The Impact of NEO PI-R Gender-Norms on the Assessment of Personality Disorder Profiles

Douglas B. Samuel,
Department of Psychiatry, Yale University School of Medicine

Emily B. Ansell,
Department of Psychiatry, Yale University School of Medicine

Christopher J. Hopwood,
Department of Psychology, Michigan State University

Leslie C. Morey,
Department of Psychology, Texas A & M University

John C. Markowitz,
Department of Psychiatry, Columbia University College of Physicians & Surgeons and New York State Psychiatric Institute

Andrew E. Skodol, and
Department of Psychiatry, University of Arizona College of Medicine and The Sunbelt Collaborative

Carlos M. Grilo
Department of Psychiatry, Yale University School of Medicine

Abstract

Many personality assessment inventories provide gender-specific norms to allow comparison of an individual's standing relative to others of the same gender. In some cases, this means that an identical raw score produces standardized scores that differ notably depending on whether the respondent is male or female. Thus, an important question is whether unisex-normed scores or gender-normed scores more validly assess personality. We examined the gender-normed and unisex-normed scores from the NEO Personality Inventory – Revised (NEO PI-R; Costa & McCrae, 1992) in a large clinical sample, using two measures of personality disorder as validating criteria. Gender-normed scores did not obtain significantly higher correlations. In fact, for two personality disorders, antisocial and narcissistic, gender-normed scores yielded significantly lower correlations, suggesting that personality disorder pathology relates most closely to one's absolute level of a personality trait rather than one's standing relative to others of the same gender. We discuss ramifications of this finding for personality research and clinical assessment.

Correspondence regarding this article should be directed to the first author via (douglas.samuel@yale.edu) or VA Connecticut Healthcare, 950 Campbell Avenue 151-D, Building 35, West Haven, CT 06516.

This publication has been reviewed and approved by the Publications Committee of the Collaborative Longitudinal Personality Disorders Study. No conflicts of interest are represented by any authors.

Keywords

Gender; t-scores; NEO PI-R; FFM; Personality Disorders

There is a long history within psychological assessment of providing normed scores based on demographic variables such as gender and age. This practice carries the explicit message that identical raw scores on assessment instruments have different meanings depending on one's biological sex or age. For example, the Wechsler Intelligence Scale for Children-IV (WISC-IV; Wechsler, 2003) is standardized on the basis of age, such that a given raw score yields very different scaled scores for a 6-year-old than for a 16-year-old. Whereas the normed score indicates where the individual stands relative to others his or her own age, the raw score assesses the individual's absolute level of ability. It would not be surprising, then, if these two metrics related differently to external measures.

Personality assessment has a long history of standardizing scales relative to gender. The Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1940) provided separate norms for men and women. Other commonly used personality measures such as the Millon Clinical Multiaxial Inventory (Millon, Millon & Davis, 1996), the NEO Personality Inventory – Revised (NEO PI-R; Costa & McCrae, 1992), and the Schedule of Nonadaptive and Adaptive Personality (SNAP, Clark, 1993) also historically provided gender-specific norms.

The use of gender norms has permeated psychological assessment, but not without controversy. For instance, although the SNAP manual (Clark, 1993) endorses the use of gender norms, it also explicitly notes that "the long-held assumption that gendered norms provide a more valid basis for assessment is being challenged" and wonders whether gender differences on traits "may result from real differences in trait level rather than from culturally based differences in trait expression" (p. 58). This concern is emphasized further in the manual for the Personality Assessment Inventory (PAI), which recommends against gender-norming as the resultant T-scores might indicate similarity among groups that epidemiological studies have shown to differ (Morey, 1991).

Recently, the field has shifted toward unisex norms for newly developed measures as well as existing instruments. For example, the revised version of the SNAP (SNAP-2; Clark, Simms, Wu, & Casillas, in press) now provides unisex, rather than gender-specific norms. Even the MMPI-2-Restructured Form (MMPI-2-RF; Ben-Porath & Tellegen, 2008) now endorses the use of combined gender norms on the clinical scales. Finally, the MCMI-III has recently provided unisex-norms to replace the previous gender-specific norms. This industry-wide shift has occurred for two primary reasons, practical and conceptual. The practical reason is that many of these instruments are applied in employment and forensic settings, and recent interpretations of the 1991 Civil Rights Act and Americans with Disabilities Act have indicated that the use of any demographic norms (gender, age, race, etc.) in these settings could unlawfully discriminate against classes of individuals (for a review, see Sackett & Wilk, 1994).

The second, more conceptual reason is a shift in the field's understanding of what gender differences mean. It has been argued that gender differences on test scores from personality instruments might reflect differences in the expression of a trait, and that separate norms may correct for biases within a test. However, an alternative perspective is that test score differences reflect actual differences on the latent trait that are obscured by the use of gender-norms. For instance, Schinka, LaLone, and Greene (1998) indicated that demographic information, including gender, provided little incremental variance over patient status in predicting MMPI-2

scale scores and concluded that "these results call into question the continued use of separate gender-based norms" (p. 209).

With regard to gender differences in personality trait scores, Feingold (1994) conducted meta-analyses using the personality literature and normative data from well-known personality inventories. The results within the personality literature revealed consistent, if small, effect sizes as men scored more highly on measures of self-esteem, while women scored more highly on measures of general anxiety. The results within the normative data from existing instruments indicated that males scored more highly on measures of assertiveness, while females scored more highly on measures of extraversion, anxiety, trust, and tendermindedness. In both cases, Feingold reported that the magnitude of these differences was generally invariant across age and education level of the respondent as well as the nation where the study was conducted.

Although differences in personality traits across gender may be small effects that are less pronounced than for cognitive abilities (e.g., Hyde & Linn, 1988), they might still be meaningful. For example, a raw score of 134 on the NEO-PI-R neuroticism scale equates to t-scores of 80 for a male and 73 for a female. Nonetheless the presence of mean differences should be carefully distinguished from slope bias. Indeed, mean differences may simply indicate valid variation between men and women on personality traits. Slope bias, on the other hand indicates that the same score differentially predicts outcomes depending on one's gender. Thus, an important question for personality assessment is whether the gender-normed score relates more highly to external criteria, as gender standardizing would only be justified if it increased criterion-related validity.

Personality pathology is one particularly important validity indicator for personality traits. Over the past two decades, numerous published studies have supported the link between the general personality traits assessed by the NEO PI-R (as well as other instruments measuring traits of the five-factor model; FFM) and personality disorders (Widiger & Costa, 2002). A meta-analysis of 16 of these studies (Samuel & Widiger, 2008), a review of this research (Livesley, 2001), and an interbattery factor analysis (O'Connor, 2005) all concluded that robust links connect the dimensions of normal personality and the personality disorders (PD) listed in the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-IV*, APA, 2000). One consistent finding has been a strong relationship between neuroticism and borderline PD (Samuel & Widiger, 2008). Yet, since an identical raw score on NEO PI-R neuroticism yields standardized scores for men and women that differ by nearly a standard deviation, one might question how the gender-normed scores relate to borderline PD (BPD). To the extent that mean differences on NEO PI-R neuroticism scores merely reflect valid variation between genders then one would expect the unisex-normed score to be most predictive of BPD. However, if the mean differences on neuroticism scores instead reflects bias in the assessment of the trait (i.e., neuroticism is overestimated in females) then the gender-normed profile would prove superior.

One way to answer this question is to compare how NEO PI-R profiles composed of gender-normed and unisex-normed facet scores, respectively, relate to prototypic FFM profiles for each PD. Lynam and Widiger (2001) asked researchers to describe a prototypic case of a given PD in terms of the 30 facets of the FFM, using a 1 (*extremely low*) to 5 (*extremely high*) Likert-type scale. They then averaged the FFM ratings to reach a mean consensus profile for each PD. For example, a prototypic case of BPD obtained a mean rating of 4.75 on the neuroticism facet of angry hostility and a mean rating of 1.88 on the conscientiousness facet of deliberation. A subsequent study by Samuel and Widiger (2004) found that FFM descriptions by practicing clinicians converged strongly with those reported by Lynam and Widiger.

Researchers have proposed that the similarity between an individual's NEO PI-R profile and these consensus FFM ratings can function as indicators of the *DSM-IV* PDs (cf., Miller, Lynam,

Widiger, & Leukefeld, 2001): the more closely an individual's FFM profile corresponds to the prototypic profile for BPD, the more likely that person will evince BPD pathology. These FFM PD prototype scores have demonstrated convergent, discriminant, and predictive validity as well as temporal stability (Miller, Reynolds, & Pilkonis, 2004) and can increment prediction of behavioral outcomes beyond measures specifically designed to assess the PD (Trull, Widiger, Lynam, & Costa, 2003). Overall, their robust relations with FFM PD prototypes make explicit PD scales useful outcome variables with which to compare the validity of gender-normed and unisex-normed NEO-PI-R score profiles.

The current study utilizes the prototype matching technique to examine the relations of PD prototypes, computed using gender-normed and unisex-normed t-scores from the NEO-PI-R, with semi-structured interview and self-report measures of the *DSM-IV* PDs. Considering the historical emphasis on gender-norms within personality, we hypothesized that these scores would show greater validity than unisex-normed scores.

Method

Participants were 668 patients recruited from multiple clinical sites for the Collaborative Longitudinal Study of Personality Disorders project. Participants underwent clinical diagnostic interviews and completed self-reports as part of a standardized assessment process across sites (Gunderson et al, 2000). Clinical interviews were conducted prior to administration of the self-report packets. The sample had 425 women (64%) and was predominantly Caucasian (76%), but included 80 (11%) African-Americans, 62 (9%) Hispanics, 11 (2%) Asian-Americans, 2 Native Americans and 9 identifying their race as *other*. The age of participants ranged from 18 to 45 years, with a mean of 32.7 (*sd* = 8.1). The sample comprises individuals diagnosed with *DSM-IV* borderline (n = 175), schizotypal (n = 86), avoidant (n = 157), and obsessive-compulsive (n = 153), as well as a comparison group meeting criteria for major depressive disorder (MDD) but without a PD diagnosis (n = 97). Diagnoses were assigned using a semi-structured interview administered by trained assessors and confirmed by at least one additional diagnostic method. All participants were seeking or had recently received psychiatric treatment at the study's outset.

The MDD group was included in the data collection for numerous reasons (Gunderson et al., 2000), but primarily to serve as a clinical comparison with relatively lower levels of personality psychopathology. This is particularly important within the current analyses as it increases the variability within the NEO PI-R and PD scale scores, which in turn improves the power to detect significant relationships. To this end, participants were excluded from the MDD group if they had met a total of 15 or more PD symptoms or were within two criteria of any particular PD diagnosis. Including these individuals extended the range of personality pathology downward to ensure a more comprehensive sampling of the total range of the PD constructs. The MDD group demographically resembled the overall sample: it was primarily female (60%) and Caucasian (77%), with age ranging from 18 to 45 (mean 32.8 [*sd* = 8.0]). Complete demographic details for each diagnostic group are available elsewhere (Gunderson et al., 2000). Participants were followed longitudinally, but the current analyses use scores only from the baseline assessment to maximize the sample size. Post-hoc power analysis ($\alpha = .05$) indicates that this sample size provides power of .90 to detect effects as small as $d = 0.12$.

Instruments

Diagnostic Interview for DSM-IV Personality Disorders (DIPD-IV)—The DIPD-IV (Zanarini, Frankenburg, Sickel, & Yong, 1996), a semi-structured interview that assesses the diagnostic criteria for each of the *DSM-IV* PD diagnoses, was administered amid a comprehensive diagnostic interview by trained clinicians. Each diagnostic criterion is scored as 0 (not present), 1 (present but of uncertain clinical significance), or 2 (definitely present).

Administered by well-trained and monitored clinical assessors, the inter-rater reliability (based on 84 pairs of raters) kappas for the 12 PDs (10 formal and 2 research categories) ranged from .58 to 1.0 and test-retest kappas (based on two direct interviews of 52 cases) ranged from .69 to .74 (Zanarini et al., 2000). Inter-rater reliability for the dimensional PD ratings ranged from .69 (schizoid) to .97 (antisocial) in this sample (Zanarini et al., 2000). Values for each diagnostic criterion were summed to create a dimensional score for each PD. For antisocial personality disorder, the 15 childhood conduct disorder items were first averaged, and then summed with the adult criteria. Current sample alpha values for the scale scores ranged from .61 (schizoid) to .86 (borderline) with a median of .79.

Schedule of Nonadaptive and Adaptive Personality – 2 (SNAP-2; Clark, Simms, Wu, & Casillas, in press)—Consisting of 390 true/false statements, the SNAP-2 provides a self-report assessment of a dimensional model of personality disorder. It also includes scales assessing the *DSM-IV* (APA, 2000) PDs. These scale scores provide a dimensional assessment of the PDs and range in length from 19 (avoidant) to 34 (antisocial) items. In the current sample, the SNAP-2 PD scale scores obtained reasonable internal consistency, ranging from .69 (OCPD) to .88 (avoidant), with an overall median of .83. The SNAP-2 contains only minor modifications of PD scales from the original SNAP (Clark, 1993), which assessed the *DSM-III-R* PD constructs. Scores on the SNAP PDs obtained stability coefficients over a nine-month interval ranging from .59 (schizotypal) to .84 (antisocial), with a median of .75 within a non-clinical sample (Melley, Oltmanns, & Turkheimer, 2002). In addition, SNAP PD scores correlate strongly with scores from other self-report inventories (see Widiger & Boyd, 2009) and scores from a structured interview measure (Samuel & Widiger, in press).

NEO Personality Inventory – Revised (NEO PI-R; Costa & McCrae, 1992)—The NEO PI-R contains 240 statements that the individual rates on a Likert-type scale with the options *strongly disagree, disagree, neutral, agree*, or *strongly agree*. It assesses five broad domains of the Five Factor Model of personality (e.g., extraversion) as well as 30 facets that underlie these domains. In the current sample, alphas ranged from .58 to .85 for the 30 facet scale scores, with a median of .74. The NEO PI-R scores have strong temporal stability, with values ranging from .76 to .84 over a seven-year period (Costa, Herbst, McCrae, & Siegler, 2000) and have shown consistency across cultures (McCrae et al., 2005).

Data Analysis

The NEO PI-R includes 30 facet scales, each composed of eight items scored on a 0–4 Likert-type scale. Thus, the raw score for each facet is the sum of its eight component items, producing a score ranging between 0 and 32. These raw scores for each subject were converted to normed scores in two different ways. The first computed standardized t-scores based on gender-specific norms provided in the manual (Costa & McCrae, 1992). The second method converted the raw scores to t-scores based on unisex-norms also provided in the NEO PI-R manual. Thus, each participant had two separate NEO PI-R profiles, one of gender-normed t-scores and the other of unisex-normed t-scores. Both the unisex and gender-norms derive from a subset of adults drawn from three different samples (total N = 2273) collected by Costa and McCrae in the late 1980's and early 1990's. Groups of 500 men and 500 women were selected from this sample to closely match 1995 U.S. census data in terms of age and race. Further detail about the normative group can be found in Costa and McCrae (1992) or Costa, McCrae, and Dye (1991).

Five-factor model prototype matching scores were calculated using procedures outlined by Miller and colleagues (2001), by correlating the complete FFM profile for each participant with the mean consensus ratings of Lynam and Widiger (2001) for each of the 10 *DSM-IV* PDs. As an illustration, the first column of Table 1 presents a selected participant's unisex-

normed NEO PI-R profile, while the second column provides the mean consensus profile for borderline PD, as presented by Lynam and Widiger (2001). The correlation between these two profiles (r = .79) also appears at the bottom of Table 1 and indicates the degree to which this individual's NEO PI-R profile matches the borderline prototype. We refer to these as profile matching indices (PMIs).

As described above, each participant had separate NEO PI-R profiles using unisex norms and gender-norms. This yielded 20 PMIs (2 profiles x 10 PDs) for each participant that assessed the degree to which his or her NEO PI-R profiles matched the FFM prototype for each PD. Previous research has typically calculated PMIs using intraclass correlations, but we used standard Pearson correlations, since McCrae (2008) suggests that these methods perform quite similarly for calculating profile agreement. This was necessary as ICCs assess elevation as well as shape and thus require equivalent metrics. This was impossible given the use of t-score profiles, which do not have discrete distributions.

Results

Since each participant had two PMIs per PD, we first correlated these with one another to determine their similarity. The two metrics were extremely similar, suggesting limited effects of gender-standardizing. The first column of Table 2 provides correlations between the unisex-normed PMIs and the gender-normed PMIs, which were all above .978, with a median of .997. To assess the convergence between the two PD assessments, we computed the relationship between the SNAP-2 and DIPD-IV measures of each PD. Correlations ranged from .500 (narcissistic) to .735 (avoidant), with a median of .633.

The primary results of interest were the relationships of the unisex and gender PMIs with the respective PD scales from the DIPD-IV and SNAP-2. The third column of Table 2 presents the convergent correlations between the unisex PMIs and the PD scales from the DIPD-IV. The correlations for the unisex PMIs with the DIPD-IV ranged from .260 (OCPD) to .542 (avoidant), with a median of .355, while the gender PMIs ranged from .261 (OCPD) to .549 (avoidant), with a median of .343. Overall, this coefficient was higher for the gender-normed profiles for 3 PDs and higher for the unisex-normed PDs for 6.

Although differences between these two values were generally minor, significance tests were conducted. Because they were dependent correlations, significance testing relied on triangulating these values with the correlations between the two profiles (i.e., the values in the first column). These analyses indicated that the gender PMI did not obtain a significantly higher correlation with the DIPD-IV scores than did the unisex PMI for any of the PDs. In fact, the contrary was true for antisocial and narcissistic PDs. Table 2 shows that although most of the effects would be considered small (Cohen, 1992), the Cohen's d values for narcissistic and antisocial are medium-sized.

We repeated these analyses using the SNAP-2 PD scale scores as the criterion measure. Results resembled those for the DIPD-IV but had greater magnitude. This is not surprising, as the SNAP-2 shares self-report method variance with the NEO-PI-R that is not shared with the DIPD-IV. Values ranged from .343 (dependent) to .667 (avoidant) with a median of .535. The fourth column of Table 2 presents the convergent correlation between the gender PMIs and the SNAP-2 PD scale scores, ranging from .341 (dependent) to .671 (avoidant), with a median of .515 across the 10 PDs. Overall, this coefficient was higher for the gender-standardized profiles for 5 PDs and higher for the unisex-standardized PDs for 4. Consistent with the results for the DIPD-IV comparisons, the unisex PMIs correlated significantly more highly with the antisocial and narcissistic PDs from the SNAP-2. The gender PMIs were not significantly larger for any PD. The effect sizes generally approached zero and even the effect size for narcissistic, although

significant, was small. The effect size for the differences in antisocial (.47) neared the threshold to be considered large.

Discussion

Current evidence suggests that men and women obtain different mean scores on various personality measures. One approach to these observed differences is to provide gender-specific norms, which serve to equate men and women in terms of standardized t-scores. A second approach is to provide unisex-norms that standardize these differences in a way that reflects the observed differences. For personality measures, these two approaches yield normed scores that differ only subtly. For example, standardizing NEO PI-R scores based on the respondent's gender introduces differences between men and women that, with a few exceptions, are relatively minor. Consistent with this fact, the PMIs generated from gender-normed scores were similar to those generated using unisex norms. Nonetheless, the results of the current study suggest that although differences between them are small, they can be meaningful.

This was particularly true for antisocial and narcissistic PDs, for which unisex PMIs obtained higher correlations with the SNAP-2 and DIPD-IV scales. Differences between the gender and unisex PMIs for the other eight PDs were not statistically significant, despite attempting multiple statistical comparisons within a relatively large clinical sample. Although generally reluctant to draw conclusions from null findings, we find this case exceptional. Providing gender norms requires a specific effort, ostensibly to increase assessment validity. The finding that this extra step not only failed to enhance, but in two cases actually detracted from validity, is important. Minimally, it suggests the use of unisex-normed or raw scores instead of gender-normed scores in future studies employing the prototype matching technique or relating FFM measures to personality pathology. This result also supports the feedback provided in a section of the NEO PI-R's computer-generated Interpretive Report (Costa & McCrae, 1992), which offers clinical hypotheses based on the similarity between one's profile and PD prototypes. It is noteworthy that these hypotheses, unlike other feedback in the report, are not based on gender-normed scores.

There are numerous ways to compare the validity and utility of gender-normed scores. Previous research has typically examined each individual scale from a measure to see how norms affect its prediction of an outcome variable (e.g., Schinka et al, 1998). A contribution of the current study is that it considered the overall profile of NEO PI-R scores rather than specific scales. It is quite possible that there would be instances in which individual gender-normed scores from the NEO PI-R might obtain higher correlations than unisex-normed scores with certain outcome measures. However, our findings indicate that when aggregated across the 30 facets, the unisex scores are equally or even more valid, and suggest that gender-norms introduce systematic variance to the assessment of the NEO PI-R that is unrelated to measures of PD. Considering the entire profile of scores permits a more global comparison of the normed scores that avoids potential idiosyncrasies related to individual traits.

Despite longstanding discussion, the field of personality assessment remains divided on the validity of gender-normed scores and their use remains controversial (cf., concerns regarding gender differences in the Fake-Bad scale [Lees-Haley, 1992] of the MMPI-2; Butcher, Aribisi, Atlis, & McNulty, 2003). Perhaps some of this debate may be attributable to the analytic strategies that have examined individual scales rather than instrument-wide profiles. Future research that explores this possibility with other instruments and constructs is crucial.

Gender Differences in Personality Disorder Scores

Scores on PD scales are merely one potential comparison for the validity of gender and unisex-normed personality scores. An inherent complexity in using any particular criterion measure

is the possibility that the criterion is itself biased. The *DSM-IV-TR* (APA, 2000) does indicate different gender prevalence rates for several of the PD diagnoses, such that borderline, dependent, and histrionic PDs occur more frequently in women, whereas antisocial and narcissistic occur more commonly in men. Some have argued that these discrepancies reflect biases in the *DSM* constructs themselves, as they represent behaviors and traits that caricature normative gender roles (Kaplan, 1983). Others have noted that they may reflect genuine and valid gender differences in the occurrence of pathology (Kass, Spitzer, & Williams, 1983). A fundamental question is whether these prevalence rates reflect simple mean differences or indicate slope bias. For example, women scoring more highly than men on a measure of borderline PD is not a cause for concern (at least with respect to validity) if a given score is equally predictive of an outcome, such as self-harm behavior. However, if BPD scores relate differently to self-harm for men and women, then this would indicate bias within the construct or the measure.

The detection of bias with the PD constructs is a thorny, multilayered issue that researchers have approached in several ways. Some have found that female vignettes are more likely to be diagnosed with histrionic PD (e.g., Warner, 1978; Ford & Widiger, 1989; Flanagan & Blashfield, 2005) and Lindsay and Widiger 1995) suggested that scores on some self-report PD items show gender differences that are unrelated to maladaptive functioning Although the current study is not concerned with gender bias, per se, the possibility of bias among PD scores prompts some caution against overgeneralizing the present results. While PDs and personality pathology remain an important outcome for the comparison of gender and unisex norms on personality measures, future studies that use additional methods can importantly extend these results.

The FFM relates to a variety of important life outcomes (Ozer & Benet-Martinez, 2006) and researchers might investigate the validity of unisex and gender scores relative to external validators such as job performance, relationship quality, or physical health outcomes. A primary contribution of the current study is a novel method of testing the validity of gender norms by considering trait profiles rather than individual scales. Future could also employ method to explore the performance of gender norms with other outcomes. Although the utility of gender-standardizing in other contexts deserves further research, the results of the current study limit optimism about the validity of gender-normed scores from other personality assessment methods.

Limitations

The current study compared gender and unisex norms within a relatively large clinical sample carefully diagnosed with relevant personality pathology. Although the sample possesses notable strengths, its composition might also be considered a limitation as it was explicitly designed to target only four diagnostic categories. This strategy usefully provides comparisons among the studied diagnostic groups (avoidant, borderline, obsessive-compulsive, and schizotypal) but limits examination of the full range of personality pathology. The resulting distribution of that pathology could be “lumpy,” and non-normal distributions may affect the levels and covariation of the various traits and symptoms. Additionally, the selection of individuals meeting criteria for specific PDs means that scores for other PDs (e.g., antisocial) must be considered in the context of the other primary PD diagnoses. Future research should replicate these findings in a sample with more evenly distributed pathology. A second limitation is that this study examined the validity of the normed scores on only one measure, the NEO PI-R. Future studies need to replicate these findings with other personality measures. While the NEO PI-R appears relatively free of gender bias, such freedom must be demonstrated on a test-by-test basis before unisex norms can be applied. Finally, the calculation of any normed scores is inexorably linked to the normative sample; hence the importance of obtaining a representative normative sample for each measure cannot be overstated. Any variation within

a given normative sample can produce differences in the norms that greatly influence the eventual validity of the respective normed scores.

Conclusions

This study considered the validity of personality trait profiles generated using gender-norms and unisex-normed scores. The findings demonstrated no advantage for the gender-normed scores in correlations with self-report and interview measures of the *DSM-IV-TR* personality disorders. To the extent that any advantage emerged, it favored unisex-normed scores. Although for specific variables the gender-normed scores might potentially provide superior predictions (McCrae, Martin, & Costa, 2005), the current study indicates that one's absolute level on a personality trait, rather than standing relative to others of the same gender, provides the most valid assessment. Minimally, these findings lead us to recommend the continued use of raw or unisex normed scores when calculating FFM prototype matching indices or relating the NEO PI-R to personality disorders. Our findings also suggest caution in employing gender-standardized scores when interpreting results from other personality assessment instruments.

Acknowledgments

Writing of this manuscript was supported by the Office of Academic Affiliations, Advanced Fellowship Program in Mental Illness Research and Treatment, Department of Veterans Affairs. Research was supported by NIMH grants MH 50837, 50838, 50839, 50840, 50850, and MH75543 (Hopwood).

References

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. Text revision. 4. Washington, DC: Author; 2000. rev

Ben-Porath, YS.; Tellegen, A. MMPI-2-RF (Minnesota Multiphasic Personality Inventory-2-Restructured Form) Manual for administration, scoring, and interpretation. Minneapolis, MN: University of Minnesota Press; 2008.

Butcher JN, Arbisi PA, Atlis MM, McNulty JL. The construct validity of the Lees-Haley fake bad scale. Does this scale measure somatic malingering and feigned emotional distress? Archives of Clinical Neuropsychology 2003;18:473–485. [PubMed: 14591444]

Clark, LA. Manual for the Schedule for Nonadaptive and Adaptive Personality. Minneapolis, MN: University of Minnesota Press; 1993.

Clark, LA.; Simms, LJ.; Wu, KD.; Casillas, A. Manual for the Schedule for Nonadaptive and Adaptive Personality (SNAP-2). Minneapolis, MN: University of Minnesota Press; (in press)

Cohen J. A power primer. Psychological Bulletin 1992;112:155–159. [PubMed: 19565683]

Costa PT Jr, Herbst JH, McCrae RR, Siegler IC. Personality at midlife: Stability, intrinsic maturation, and response to life events. Assessment 2000;7:365–378. [PubMed: 11151962]

Costa, PT., Jr; McCrae, RR. Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual. Odessa, FL: Psychological Assessment Resources; 1992.

Costa PT Jr, McCrae RR, Dye DA. Facet scales for Agreeableness and Conscientiousness: A revision of the NEO Personality Inventory. Personality and Individual Differences 1991;12:887–898.

Feingold A. Gender differences in personality: A meta-analysis. Psychological Bulletin 1994;116:429–456. [PubMed: 7809307]

Flanagan EH, Blashfield RK. Gender acts as a context for interpreting diagnostic criteria. Journal of Clinical Psychology 2005;61:1485–1498. [PubMed: 16173082]

Ford M, Widiger TA. Sex bias in the diagnosis of histrionic and antisocial personality disorders. Journal of Consulting and Clinical Psychology 1989;57:301–305. [PubMed: 2708619]

Gunderson JG, Shea MT, Skodol AE, McGlashan TH, Morey LC, Stout RL, Zanarini MC, Grilo CM, Oldham JM, Keller MB. The Collaborative Longitudinal Personality Disorders Study: development, aims, design, and sample characteristics. Journal of Personality Disorders 2000;14:300–315. [PubMed: 11213788]

Hathaway, SR.; McKinley, JC. The MMPI manual. New York: Psychological Corporation; 1940.

Hyde JS, Linn MC. Gender differences in verbal ability A meta-analysis. Psychological Bulletin 1988;104:53–69.

Kaplan M. A woman's view of DSM-III. American Psychologist 1983;38:786–792. [PubMed: 6614624]

Kass F, Spitzer RL, Williams JB. An empirical study of the issue of sex bias in the diagnostic criteria of DSM-III Axis II personality disorders. American Psychologist 1983;38:799–801. [PubMed: 6614626]

Lees-Haley PR. Efficacy of the MMPI-2 validity scales for detecting spurious PTSD claims: F, F-K, Fake Bad scale, Ego Strength, Subtle-Obvious subscales, DIS and DEB. Journal of Clinical Psychology 1992;48:681–689. [PubMed: 1401155]

Lindsay KA, Widiger TA. Sex and gender bias in self-report personality disorder inventories: Item analyses of the MCMI-II, MMPI, and PDQ-R. Journal of Personality Assessment 1995;65:1–20. [PubMed: 16367643]

Livesley, WJ. Conceptual and taxonomic issues. In: Livesley, WJ., editor. Handbook of personality disorders: Theory, research, and treatment. New York: Guilford; 2001. p. 3-38.

Lynam DR, Widiger TA. Using the five-factor model to represent the *DSM-IV* personality disorders: An expert consensus approach. Journal of Abnormal Psychology 2001;110:401–412. [PubMed: 11502083]

McCrae RR. A note on some measures of profile agreement. Journal of Personality Assessment 2008;90:105–109. [PubMed: 18444102]

McCrae RR, Terracciano A. 78 members of the Personality Profiles of Culture Project. Universal features of personality traits from the observer's perspective. Journal of Personality and Social Psychology 2005;88:547–561. [PubMed: 15740445]

McCrae RR, Martin TA, Costa PT Jr. Age trends and age norms for the NEO Personality Inventory – 3 in adolescents and adults. Assessment 2005;12:363–373. [PubMed: 16244117]

Melley AH, Oltmanns TF, Turkheimer E. The Schedule for Nonadaptive and Adaptive Personality (SNAP): Temporal stability and predictive validity of the diagnostic scales. Assessment 2002;9:181–187. [PubMed: 12066833]

Miller JD, Lynam DR, Widiger TA, Leukefeld C. Personality disorders as extreme variants of common personality dimensions. Can the Five-factor model of personality adequately represent psychopathy? Journal of Personality 2001;69:253–276. [PubMed: 11339798]

Miller JD, Reynolds SK, Pilkonis PA. The validity of the five-factor model prototypes for personality disorders in two clinical samples. Psychological Assessment 2004;16:310–322. [PubMed: 15456386]

Millon, T.; Millon, C.; Davis, R. Millon Clinical Multiaxial Inventory: MCMI- III. Upper Saddle River, NJ: Pearson Assessments; 1996.

Morey, LC. The Personality Assessment Inventory: Professional manual. Odessa, FL: Psychological Assessment Resources; 1991.

O'Connor BP. A search for consensus on the dimensional structure of personality disorders. Journal of Clinical Psychology 2005;61:323–345. [PubMed: 15468325]

Ozer DJ, Benet-Martinez V. Personality and the prediction of consequential outcomes. Annual Review of Psychology 2006;57:401–421. [PubMed: 16318601]

Sackett PR, Wilk SL. Within-group norming and other forms of score adjustment in preemployment testing. American Psychologist 1994;49:929–954. [PubMed: 7985886]

Samuel DB, Widiger TA. Clinicians' descriptions of prototypic personality disorders. Journal of Personality Disorders 2004;18:286–308. [PubMed: 15237048]

Samuel DB, Widiger TA. 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 2008;28:1326–1342. [PubMed: 18708274]

Samuel DB, Widiger TA. Comparing Personality Disorder Models: Cross-Method Assessment of the FFM and DSM-IV-TR. Journal of Personality Disorders. (in press).

Schinka JA, LaLone L, Greene RL. Effects of psychopathology and demographic characteristics on MMPI-2 scale scores. Journal of Personality Assessment 1998;70:197–211. [PubMed: 9697327]

Trull TJ, Widiger TA, Lynam DR, Costa PT Jr. Borderline personality disorder from the perspective of general personality functioning. Journal of Abnormal Psychology 2003;112:193–202. [PubMed: 12784828]

Warner R. The diagnosis of antisocial and hysterical personality disorders. Journal of Nervous and Mental Disease 1978;166:839–845. [PubMed: 722306]

Wechsler, D. Wechsler Intelligence Scale for Children. 4. San Antonio, TX: Harcourt Assessment, Inc; 2003.

Widiger, TA.; Boyd, S. Assessing personality disorders. In: Butcher, JN., editor. Oxford handbook of personality assessment. 3. New York: Oxford University Press; 2009. p. 336-363.

Widiger, TA.; Costa, PT, Jr. Five-Factor model personality disorder research. In: Costa, Paul T., Jr; Widiger, Thomas A., editors. Personality disorders and the five-factor model of personality. 2. Washington, DC, US: American Psychological Association; 2002. p. 59-87.

Zanarini, MC.; Frankenburg, FR.; Sickel, AE.; Yong, L. The Diagnostic Interview for DSM-IV Personality Disorders (DIPD-IV). McLean Hospital; Belmont, MA: 1996.

Zanarini MC, Skodol AE, Bender D, Dolan R, Sanislow C, Schaefer E, Morey LC, Grilo CM, Shea MT, McGlashan TH, Gunderson JG. The collaborative longitudinal personality disorders study: reliability of axis I and axis II diagnoses. Journal of Personality Disorders 2000;14:291–299. [PubMed: 11213787]

Table 1

Illustrative Example of Five-Factor Model Prototype Matching Technique

	Subject 1009 (t-scores)	Borderline Prototype (mean ratings)
(n1) Anxiousness	66.42	4.04
(n2) Angry Hostility	77.39	4.75
(n3) Depressiveness	75.37	4.17
(n4) Self-Consciousness	72.05	3.17
(n5) Impulsiveness	73.18	4.79
(n6) Vulnerability	96.15	4.17
(e1) Warmth	57.75	3.21
(e2) Gregariousness	46.88	2.92
(e3) Assertiveness	44.04	3.17
(e4) Activity	64.55	3.29
(e5) Excitement Seeking	55.31	3.88
(e6) Positive Emotions	60.67	2.63
(o1) Fantasy	58.98	3.29
(o2) Aesthetics	62.08	2.96
(o3) Feelings	71.75	4.00
(o4) Actions	59.73	4.00
(o5) Ideas	68.00	3.21
(o6) Values	53.17	2.88
(a1) Trust	46.90	2.21
(a2) Straightforwardness	26.82	2.08
(a3) Altruism	51.14	2.46
(a4) Compliance	40.25	2.00
(a5) Modesty	57.38	2.83
(a6) Tendermindedness	37.14	2.79
(c1) Competence	38.00	2.71
(c2) Order	30.95	2.38
(c3) Dutifulness	21.28	2.29
(c4) Achievement Striving	28.75	2.50
(c5) Self-Discipline	6.28	2.33
(c6) Deliberation	26.83	1.88
	correlation = .79	

Notes: Column 1 presents a selected NEO PI-R profile composed of unisex-normed t-scores. Below the columns is the Pearson correlation between the profiles. The data in the final column present the mean FFM ratings (1–5 scale) for a prototypic case of borderline personality disorder adapted from "Using the Five-Factor Model to Represent the *DSM-IV* Personality Disorders: An Expert Consensus Approach" by D.R. Lynam and T.A. Widiger, 2001, *Journal of Abnormal Psychology, 110,* p. 404. Copyright by the American Psychological Association.

Table 2

FFM Prototype Matching Scores Correlated with SNAP-2 and DIPD-IV Scales

			Correlation with DIPD-IV scales				Correlation with SNAP-2 scales			
	Unisex with Gender	**SNAP- 2 with DIPD**	**Unisex**	**Gender**	***t(645)***	***d***	**Unisex**	**Gender**	***t(645)***	***d***
Paranoid	.994	.650	.466	.464	0.52	.04	.575	.578	−0.85	−.07
Schizoid	.999	.533	.355	.355	0	.00	.597	.596	0.76	.06
Schizotypal	.999	.593	.261	.263	−1.18	−.09	.427	.430	−1.89	−.15
Antisocial	.990	.638	.355	.330	4.87***	.38	.524	.496	6.01***	.47
Borderline	.999	.679	.424	.422	1.26	.10	.589	.591	−1.41	−.11
Histrionic	.999	.540	.294	.292	1.19	.09	.468	.469	−0.64	−.05
Narcissistic	.983	.500	.454	.425	4.52***	.36	.546	.533	2.14*	.17
Avoidant	.993	.735	.542	.549	−1.8	−.14	.667	.671	−1.16	−.09
Dependent	.979	.634	.285	.274	1.42	.11	.343	.341	0.26	.02
Obsessive	.999	.632	.260	.261	−0.59	−.05	.378	.378	0	.00

Notes: Unisex = Correlation with PD scale for FFM prototype matching scores calculated using unisex-normed t-scores; Gender = Correlation with PD scale for FFM prototype matching scores calculated using facets gender-normed t-scores; Raw with Unisex = the Pearson correlation among the two profiles; SNAP-2 with DIPD-IV= the Pearson correlation between the SNAP-2 and DIPD-IV scales; All prototype matching scores were calculated using Pearson correlations. *d* = Cohen's *d* value effect size.

* p <.05,

** p <.01,

*** p <.001.

All t-tests are two tailed. Lowest *n* for each comparison = 648.