, SRMR = .05). BIC decreased 98 from metric invariance, strong support in favor of the more parsimonious model. The chi-square difference test was significant ($\Delta\chi^2$ = 57.886, $\Delta df$ = 25, $p < .001$), but the change in CFI was less than .01. We concluded based on these fit statistics that the weight of the

**Table 4.** Unstandardized Parameter Estimates for Strong Invariance Model.

| FFMRF Item | Neuroticism | Extraversion | Openness | Conscientiousness | Agreeableness |
|---|---|---|---|---|---|
| Anxiousness (N1) | **0.66** | −0.01 | **−0.13** | 0.01 | 0.08 |
| Angry Hostility (N2) | **0.50** | −0.01 | 0.05 | −0.01 | **−0.27** |
| Depressiveness (N3) | **0.65** | **−0.41** | 0.08 | 0.05 | −0.04 |
| Self-Consciousness (N4) | **0.49** | **−0.24** | −0.02 | 0.02 | **0.27** |
| Impulsivity (N5) | **0.30** | 0.10 | **0.31** | **−0.25** | −0.17 |
| Vulnerability (N6) | **0.48** | −0.01 | −0.02 | **−0.25** | **0.27** |
| Warmth (E1) | 0.05 | **0.43** | −0.03 | 0.05 | **0.37** |
| Gregariousness (E2) | −0.06 | **0.69** | −0.02 | 0.01 | 0.00 |
| Assertiveness (E3) | 0.04 | **0.37** | **0.09** | **0.21** | **−0.38** |
| Activity (E4) | **−0.09** | **0.38** | **0.11** | **0.25** | −0.05 |
| Excitement Seeking (E5) | −0.06 | **0.26** | **0.30** | 0.04 | **−0.16** |
| Positive Emotions (E6) | **−0.25** | **0.45** | **0.11** | **0.10** | **0.10** |
| Fantasy (O1) | **0.14** | 0.02 | **0.54** | **−0.13** | **0.20** |
| Aesthetics (O2) | 0.03 | **0.08** | **0.25** | **0.16** | 0.08 |
| Feelings (O3) | −0.03 | **0.11** | **0.16** | **0.23** | **0.14** |
| Actions (O4) | −0.01 | 0.03 | **0.53** | −0.06 | −0.03 |
| Ideas (O5) | 0.07 | −0.02 | **0.55** | 0.03 | 0.04 |
| Values (O6) | −0.08 | −0.08 | **0.48** | 0.04 | −0.02 |
| Competence (C1) | 0.03 | 0.00 | −0.02 | **0.53** | −0.02 |
| Order (C2) | **0.09** | **0.19** | **−0.23** | **0.53** | −0.04 |
| Dutifulness (C3) | −0.01 | 0.03 | **−0.12** | **0.51** | 0.07 |
| Achievement (C4) | −0.02 | **0.10** | 0.01 | **0.56** | −0.04 |
| Self-discipline (C5) | **−0.09** | 0.01 | −0.01 | **0.52** | 0.09 |
| Deliberation (C6) | 0.05 | **−0.08** | −0.02 | **0.42** | **0.10** |
| Trust (A1) | −0.06 | **0.24** | 0.06 | **−0.18** | **0.39** |
| Straightforwardness (A2) | −0.04 | **0.11** | 0.06 | **0.30** | **0.11** |
| Altruism (A3) | −0.04 | **0.11** | **0.11** | **0.17** | **0.31** |
| Compliance (A4) | −0.07 | −0.02 | −0.04 | **0.15** | **0.42** |
| Modesty (A5) | −0.01 | **−0.16** | **0.10** | **0.15** | **0.41** |
| Tender-mindedness (A6) | 0.08 | 0.13 | 0.04 | 0.01 | **0.62** |

*Note*. FFMRF= Five-Factor Model Rating Form. All loadings significant at $p < .05$ are shown in boldface. Total $N = 699$; men = 357, women = 342.

evidence supported the strong invariance model, allowing the comparison of latent factor means across groups.

The final form of invariance tested was strict invariance. If strict invariance is achieved, differences in means and variances of the FFMRF items can be explained by gender differences in the means and variances of the factors. This model adds the constraint of invariant unique variances across groups to the constraints of equal loadings and intercepts. The strict invariance test resulted in acceptable but slightly worse values for RMSEA, CFI, TLI, and SRMR ($\chi^2 = 1195.80$, RMSEA = .04, CFI=.91, TLI = .90, SRMR = .06). BIC did decrease by 104, but both the chi-square difference test ($\Delta\chi^2 = 85.009$, $\Delta df = 30$, $p < .001$), and change in CFI ($\Delta$CFI = .012) supported the strong invariance model over the strict invariance model. Given that strict invariance is an extremely conservative test, and not necessary to compare latent mean differences (Brown, 2006), we determined that the evidence was better in terms of strong invariance over strict invariance. Factor loadings for the best-fitting strong invariance model are presented in Table 4.

Although not typically included in tests of measurement invariance, we next moved to examining whether the factor variance, covariances, and means could be held equal across gender. Starting from the constraints of the strong invariance model, we fixed factor variances to 1.0 in both groups and constrained the factor covariances to be equal in men and women. This yielded an acceptable model fit according to all indices and a decrease in BIC of 62 and a change in CFI of less than .01 (see Model 5, Table 3). Finally, we tested a model that constrained the factor means to be equal across men and women (by fixing them to 0.0 in all groups, see Muthén & Muthén, 1998-2012). Compared with Model 5, this model led to a significant decrement in fit, with RMSEA raised by .005, SRMR increased by .01, CFI decreased by .024, TLI decreased by .03, and BIC increased by 94 (see Model 6, Table 3). This led us to conclude that the latent factor means were different in men and women. To compare these differences, we examined the best fitting strong invariance model, where the means of the five factors are set to 0 in women and freely estimated in men. Men were significantly lower on the first factor (Neuroticism, .46 standard deviations below women; $Z = -3.67$, $p < .001$), the second factor (Extraversion, .68 standard deviations, $Z = -5.11$, $p < .001$), the fourth factor (Conscientiousness, .20 standard deviations, $Z = -2.00$, $p = .045$), and the fifth factor (Agreeableness, .47 standard deviations, $Z = -3.90$, $p < .001$), and significantly higher than women on the third factor (Openness, .33 standard deviations above women, $Z = 2.91$, $p = .004$).

## Discussion

The FFMRF demonstrated a number of differences across gender. Specifically, men were notably lower than women on neuroticism, extraversion, agreeableness, and conscientiousness but slightly higher on openness. Results for the facets were consistent with these domain trends, with only a few exceptions. Many of these differences were significant but most effect sizes would be considered small. These mean-level gender differences were generally consistent with prior findings for the FFM across a variety instruments (Feingold, 1994) and the NEO PI-R domains and facets more specifically (Costa, Terracciano, & McCrae, 2001; Furnham et al., 2013). For example, Feingold's meta-analysis indicated women scored higher on anxiety, trust, and tendermindedness and the present results indicated women having higher scores on the facets of anxiousness ($d = .42$), trust ($d = .31$), and tendermindedness ($d = .44$). Feingold also suggested women were slightly higher than men on extraversion (particularly gregariousness), and these findings were robust at the domain level ($d = .30$) and for the facets of gregariousness ($d = .41$) and warmth ($d = .40$). Similarly, Furnham et al. (2013) found small to moderate differences on the NEO PI-R facets of vulnerability, positive emotions, and altruism; all of which obtained effect sizes larger than .20 in our sample.

There were, however, a few exceptions worth noting. Whereas Feingold (1994) reported negligible gender differences on the trait of orderliness, the conscientiousness facet of order (c2) was one of the largest effects ($d = .44$; women higher than men) for the FFMRF. Costa et al. (2001) also reported that women scored higher on the order facet on the NEO PI-R, but the effect was much more modest. Additionally, Feingold (1994) as well as Costa et al. (2001) reported that men scored more highly on assertiveness, whereas we found that females obtained slightly higher scores on the FFMRF facet of assertiveness ($d = .13$). We found little difference on the FFMRF openness domain scores ($d = -.05$), but this overall domain effect appears to be the product of divergent facet-level findings. Specifically, like Costa et al. (2001), we found that women tended to be higher than men on the facets of aesthetics and feelings, whereas men were modestly higher on values and ideas. Interestingly, though, the facet of actions in our sample was higher for males ($d = -.20$); whereas Costa et al. (2001) found that females scored higher on this facet. Future research that compares the nomological network of openness to actions as operationalized by these two measures is warranted.

Although not as frequent or pronounced as for openness, there were also examples from other domains where gender differences for individual facets trended in a direction opposite the broad domain. For example, women generally scored higher than men on neuroticism but men scored significantly higher on the facet of impulsiveness ($d = -.18$) and trended this way for depressiveness ($d = -.07$). Similarly, women were generally higher than men on extraversion and agreeableness, but the extraversion facet of excitement-seeking ($d = -.10$) and agreeableness facet of modesty ($d = -.02$), trended higher for men. If these patterns were pronounced they might have important implications for investigations of gender differences at the domain level. Indeed, this illustrates the value of a facet-level assessment as otherwise the broad domain analyses might mask more subtle, but still important, differences across gender. Future studies of gender differences in personality would be well-served to focus on the lower-order traits.

### *Factor Structure and Invariance Across Gender*

Beyond mean-level gender differences on the FFMRF, a novel and significant contribution of the present study was the investigation of measurement invariance models. Initial results confirmed a good fit of the FFM to a five-factor structure when using ESEM (Samuel et al., 2013). Not surprisingly the use of CFA provided a poor fit for the FFM (Hopwood & Donnellan, 2010; Marsh et al., 2010). Building on these ESEM findings we noted that the FFMRF obtained configural, metric, and strong invariance across genders. This suggests that mean scores can be effectively compared across genders (Brown, 2006). Although men and women do obtain significantly different mean scores on several domains and facets, these reflect true differences and not a systematic artifact or bias of the measure itself. Strong invariance indicates that the FFMRF is measuring the same latent trait domains in men and women, such that its hierarchical structure appears to be the same across gender.

The finding of strong factorial invariance across gender provides additional support for the use of the FFMRF as it suggests that its scores can be usefully applied in a variety of settings. These results also indicate areas where the FFMRF may show larger gender differences compared with prior findings from other FFM instruments (Costa et al., 2001). Specifically, the conscientiousness facet of order has typically been relatively balanced across the genders, with perhaps a small trend toward women being higher than men (Furnham et al., 2013). However, in the present sample, we found this to be among the largest differences ($d = .44$). This discrepancy between gender differences on the FFMRF, compared with those on the NEO PI-R (Costa et al., 2001) is all the more surprising as that facet obtained the strongest convergence between these two instruments ($r = .64$; Samuel et al., 2013). In addition, an inspection of the adjective descriptors on this FFMRF item (ordered, methodical, organized vs. haphazard, disorganized, sloppy) does not reveal a term that carries strong gender expectancy. Future research testing the gender invariance of other FFM instruments would be useful for quantifying the degree of difference between men and women on this personality trait.

The assertiveness facet from extraversion was another that obtained a gender difference in the present study that contrasted with prior studies. It was higher for females in our sample, but prior findings have indicated males score more highly on this facet (Feingold, 1994). The magnitude of the effect was small in our study and in the report from Costa et al. (2001), but it is nonetheless striking that the direction of the sign would be reversed. An inspection of the FFMRF item content did not immediately suggest wording that might account for this difference (dominant, forceful vs. unassuming, quiet, resigned), but future research, perhaps investigating gender differences on each specific adjective, might help to further explicate this finding.

## Limitations and Future Directions

Our student sample, not surprisingly, was more restricted in terms of age and educational attainment than the general population, which might limit the generalizability of our findings. Nonetheless, existing findings do not give reason to suspect that measurement invariance on any personality measure would be fundamentally affected by using college versus adult samples. Feingold (1994) examined this question for mean-level differences across a number of personality measures and found that "gender differences were essentially the same for college students as for general adults" and that the absolute values of any differences in effect sizes were "trivial" (p. 446). Nonetheless, it will be important to explore this question within a community sample to determine if our findings for the FFMRF replicate. For example, it is possible that the personality trait of assertiveness, as mentioned above, might be differentially predictive of enrolling in college for women and men. Thus, our findings might reflect differences in which individuals are in our sample (by virtue of being in college), rather than absolute differences in men and women. It should also be noted that our analyses focused on gender invariance within the FFMRF instrument. Although beyond the scope of the present report, future research should investigate predictive bias and might consider using a regression approach using external criteria. This would examine, for example, whether scores on FFMRF variables for men and women were equally predictive of (nonbiased) criteria.

## Conclusions

Within a large sample of undergraduate students that was evenly split between men and women, the FFMRF displayed mean differences across gender that were largely consistent with prior findings for other FFM measures (e.g., Costa et al., 2001; Feingold, 1994). Specifically, men were notably lower than women on neuroticism, extraversion, agreeableness, and conscientiousness but slightly higher on openness. Nonetheless, there were also gender differences that emerged at the facet level in the present study that had not been previously reported and will form the basis for additional research. Importantly, though, we found that the FFMRF displayed a five-factor structure that was invariant across gender. This factorial invariance adds important support for the validity of the FFMRF as a self-report measure as it indicates that the scores assess the same latent constructs in men and women.

**Declaration of Conflicting Interests**

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

**Funding**

The authors received no financial support for the research, authorship, and/or publication of this article.

**References**

Anderson, K. G., Sankis, L. M., & Widiger, T. A. (2001). Pathology versus statistical infrequency: Potential sources of gender bias in personality disorder criteria. *Journal of Nervous and Mental Disease*, *189*, 661-668. doi:10.1097/00005053-200110000-00002

Ashton, M. C., Lee, K., & Goldberg, L. R. (2004). A hierarchical analysis of 1,710 English personality-descriptive adjectives. *Journal of Personality and Social Psychology*, *87*, 707-721. doi:10.1037/0022-3514.87.5.707

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397-438.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment*, *9*, 9-30. doi:10.1111/1468-2389.00160

Block, J. (1995). A contrarian view of the 5-factor approach to personality description. *Psychological Bulletin*, *117*, 187-215. doi:10.1037/0033-2909.117.2.187

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

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.

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology*, *56*, 453-484. doi: 10.1146/annurev.psych.090902.141913

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159. doi: 10.1037/0033-2909.112.1.155

Costa, P. T., Jr., & McCrae, R. R. (1992). *Professional manual: Revised NEO Personaltiy Inventory and NEO Five-Factor Inventory*. Odessa, FL: PAR.

Costa, P. T., Jr., & McCrae, R. R. (2010). *NEO Inventories Professional Manual: NEO PI-3, NEO-FFI-3, NEO PI-R*. Lutz, FL: PAR.

Costa, P. T., Jr., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*, 322-331. doi: 10.1037/0022-3514.81.2.322

De Raad, B., & Perugini, M. (Eds.). (2002). *Big Five assessment*. Gottingen, Germany: Hogrefe & Huber.

Deary, I. J., Weiss, A., & Batty, G. D. (2010). Intelligence and personality as predictors of illness and dealth: How researchers in differential psychology and chronic disease epidemiology are collaborating to understand and address health inequalities. *Psychological Science in the Public Interest*, *11*(2), 53-79.

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

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical & Statistical Psychology*, *47*, 309-326.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (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

Eisenlohr-Moul, T. A., Walsh, E. C., Charnigo, R. J., Lynam, D. R., & Baer, R. A. (2012). The "what" and the "how" of dispositional mindfulness: Using interactions among subscales of the Five-Facet Mindfulness Questionnaire to understand its relation to substance use. *Assessment*, *19*, 276-286. doi:10.1177/1073191112446658

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, *116*, 429-456. doi:10.1037/0033-2909.116.3.429

Ferguson, C. J. (2010). A meta-analysis of normal and disordered personality across the life span. *Journal of Personality and Social Psychology*, *98*, 659-667. doi:10.1037/A0018770

Follingstad, D. R., Rogers, M. J., & Duvall, J. L. (2012). Factors predicting relationship satisfaction, investment, and commitment when women report high prevalence of psychological abuse. *Journal of Family Violence*, *27*, 257-273. doi:10.1007/s10896-012-9422-8

Furnham, A., Guenole, N., Levine, S. Z., & Chamorro-Premuzic, T. (2013). The NEO Personality Inventory–Revised: Factor structure and gender invariance from exploratory structural equation modeling analyses in a high-stakes setting. *Assessment*, *20*, 14-23. doi:10.1177/1073191112448213

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, 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

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

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117-144.

Howell, A. J., Dopko, R. L., Turowski, J. B., & Buro, K. (2011). The disposition to apologize. *Personality and Individual Differences*, *51*, 509-514. doi:10.1016/j.paid.2011.05.009

Hu, L., & Bentler, P. M. (1993). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 16-99). Newbury Park, CA: Sage.

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

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory: Versions 4a and 54*. University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy. In O. P. John, R. W. Robins & L. A. Pervin (Eds.), *Handbook of personality* (3rd ed., pp. 114-158). New York, NY: Guilford Press.

Kaiser, A. J., Milich, R., Lynam, D. R., & Charnigo, R. J. (2012). Negative urgency, distress tolerance, and substance abuse among college students. *Addictive Behaviors*, *37*, 1075-1083. doi:10.1016/j.addbeh.2012.04.017

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin*, *136*, 768-821. doi:10.1037/A0020327

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*, 202-220. doi: 10.1177/1094428105284919

Lowe, J. R., & Widiger, T. A. (2009). Clinicians' judgments of clinical utility: A comparison of DSM-IV with dimensional models of general personality. *Journal of Personality Disorders*, *23*, 211-229. doi:10.1521/pedi.2009.23.3.211

Lynam, D. R., & Widiger, T. A. (2001). Using the five-factor model to represent the DSM-IV personality disorders: An expert consensus approach. *Journal of Abnormal Psychology*, *110*, 401-412. doi:10.1037//0021-843x.110.3.401

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

Marsh, H. W., Ludtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the Big Five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*, 471-491. doi:10.1037/A0019227

McCrae, R. R., Terracciano, A., & Pro, P. P. C. (2005). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology*, *89*, 407-425. doi:10.1037/0022-3514.89.3.407

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525-543.

Miller, J. D., & Campbell, W. K. (2008). Comparing clinical and social-personality conceptualizations of narcissism. *Journal of Personality*, *76*, 449-476. doi:10.1111/j.1467-6494.2008.00492.x

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.

Mullins-Sweatt, S. N., & Widiger, T. A. (2009). Clinical utility and DSM-V. *Psychological Assessment*, *21*, 302-312. doi:10.1037/a0016607

Mullins-Sweatt, S. N., & Widiger, T. A. (2011). Clinician's judgments of the utility of the DSM-IV and five-factor models for personality disordered patients. *Journal of Personality Disorders*, *25*, 463-477. doi:10.1521/pedi.2011.25.4.463

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, *57*, 401-421. doi:10.1146/annurev.psych.57.102904.190127

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

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111-163.

Reynolds, S. K., & Clark, L. A. (2001). Predicting dimensions of personality disorder from domains and facets of the Five-Factor Model. *Journal of Personality*, *69*, 199-222. doi:10.1111/1467-6494.00142

Samuel, D. B., Mullins-Sweatt, S. N., & Widiger, T. A. (2013). An investigation of the factor structure and convergent and discriminant validity of the five-factor model rating form. *Assessment*, *20*, 24-35. doi:10.1177/1073191112455455

Samuel, D. B., & Widiger, T. A. (2004). Clinicians' personality descriptions of prototypic personality disorders. *Journal of Personality Disorders*, *18*, 286-308.

Samuel, D. B., & Widiger, T. A. (2006). Clinicians' judgments of clinical utility: A comparison of the DSM-IV and five-factor models. *Journal of Abnormal Psychology*, *115*, 298-308. doi:10.1037/0021-843X.115.2.298

Samuel, D. B., & Widiger, T. A. (2008). A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: A facet level analysis. *Clinical Psychology Review*, *28*, 1326-1342. doi:10.1016/j.cpr.2008.07.002

Samuel, D. B., & Widiger, T. A. (2009). Comparative gender biases in models of personality disorder. *Personality and Mental Health*, *3*, 12-25. doi:10.1002/pmh.61

Samuel, D. B., & Widiger, T. A. (2010a). Comparing personality disorder models: Cross-method assessment of the FFM and DSM-IV-TR. *Journal of Personality Disorders*, *24*, 721-745. doi:10.1521/pedi.2010.24.6.721

Samuel, D. B., & Widiger, T. A. (2010b). A comparison of Obsessive-Compulsive Personality Disorder scales. *Journal of Personality Assessment*, *92*, 232-240. doi:10.1080/00223891003670182

Samuel, D. B., & Widiger, T. A. (2011). Clinicians' use of personality disorder models within a particular treatment setting: A longitudinal comparison of temporal consistency and clinical utility. *Personality and Mental Health*, *5*, 12-28. doi:10.1002/pmh.152

Schenk, A. M., Ragatz, L. L., & Fremouw, W. J. (2012). Vicious dogs part 2: Criminal thinking, callousness, and personality styles of their owners. *Journal of Forensic Sciences*, *57*, 152-159. doi:10.1111/j.1556-4029.2011.01961.x

Schuettler, D., & Boals, A. (2011). The path to posttraumatic growth versus posttraumatic stress disorder: Contributions of event centrality and coping. *Journal of Loss & Trauma*, *16*, 180-194. doi:10.1080/15325024.2010.519273

Schwartz, R. L., Fremouw, W., Schenk, A., & Ragatz, L. L. (2012). Psychological profile of male and female animal abusers. *Journal of Interpersonal Violence*, *27*, 846-861. doi:10.1177/0886260511423254

Soto, C. J., & John, O. P. (2009). Ten facet scales for the Big Five Inventory: Convergence with NEO PI-R facets, self-peer agreement, and discriminant validity. *Journal of Research in Personality*, *43*, 84-90. doi:10.1016/j.jrp.2008.10.002

Thomas, K. M., Yalch, M. M., Krueger, R. F., Wright, A. G. C., Markon, K., & Hopwood, C. J. (2013). The convergent structure of DSM-5 personality trait facets and Five-Factor Model trait domains. *Assessment*, *20*, 308-311.

Vandenburg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70.