Psychopathology

Research Article

Psychopathology 2020;53:168–178
DOI: 10.1159/000507377

Received: December 1, 2019
Accepted after revision: March 19, 2020
Published online: May 19, 2020

A Common Metric for Self-Reported Severity of Personality Disorder

Johannes Zimmermann[a] Steffen Müller[a] Bo Bach[b] Joost Hutsebaut[c]
Benjamin Hummelen[d] Felix Fischer[e]

[a]Department of Psychology, University of Kassel, Kassel, Germany; [b]Center for Personality Disorder Research, Psychiatric Research Unit, Region Zealand, Slagelse, Denmark; [c]Viersprong Institute for Studies on Personality Disorders Halsteren and Centre of Expertise on Personality Disorders, Utrecht, The Netherlands; [d]Department of Research and Development, Clinic of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway; [e]Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health, Berlin, Germany

Keywords
Severity · Personality disorder · Personality functioning · Item response theory · Patient-reported outcomes · Bifactor model

Abstract
Introduction: Dimensional models of personality disorders (PD) in the DSM-5 and ICD-11 share a focus on impairments in self and interpersonal functioning to represent the general features and severity of PD. This new perspective has led to the development of numerous measures for assessing individual differences in PD severity. While this improves choices for researchers and practitioners, it also poses the challenge of an increasing lack of standardization. ***Objective:*** The aim of this study is to establish a common metric across 6 widely used self-report measures of PD severity using item response theory models. ***Methods:*** 849 participants completed a survey including the Inventory of Personality Organization – 16-item version (IPO-16), the Level of Personality Functioning Scale – Brief Form 2.0, the Level of Personality Functioning Scale – Self-Report, the Operationalized Psychodynamic Diagnosis – Structure Questionnaire Short Form, the Personality Inventory for DSM-5 – Brief Form Plus and the Standardized Assessment of Severity of Personality Disorder (SASPD). We fitted exploratory multidimensional graded response models and used bifactor rotation to extract a general factor across measures. Factor scores were linked to representative T scores using data from a representative survey of 2,502 participants who completed the IPO-16. ***Results:*** When using bifactor rotation in a 7-factor model, all items loaded positively on the general factor, and the general factor explained 65.5% of the common variance. With the exception of the SASPD, all measures provided highly discriminating items (factor loadings >0.70) for measuring the general factor and reached an acceptable reliability (>0.80) across a wide range of the latent continuum. We constructed a crosswalk table linking total scores of the 6 measures to each other and to representative T scores. ***Conclusions:*** Our results suggest that 6 different self-report measures of the severity of PD capture a strong common factor and can therefore be scaled along a single latent continuum. Our results may facilitate instrument-independent assessment of severity of PD and increase comparability across studies.

karger@karger.com
www.karger.com/psp

Karger

Johannes Zimmermann
Department of Psychology, University of Kassel
Holländische Strasse 36–38
DE–34127 Kassel (Germany)
jz@uni-kassel.de

Introduction

The field of personality disorders (PDs) is currently undergoing a paradigm shift from a categorical to a dimensional approach to the classification of personality pathology. The most prominent examples of this process are the Alternative DSM-5 Model for Personality Disorders (AMPD) [1] and the PD chapter in ICD-11 [2]. Both models focus on impairments in self and interpersonal functioning to represent the general features and severity of PD. This new perspective has led to the development of numerous self-report measures for assessing individual differences in the severity of PD. While this improves choices for researchers and practitioners, it also poses the challenge of an increasing lack of standardization. That is, it is not fully clear whether the available measures assess the same construct and how scores obtained from these measures can be compared. The aim of this study is to establish a common metric across 6 widely used measures of PD severity using item response theory (IRT) models. This may facilitate instrument-independent assessment of the severity of PD and increase comparability across studies.

Severity of PDs

Explicitly considering the severity of PDs in the DSM-5 and ICD-11 is an important step forward for several reasons. First, general severity has been shown to be a strong predictor of current or future functioning [3, 4] and has also proved to be a good clinical tool in treatment planning and implementation [5–8]. Moreover, a general severity factor can account for the high comorbidity among PD diagnoses [9], and individual PD criteria that are central to this factor may inform or replace the general criteria for PD in section II of the DSM-5. It should also be noted that general severity seems to be sensitive to change, as was illustrated by a recent report from the Collaborative Longitudinal Personality Disorders Study [10].

Severity of PD can be conceptualized in different ways [11–13]. Kernberg [14] provided an early and highly influential account of conceptualizing personality pathology according to severity. He suggested that the level of personality organization manifests itself in three functional areas: (1) the integration of one's identity (i.e., the ability to establish nuanced and stable images of the self and others), (2) the maturity of defense mechanisms (i.e., the ability to process threatening internal and external stimuli in an adaptive way) and (3) the integrity of reality testing (i.e., the ability to distinguish between internal and external stimuli and establish contact with a socially shared reality). Kernberg further identified three levels of severity based on the level of impairment in these functional areas, namely neurotic, borderline and psychotic personality organization. More recent psychodynamic conceptualizations of severity (e.g., the Level of Structural Integration Axis of Operationalized Psychodynamic Diagnosis) [15, 16] tend to be similar to Kernberg's model in that they refer to impairments in basic psychological capacities and differentiate among several prototypical levels of functioning [17, 18].

Some aspects of Kernberg's model can also be found in the Level of Personality Functioning Scale (LPFS) [19], representing criterion A in the AMPD. Based on an exhaustive review of 5 psychodynamic measures, Bender et al. [19] concluded that the central feature of PD is the inability to understand and regulate oneself and one's interactions with others. In line with this idea, the LPFS proposes one dimension of generalized severity that is expressed in different areas of personality functioning, namely identity and self-direction (self-functioning) as well as empathy and intimacy (interpersonal functioning). Clinicians are expected to assess the general level of impairment on an ordinal scale ranging from 0 (no or little impairment) to 4 (extreme impairment).

Severity of PD can also be defined by the pervasiveness of personality pathology across different disorders, clusters or trait domains [11, 20]. For example, the AMPD includes a trait model, referred to as criterion B, which includes 25 pathological personality trait facets organized within 5 broad domains (i.e., negative affectivity, detachment, antagonism, disinhibition and psychoticism). This trait model could be used as a measure of severity by simply summing up the number of pathological traits [21]. Accordingly, the official instructions for the Personality Inventory for DSM-5 (PID-5) [22] imply that a high trait score indicates severity in terms of problematic areas for the individual receiving care [1]. Similarly, the ICD-11 classification of PDs infers that the traits are used to describe the personality features that contribute to personality disturbance and that individuals with more severe personality disturbance tend to have a greater number of prominent trait domains [2].

Common Metrics

The shift to dimensional approaches to PD has resulted in the development of various new self-report measures of PD severity [23]. Although this is a necessary and important process that expands researchers' and practitioners' choices, it comes with the drawback of an in-

creasing lack of standardization. The new instruments differ in various ways (e.g., underlying theoretical conceptualizations, emphasis on different aspects of the construct, length and precision, etc.), and one of the main challenges is that data obtained through different measures are hard to compare. For example, measures are available that are based on psychodynamic conceptualizations (e.g., Operationalized Psychodynamic Diagnosis – Structure Questionnaire Short Form [OPD-SQS] [24, 25]; Inventory of Personality Organization [IPO] [26, 27]), AMPD criterion A (LPFS – Brief Form [LPFS-BF] [28]; LPFS – Self-Report [LPFS-SR] [29]), AMPD criterion B (PID-5) [22] and ICD-11 (Standardized Assessment of Severity of Personality Disorder [SASPD] [30]). Despite their semantic similarities [31], it is not clear whether these measures assess the same construct and how scores obtained from these measures can be compared.

In recent years, IRT has been used to develop instrument-independent scales, calibrating different measures of the same outcome on a common metric. By explicitly formulating the relationship between observed item responses and an unobservable, latent variable, one defines the underlying trait and is further able to estimate the level of the latent trait for individual persons (i.e., factor scores or theta estimates) using any subset of items included in the model. Therefore, the level of the latent trait can be estimated using different questionnaires and, ideally, comparable estimates can be obtained. Usually, such models are estimated in large calibration samples, often calibrated to some reference population, and can later be applied in practice using crosswalk tables of sum scores or directly estimating factor scores [32]. This approach has been successfully applied for different self-reported outcomes, including depression [33], distress [34], physical function [35] and fatigue [36]. Using the same approach, legacy measures have been calibrated to the PROMIS scales within the landmark PROSETTA stone project [37, 38]. First validation studies have shown that common metrics can be indeed used to obtain comparable group level scores using different questionnaires [39, 40].

The aim of this study is to establish such a common metric for self-reported severity of PD. To this end, we collected data from 6 widely used measures or their short forms in a German community sample and estimated a joint IRT model to link item responses to an underlying general factor. This may facilitate the instrument-independent assessment of severity of PD and increase comparability across studies.

Materials and Methods

Procedure

Participants were recruited via the survey provider clickworker.de, which allowed for representative sampling in terms of age and gender. After providing informed consent, participants completed a survey that included sociodemographic questions, a brief measure of current symptom distress as well as 6 measures of PD (see below). To ensure data integrity, bogus items were implemented in the survey (e.g., "Please select 'not at all' here"). Participants were automatically excluded from the survey if they answered 2 out of 4 bogus items incorrectly. Participants received 5 EUR as monetary compensation for completing the full survey.

Sample

A total of 924 participants aged 18 or older successfully completed the survey. We excluded 34 participants who took less than 8 min to complete the full survey (less than 2.7 s per questionnaire item) and 5 participants who were identified as careless responders by Mahalanobis distance scores in excess of 3 standard deviations from the sample average. Thirty-six entries were excluded because they were made from the same IP address within a time frame of 3 h, which we deemed indicative of fraudulent software usage.

The final sample consisted of 849 individuals. Their age (mean = 42.6; SD = 16.1; range = 18–82) and gender (50% were female) distributions were roughly representative of the German population (Table 1). Participants aged 21 and younger were oversampled to ensure adequate variance within this age group. The level of education was skewed towards highly educated individuals, with 59.0% having completed an A-level degree, as opposed to 31.9% in the general population. Current symptom distress, assessed by a short form of the Symptom Checklist (SCL-K-9) [41], was more than a standard deviation (d = 1.18) above the population average. Similarly, compared to estimates for the general population in Germany [42], a relatively high number of participants (13.1%) stated that they were currently receiving psychotherapeutic treatment. More detailed information on sample characteristics can be found in online supplementary Table S1 (for all online suppl. material, see www.karger.com/doi/10.1159/000507377).

Measures

Inventory of Personality Organization – 16-Item Version

The IPO-16 [43, 44] is a short form of the IPO [26], which assesses the level of personality organization according to Kernberg [14]. It contains 16 items describing impairments in three domains of functioning: identity, defense and reality testing. Items are presented with a 5-point response scale ranging from 1 ("never true") to 5 ("always true"), with higher scores representing greater levels of personality pathology. In the general population, the internal consistency of the total (average) score was very high (Cronbach's α = 0.91).

Level of Personality Functioning Scale – Brief Form 2.0

The LPFS-BF [28, 45] is a brief self-report questionnaire for assessing criterion A of the AMPD [1]. It consists of 12 items corresponding to the 12 subdomains of the LPFS. Respondents are asked to rate the 12 items on a 4-point Likert scale from 1 ("completely untrue") to 4 ("completely true"). The LPFS-BF items capture both self-functioning and interpersonal functioning to an equal extent. In patient samples, the internal consistency of the total (sum) score was high (Cronbach's α = 0.82).

Table 1. Sample characteristics in comparison to the general population

	Sample		German population	
	male, % (*n*)	female, % (*n*)	male, %	female, %
Age groups				
18–20 years	5.8 (49)	5.8 (49)	2.0	1.8
21–30 years	8.6 (72)	8.4 (71)	7.6	7.0
31–40 years	7.4 (62)	8.9 (75)	7.6	7.3
41–50 years	8.6 (72)	8.9 (75)	8.1	8.0
51–60 years	10.5 (88)	10.7 (90)	9.5	9.5
61–70 years	7.5 (63)	5.5 (46)	6.8	7.3
71+ years	2.1 (18)	1.4 (12)	7.4	10.3
Education				
Did not graduate	0.1 (1)	0.0 (0)	1.8	1.7
General secondary school	4.5 (38)	3.9 (33)	16.5	16.1
Intermediate secondary school	15.3 (129)	16.0 (145)	14.9	17.1
A levels (Abitur/Fachabitur)	30.4 (256)	28.5 (240)	16.9	15.2

Seven participants identified their gender as nonbinary. Population data from Statista – The Statistics Portal and from Demografieportal.

Level of Personality Functioning Scale – Self-Report

The LPFS-SR [29] is a comprehensive self-report measure for assessing criterion A of the AMPD. It captures descriptions of 5 different levels of impairment in the domains of identity, self-direction, empathy and intimacy. It includes 80 items that are rated on 4-point Likert scales ranging from 1 ("totally false, not at all true") to 4 ("very true"). In the present study, we simplified the scoring scheme and used the total (average) score, reverse-coding the 12 items that describe adaptive aspects of personality functioning. In the construction sample, the internal consistency of a weighted sum score was very high (Cronbach's α = 0.97), and a principal component analysis of the 4 domain scores provided evidence for a large first component explaining 85.5% of the variance.

Operationalized Psychodynamic Diagnosis – Structure Questionnaire Short Form

The OPD-SQS [25] is a brief self-report questionnaire for assessing impairments in structural capacities, as described in the OPD system [15]. It consists of 12 items that are rated on a 5-point Likert scale ranging from 0 ("not true at all") to 4 ("fully true"). It captures 3 subdomains of personality functioning, including self-perception, interpersonal contact and relationship model. In a sample of psychosomatic out- and inpatients, the internal consistency of the total (sum) score was high (Cronbach's α = 0.89), and the general factor explained roughly 70% of the common variance [46].

Personality Inventory for DSM-5 – Brief Form Plus

The PID-5 Brief Form Plus (PID5BF+) [47] is a brief self-report measure for assessing the 6 pathological trait domains described in AMPD criterion B and the ICD-11: negative affectivity, detachment, antagonism/dissociality, disinhibition, anankastia and psychoticism. The 34 items were selected from 17 facet scales of the full PID-5 [22] using ant colony optimization algorithms. Items are rated on a 4-point scale ranging from 0 ("very false or often false") to 3 ("very true or often true"). The internal consistency of the domain scores in 3 large samples was high (mean McDonald's ω = 0.81), and domain scores were substantially positively correlated with each other.

Standardized Assessment of Severity of Personality Disorder

The SASPD [30] is a 9-item self-report measure that provides an index of PD severity. The SASPD is substantially derived from the 8-item Standardized Assessment of Personalty-Abbreviated Scale [48], with the exception of one item covering callousness, which was added to the SASPD. The SASPD is somewhat unusual because it includes 9 items that are rated using 0–3 response options with unique descriptions. The SASPD may be considered an index of severity in terms of PD complexity rather than a unidimensional scale of impairment because it captures 9 distinct and sometimes opposing PD features, which are separately rated in terms of severity. Thus, the internal consistency of the total (sum) score was rather modest in the construction sample (Cronbach's α = 0.76).

Symptom Checklist – Short Form

The SCL-K-9 [41] is a brief screening measure for the severity of global psychological distress during the past week. It was constructed by selecting the item with the highest correlation with the total score from each of the 9 scales of the full SCL-90-R [49]. Items are rated on 5-point Likert scales ranging from 0 ("not at all") to 4 ("extremely"). In a representative sample from the general population, the SCL-K-9 appeared to be unidimensional, and the internal consistency of the total (sum) score was high (Cronbach's α = 0.87).

Statistical Analyses

We adopted an exploratory approach to determine the underlying structure of the 163 items from 6 PD measures. To this end, we fitted a series of multidimensional IRT models with the "mirt" package [50] in the statistical environment R [51]. In particular, we estimated graded response models [52] assuming a multivariate Gaussian distribution with an increasing number of latent factors,

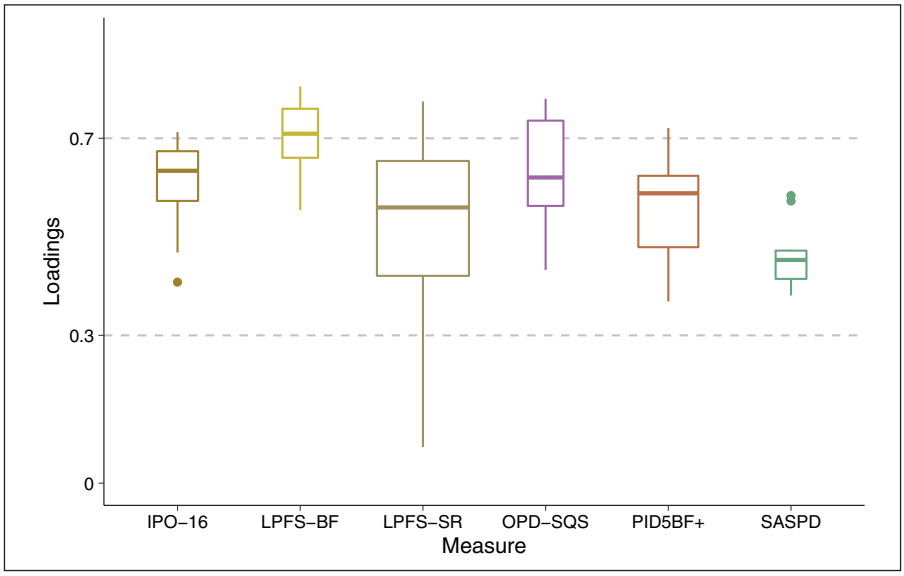

Fig. 1. Box and whisker plots summarizing the factor loadings on the general factor for 6 different measures of PD severity. The lower, middle and upper hinges correspond to the 25th, 50th and 75th percentiles. The upper and lower whiskers extend from the hinges to the largest and smallest values no further than 1.5 × interquartile range.

ranging from 1 to 10. Models were estimated using the quasi-Monte Carlo expectation-maximization algorithm with a total of 5,000 quasi-Monte Carlo integration nodes. The optimal number of factors was selected using the Bayesian information criterion as well as considering evidence from parallel analysis. Model fit was further evaluated using the collapsed M2* statistic [53] and several derivatives, including the root mean square error of approximation (<0.06), the comparative fit index (>0.95) and the standardized root mean square residual (<0.08) [54]. We used bifactor rotation because we assumed a general underlying factor driving most of the differences in item responses and evaluated the strength and reliability of the general factor using explained common variance (>0.60) and omega hierarchical (>0.70) [55].

We then estimated the expected a posteriori factor scores (theta estimates) for each possible total score for each measure [56] to allow easy transformation of total scores to the common metric.[1] Because factor score estimation in multidimensional IRT models is computationally expensive given the exponentially growing number of quadrature points, we fixed the parameters of the specific factors at zero and kept the parameters of the general factor fixed at their estimated value for this purpose (i.e., we used a unidimensional IRT model for score estimation) [36]. The standard errors of the theta estimates were plotted as a function of theta to provide information on the ranges in which measures provide acceptable (>0.80) reliability. We also linked theta estimates to T scores (mean = 50, SD = 10) representing the distribution of PD severity in the general population by estimating the latent mean and variance with fixed item parameters in an additional sample of 2,502 participants who completed the IPO-16 [44]. In order to investigate agreement among the theta estimates based on different measures, we compared these using Bland-Altman plots [57]. Finally, we explored associations between theta estimates and sociodemographic and clinical variables.

[1] In accordance with previous scoring practice, we used average scores for the IPO-16 and PID5BF+ and sum scores for the LPFS-BF, OPD-SQS and SASPD. To avoid confusion with the original weighted sum score for the LPFS-SR, we also used the average score for this measure.

Results

Parallel analysis suggested extracting up to 11 factors, the Bayesian information criterion suggested to extract 7 factors (see online suppl. Fig. S1). Further model fit indices confirmed an acceptable fit of the 7-factor model, $M2^*(11,729) = 23{,}216.05$, $p < 0.001$, root mean square error of approximation = 0.034 (90% confidence interval: 0.033–0.035), comparative fit index = 0.983, standardized root mean square residual = 0.036.

When using bifactor rotation in a 7-factor model, all items loaded positively on the general factor (Fig. 1). The general factor was relatively strong, with 65.5% explained common variance. Moreover, the omega hierarchical was 0.988, indicating that almost the entire variance in total scores can be attributed to a single general factor. Except SASPD, all measures contributed items with very high (>0.70) factor loadings (see online suppl. Tables S2 and S3 for details). The 22 highly discriminating items covered impairments in all 4 DSM-5 domains of functioning (LPFS-SR, 8 items; LPFS-BF, 7 items), as well as problems in self-perception (OPD-SQS, 4 items), identity diffusion (IPO-16, 1 item), and the DSM-5 trait facets anxiousness and anhedonia (PID5BF+, 2 items). The 11 items with rather low factor loadings (<0.30) all came from the LPFS-SR and were often those with positive item content (i.e., reverse-coded). Note that the factor loadings of the general factor had a 0.99 correlation with factor loadings from a one-dimensional IRT model, suggesting that the meaning of the general factor is robust across models with different numbers of factors.

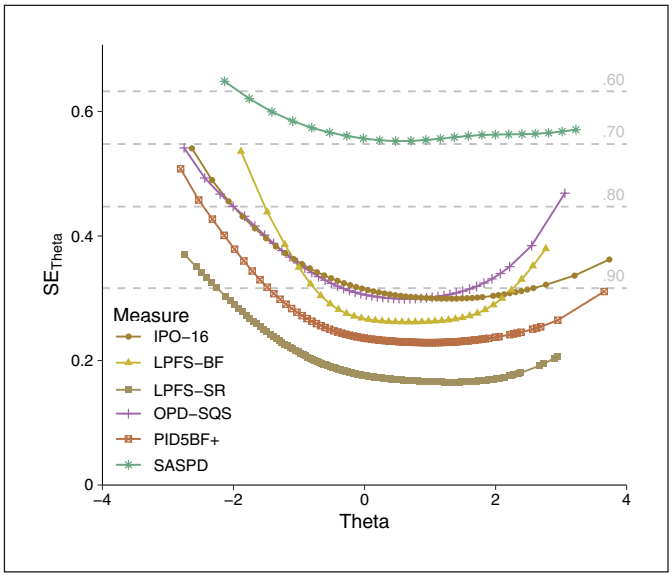

Fig. 2. Standard error of theta estimates as a function of theta for 6 different measures of PD severity. Dotted gray lines indicate corresponding levels of reliability.

Results for the 6 specific factors (SF) suggested the presence of at least two broader stylistic dimensions beyond the general factor. SF2 was defined by 11 items (with absolute loadings >0.40) from various measures capturing the bipolar dimension of fearless egocentrism versus anxiousness, and SF3 captured intimacy problems and social withdrawal (11 items). The remaining 4 SFs are rather weakly saturated, method-specific or difficult to interpret. In particular, SF1 could be interpreted as a method factor capturing mostly LPFS-SR items with positive content (6 items); SF4 contained mostly LPFS-SR items representing impairments in self-direction (5 items); SF6 was defined by LPFS-SR items representing an interdependent self-construal (2 items); and SF5 did not show any substantial loadings at all. Since we were only interested in measuring the general factor, we do not take the SFs any further into account.

Reliability of total scores of all measures was acceptable (>0.80) across a wide range of the latent severity continuum, with the exception of the SASPD (Fig. 2). The highest reliability was achieved by LPFS-SR and PID5BF+, while the reliability of LPFS-BF, IPO-16 and OPD-SQS approached 0.90 at least in the middle range of severity ($0 < \theta < 2$). Reliability was generally lower in the healthier spectrum ($\theta < -1$), and also decreased to some extent in the more severe spectrum ($\theta > 2$). Reliability of the SASPD's total score was consistently below 0.70.

Color version available online

Figure 3 provides a crosswalk between total scores of different measures of PD severity. A theta estimate of 0 corresponded to a T score of 55.7 in the general population, indicating that personality problems were slightly elevated in the current sample.[2] Adopting a normative perspective, an average amount of personality pathology (T = 50) corresponds with the following scores on the 6 measures: IPO-16_{1-5} = 1.69, LPFS-BF_{12-48} = 17, LPFS-SR_{1-4} = 1.66, OPD-SQS_{0-48} = 13, PID5BF+$_{0-3}$ = 0.50 and $SASPD_{0-27}$ = 4. In contrast, a highly elevated severity of PD (T = 70) corresponds with the following scores: IPO-16_{1-5} = 3.56, LPFS-BF_{12-48} = 40, LPFS-SR_{1-4} = 2.90, OPD-SQS_{0-48} = 41, PID5BF+$_{0-3}$ = 1.91 and $SASPD_{0-27}$ = 15. More detailed information on the crosswalk can be found in online supplementary Table S4 and Figure S2.

To help judge the accuracy of the links between the 6 measures, we present Bland-Altman plots in online supplementary Figure S3. The mean of the differences (i.e., bias) in theta estimates between each pair of measures for the full sample was often close to zero and peaked at 0.07 for the comparison of PID5BF+ and OPD-SQS. This suggests that even the largest systematic distortion between two linked questionnaires is less than 7% of a standard deviation in theta estimates. However, the 95% limits of agreement ranged from –1.07 to 1.09 to –1.75 to 1.77, indicating that two estimates of PD severity for the same individual measured using two questionnaires can be spread widely over the continuum. This is, however, at least partly due to the imperfect reliability of the questionnaires and suggests that converting scores in individual cases is associated with considerable uncertainty.

Finally, theta estimates based on all items were not associated with being male, $r = 0$, 95% confidence interval (–0.07, 0.06), or having an A-level degree, $r = -0.04$ (–0.10, 0.03), but were negatively associated with age, $r = -0.29$ (–0.34, –0.21). Moreover, theta estimates were positively associated with lifetime psychotherapy, $r = 0.27$ (0.20, 0.33), and current symptom distress, $r = 0.77$ (0.75, 0.80). The effect sizes (R^2) of the significant associations were slightly lower when using theta estimates from specific measures, with average relative decreases in R^2 ranging from 8.4% for LPFS-SR to 34.6% for SASPD (mean = 18.3%; see online suppl. Fig. S4).

[2] The freely estimated latent mean and variance in the general population sample using the fixed item parameters for the IPO-16 items from the current sample were –0.732 and 1.658, respectively. This means that in the general population, the average level of severity is lower and the variation of severity is larger compared to the current sample. Conversely, the mean and standard deviation of the current sample can be expressed in population-based T scores, with mean = 55.7 and SD = 7.8.

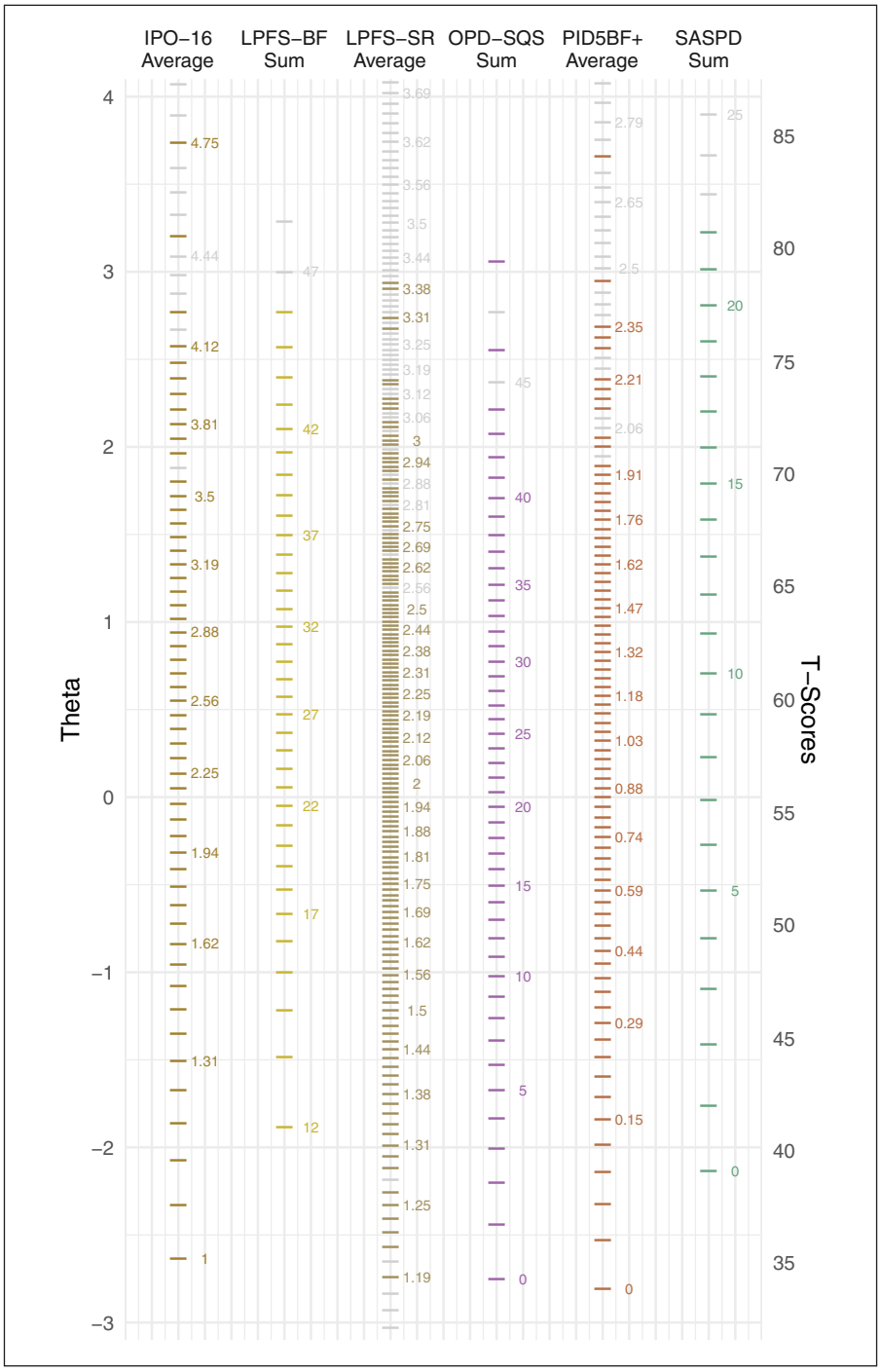

Fig. 3. Crosswalk of total scores between 6 different measures of PD severity. Gray total scores were not observed in this sample.

Discussion/Conclusion

This study aimed to establish a common metric across 6 self-report measures of PD severity. Our results suggest that all instruments assess a strong common factor and can therefore be scaled along a single latent continuum. The general factor was broadly defined by impairments in self- and interpersonal functioning, with a slight preponderance of internalizing personality pathology (e.g., anxiousness, low self-esteem). This suggests that PD severity based upon psychodynamic concepts (IPO-16, OPD-SQS), criterion A (LPFS-SR, LPFS-BF) and criterion B (PID5BF+) of the AMPD and the ICD-11 (SASPD) converge to a large extent when implemented in a self-

report format. Thus, our findings extend the family of common metrics to the field of PD assessment [32].

Among the instruments selected, the LPFS-SR provides the most comprehensive assessment of PD severity, including 80 items. Although its factor structure has raised discussions [58], the present study suggests that its large and heterogeneous pool of items reflects a strong general factor. Due to its high precision along the full range of the latent continuum (i.e., healthy to severe), it seems especially suited for individual assessment. The IPO-16, LPFS-BF and OPD-SQS are considerably shorter (12–16 items), but still provide sufficient precision for screening purposes. In contrast, the SASPD seems suboptimal in terms of reliability, which confirms recent findings [59, 60]. One explanation for this may be that the SASPD predominantly captures externalizing and other-related problems (e.g., potential harm to others), whereas the general factor extracted in this study was slightly more saturated with internalizing and self-related problems (see also Bach and Anderson [59]).

Although originally developed to assess pathological trait domains, we found that the PID5BF+ can also be used for assessing severity of PD. This is in line with the ICD-11 model's notion that the pervasiveness or complexity of impairments across different trait domains may be an important indicator of severity [20]. Moreover, this conceptualization is also reflected in the official user instructions for the PID-5 [1], which state that the PID-5 may be used to "track change in the severity of the individual's personality dysfunction over time." From an empirical perspective, this is also consistent with previous findings showing that PID-5 scores align with various measures of functioning [23].

Practical and Theoretical Implications

The study provides tentative norms for each of the 6 measures based on the German general population. For individual cases, practitioners can use the web platform (http://www.common-metrics.org/) to estimate T scores (general PD severity) including 95% confidence intervals based on individual item data. This may inform practitioners as to whether a patient reported elevated (T = 60) or highly elevated (T = 70) personality problems compared to the general population. One advantage of this approach is that missing item responses can be handled more efficiently. Additionally, the common metrics approach provides a crosswalk for converting scores from one measure to another. However, due to considerable uncertainty in individual cases, it is recommended to use this approach mainly for combining whole data sets from studies that applied different measures (e.g., individual participant data meta-analyses [61]).

Conceptually, our results question the assumed distinction between criteria A and B of the AMPD. In line with previous studies showing strong overlap between the two criteria [23], our results suggest that information about the extent of pathological traits may also indicate the general severity of PD. The implications of this finding for a future revision of the classification system are ambiguous: a more conservative conclusion would be that the two criteria reflect the same phenomena from two different clinical perspectives and traditions, both of which are clinically useful and justified. But there are also critical perspectives that find the lack of parsimony problematic: while some scholars argue that criterion A can be dispensed with due to its low incremental validity [e.g., 58], other scholars suggest replacing the pathological personality traits of criterion B with normal personality traits (e.g., the Big Five) in order to better capture the stylistic expression of personality regardless of the severity of the disorder [21, 62]. In any case, our findings support the notion of an "essentially unidimensional" continuum of generalized PD severity, which underlies the initial design of the LPFS [19].

Limitations and Future Directions

We must be transparent about the fact that our results may depend to some extent on our methodological decisions while analyzing the data. For example, we included all items, extracted several factors to achieve good model fit and used bifactor rotation to scale the general factor. A different strategy would be to select a core set of optimal items (e.g., using ant colony optimization) [63], and then integrate the remaining items using linkage techniques [33]. Although we expect that such a procedure would lead to similar results (e.g., when estimating theta), we did not actually test this. Moreover, we used bifactor rotation because we wanted to focus on the general factor. However, it is also possible to use oblique rotation, which may be better suited for assessing the (correlated) pathological trait domains capturing individual differences at a lower level of the hierarchy [22]. Our results do not contradict but are fully consistent with such a perspective. Future studies may also explore the possibility of nonmonotonic item response functions (e.g., generalized graded unfolding models) [64] or nonnormal latent distributions (e.g., semiparametric factor analyses) [65], both of which seem to be plausible candidate models for the severity of PD.

This study provides a proof of concept but is limited in several ways. First, although our sample was fairly rep-

resentative of the German population in terms of age and gender, the participants were better educated and more psychologically distressed than average. This is typical for samples recruited from online platforms [66, 67] and leads to doubts regarding the generalizability of the findings to the general population. At the same time, compared to clinical and treatment-seeking populations, severe personality pathology was probably underrepresented in our sample, which may have reduced the amount of multidimensionality in item responses [but see 68, 69]. Moreover, the sample size was rather small given the large number of parameters; thus, we were not able to test for differential item functioning (e.g., by age or gender). Future studies refining a common metric approach to PD should include large, representative samples of general and clinical populations from multiple countries.

Second, we only included a somewhat arbitrary selection of PD measures, and several other instruments are available, e.g., the Severity Indices of Personality Problems-118 [70], the General Assessment of Personality Disorder [71] and the DSM-5 Levels of Personality Functioning Questionnaire [72]. Future studies may wish to increase the item pool by combining several samples with different measures and estimating joint IRT models using multiple imputation techniques.

Third, it seems important to investigate the convergent and discriminant validity of a common metric for PD severity. As a preliminary step, we explored associations with sociodemographic and clinical variables and found that effect sizes differed somewhat depending on the measure used. However, across all measures, we found a very high correlation between theta estimates and current symptom distress, suggesting that self-reported PD measures may capture, at least to some extent, rather unspecific complaints or stress reactions caused by current circumstances. Longitudinal studies including measures of constructs that are conceptually distinct from PD are needed to disentangle these sources empirically [e.g., 10].

Fourth, a general concern regarding the assessment of PD severity with self-reports is that typical self-report items can be easily faked. This may not be a great problem in many clinical and research settings, but unsatisfying in (e.g.) forensic contexts. To address this problem, the development of forced-choice assessments that balance the social desirability of response options seems promising [73]. However, even honest respondents may lack insight into some of their behavioral patterns that others may regard as highly maladaptive. This calls for the use of expert ratings [74], informant reports [75] and performance-based measures [76] that may compensate for some of the limitations inherent to the self-report measures of PD severity applied in this study. Thus, a common metrics approach based on self-reports can only support but not replace the comprehensive multimethod assessment of PD that is recommended for clinical practice [12].

Statement of Ethics

This study was conducted ethically in accordance with the World Medical Association Declaration of Helsinki. All participants gave their informed consent to participate in the study.

Disclosure Statement

The authors have no conflicts of interest to declare.

Funding Sources

The data collection for this project was funded by the Institut für Verhaltenstherapie-Ausbildung Hamburg (IVAH) as part of a collaborative project on evaluating instruments for assessing personality pathology. The IVAH did not take part in data preparation, data analyses or manuscript preparation.

Author Contributions

J.Z. conceptualized the study, analyzed the data and drafted and finalized the manuscript, S.M. collected the data and wrote the sample description, F.F. supported the data analyses and critically revised the first draft, and B.H., J.H. and B.B. wrote parts of the introduction and discussion and gave detailed feedback on the first draft.

References

1 American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. 5th ed. Arlington (VA): American Psychiatric Association; 2013.

2 World Health Organization. ICD-11 Clinical Descriptions and Diagnostic Guidelines for Mental and Behavioural Disorders. Geneva: World Health Organization; 2019.

3 Hopwood CJ, Malone JC, Ansell EB, Sanislow CA, Grilo CM, McGlashan TH, et al. Personality assessment in DSM-5: empirical support for rating severity, style, and traits. J Pers Disord. 2011 Jun;25(3):305–20.

4 Buer Christensen T, Eikenaes I, Hummelen B, Pedersen G, Nysæter TE, Bender DS, et al. Level of personality functioning as a predictor of psychosocial functioning: concurrent validity of criterion A. Pers Disord. 2020 Mar;11(2):79–90.

5 Hopwood CJ. A framework for treating DSM-5 alternative model for personality disorder features. Pers Ment Health. 2018 May;12(2):107–25.

6 Skodol AE, Morey LC, Bender DS, Oldham JM. The alternative DSM-5 model for personality disorders: a clinical application. Am J Psychiatry. 2015 Jul;172(7):606–13.
7 Bach B, Markon K, Simonsen E, Krueger RF. Clinical utility of the DSM-5 alternative model of personality disorders: six cases from practice. J Psychiatr Pract. 2015 Jan;21(1):3–25.
8 Bateman AW, Gunderson J, Mulder R. Treatment of personality disorder. Lancet. 2015 Feb;385(9969):735–43.
9 Sharp C, Wright AG, Fowler JC, Frueh BC, Allen JG, Oldham J, et al. The structure of personality pathology: both general ('g') and specific ('s') factors? J Abnorm Psychol. 2015 May;124(2):387–98.
10 Wright AG, Hopwood CJ, Skodol AE, Morey LC. Longitudinal validation of general and specific structural features of personality pathology. J Abnorm Psychol. 2016 Nov;125(8):1120–34.
11 Crawford MJ, Koldobsky N, Mulder R, Tyrer P. Classifying personality disorder according to severity. J Pers Disord. 2011 Jun;25(3):321–30.
12 Hengartner MP, Zimmermann J, Wright AG. Personality pathology. In: Zeigler-Hill V, Shackelford T, editors. The SAGE handbook of personality and individual differences. London: Sage; 2018. Vol III: Applications of personality and individual differences. p. 3–35. https://doi.org/10.4135/9781526451248.n1.
13 Clark LA, Nuzum H, Ro E. Manifestations of personality impairment severity: comorbidity, course/prognosis, psychosocial dysfunction, and 'borderline' personality features. Curr Opin Psychol. 2018 Jun;21:117–21.
14 Kernberg OF. Severe personality disorders. New Haven (CT): Yale University Press; 1984.
15 OPD Task Force. Operationalized Psychodynamic Diagnosis OPD-2: Manual of diagnosis and treatment planning. Cambridge (MA): Hogrefe & Huber; 2008.
16 Zimmermann J, Ehrenthal JC, Cierpka M, Schauenburg H, Doering S, Benecke C. Assessing the level of structural integration using operationalized psychodynamic diagnosis (OPD): implications for DSM-5. J Pers Assess. 2012;94(5):522–32.
17 Westen D, Gabbard GO, Blagov PS. Back to the future: personality structure as a context for psychopathology. In: Krueger RF, Tackett JL, editors. Personality and psychopathology. New York: Guilford Press; 2006. pp. 335–84.
18 Lingiardi V, McWilliams N, editors. Psychodynamic diagnostic manual: PDM-2. New York, London: The Guilford Press; 2017.
19 Bender DS, Morey LC, Skodol AE. Toward a model for assessing level of personality functioning in DSM-5, part I: a review of theory and methods. J Pers Assess. 2011 Jul;93(4):332–46.
20 Tyrer P, Reed GM, Crawford MJ. Classification, assessment, prevalence, and effect of personality disorder. Lancet. 2015 Feb;385(9969):717–26.
21 Leising D, Scherbaum S, Packmohr P, Zimmermann J. Substance and evaluation in personality disorder diagnoses. J Pers Disord. 2018 Dec;32(6):766–83.
22 Krueger RF, Derringer J, Markon KE, Watson D, Skodol AE. Initial construction of a maladaptive personality trait model and inventory for DSM-5. Psychol Med. 2012 Sep;42(9):1879–90.
23 Zimmermann J, Kerber A, Rek K, Hopwood CJ, Krueger RF. A brief but comprehensive review of research on the alternative DSM-5 model for personality disorders. Curr Psychiatry Rep. 2019 Aug;21(9):92.
24 Ehrenthal JC, Dinger U, Horsch L, Komo-Lang M, Klinkerfuss M, Grande T, et al. Der OPD-Strukturfragebogen (OPD-SF): Erste Ergebnisse zu Reliabilität und Validität. Psychother Psychosom Med Psychol. 2012 Jan;62(1):25–32.
25 Ehrenthal JC, Dinger U, Schauenburg H, Horsch L, Dahlbender RW, Gierk B. Entwicklung einer Zwölf-Item-Version des OPD-Strukturfragebogens (OPD-SFK). Z Psychosom Med Psychother. 2015;61(3):262–74.
26 Lenzenweger MF, Clarkin JF, Kernberg OF, Foelsch PA. The Inventory of Personality Organization: psychometric properties, factorial composition, and criterion relations with affect, aggressive dyscontrol, psychosis proneness, and self-domains in a nonclinical sample. Psychol Assess. 2001 Dec;13(4):577–91.
27 Hörz-Sagstetter S, Volkert J, Rentrop M, Benecke C, Gremaud-Heitz DJ, Unterrainer HF, et al. A bifactor model of personality organization. J Pers Assess. Epub 2020 Jan.
28 Weekers LC, Hutsebaut J, Kamphuis JH. The Level of Personality Functioning Scale-Brief Form 2.0: update of a brief instrument for assessing level of personality functioning. Pers Ment Health. 2019 Feb;13(1):3–14.
29 Morey LC. Development and initial evaluation of a self-report form of the DSM-5 Level of Personality Functioning Scale. Psychol Assess. 2017 Oct;29(10):1302–8.
30 Olajide K, Munjiza J, Moran P, O'Connell L, Newton-Howes G, Bassett P, et al. Development and psychometric properties of the Standardized Assessment of Severity of Personality Disorder (SASPD). J Pers Disord. 2018 Feb;32(1):44–56.
31 Waugh MH, McClain CM, Mariotti EC, Mulay AL, DeVore EN, Lenger KA, et al. Comparative content analysis of self-report scales for level of personality functioning. J Pers Assess. Epub 2020 Jan.
32 Fischer HF, Rose M. www.common-metrics.org: a web application to estimate scores from different patient-reported outcome measures on a common scale. BMC Med Res Methodol. 2016 Oct;16(1):142. Available from: www.common-metrics.org
33 Wahl I, Löwe B, Bjorner JB, Fischer F, Langs G, Voderholzer U, et al. Standardization of depression measurement: a common metric was developed for 11 self-report depression measures. J Clin Epidemiol. 2014 Jan;67(1):73–86.
34 Batterham PJ, Sunderland M, Slade T, Calear AL, Carragher N. Assessing distress in the community: psychometric properties and crosswalk comparison of eight measures of psychological distress. Psychol Med. 2018 Jun;48(8):1316–24.
35 Oude Voshaar MA, Vonkeman HE, Courvoisier D, Finckh A, Gossec L, Leung YY, et al. Towards standardized patient reported physical function outcome reporting: linking ten commonly used questionnaires to a common metric. Qual Life Res. 2019 Jan;28(1):187–97.
36 Friedrich M, Hinz A, Kuhnt S, Schulte T, Rose M, Fischer F. Measuring fatigue in cancer patients: a common metric for six fatigue instruments. Qual Life Res. 2019 Jun;28(6):1615–26.
37 Schalet BD, Revicki DA, Cook KF, Krishnan E, Fries JF, Cella D. Establishing a common metric for physical function: linking the HAQ-DI and SF-36 PF subscale to PROMIS® physical function. J Gen Intern Med. 2015 Oct;30(10):1517–23.
38 Choi SW, Schalet B, Cook KF, Cella D. Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. Psychol Assess. 2014 Jun;26(2):513–27.
39 Liegl G, Wahl I, Berghöfer A, Nolte S, Pieh C, Rose M, et al. Using Patient Health Questionnaire-9 item parameters of a common metric resulted in similar depression scores compared to independent item response theory model reestimation. J Clin Epidemiol. 2016 Mar;71:25–34.
40 Sunderland M, Batterham P, Calear A, Carragher N. Validity of the PROMIS depression and anxiety common metrics in an online sample of Australian adults. Qual Life Res. 2018 Sep;27(9):2453–8.
41 Petrowski K, Schmalbach B, Kliem S, Hinz A, Brähler E. Symptom-Checklist-K-9: norm values and factorial structure in a representative German sample. PLoS One. 2019 Apr;14(4):e0213490.
42 Larisch A, Heuft G, Engbrink S, Brähler E, Herzog W, Kruse J. Behandlung psychischer und psychosomatischer Beschwerden – Inanspruchnahme, Erwartungen und Kenntnisse der Allgemeinbevölkerung in Deutschland. Z Psychosom Med Psychother. 2013;59(2):153–69.
43 Zimmermann J, Benecke C, Hörz S, Rentrop M, Peham D, Bock A, et al. Validierung einer deutschsprachigen 16-Item-Version des Inventars der Persönlichkeitsorganisation (IPO-16). Diagnostica. 2013;59(1):3–16.
44 Zimmermann J, Benecke C, Hörz-Sagstetter S, Dammann G. Normierung der deutschsprachigen 16-Item-Version des Inventars der Persönlichkeitsorganisation (IPO-16). Z Psychosom Med Psychother. 2015;61(1):5–18.
45 Hutsebaut J, Feenstra DJ, Kamphuis JH. Development and preliminary psychometric evaluation of a brief self-report questionnaire for the assessment of the DSM-5 Level of Personality Functioning Scale: The LPFS Brief Form (LPFS-BF). Pers Disord. 2016 Apr;7(2):192–7.

46 Obbarius A, Obbarius N, Fischer F, Liegl G, Rose M. Evaluation der Faktorenstruktur und Konstruktvalidität der 12-Item-Kurzversion des OPD-Strukturfragebogens (OPD-SFK) an psychosomatischen Patienten. Psychother Psychosom Med Psychol. 2019 Jan;69(1):38–48.
47 Kerber A, Schultze M, Müller S, Rühling RM, Wright AGC, Spitzer C, et al. Development of a short and ICD-11 compatible measure for DSM-5 maladaptive personality traits using ant colony optimization algorithms. Submitted for publication. 2019. DOI: 10.31234/osf.io/rsw54.
48 Moran P, Leese M, Lee T, Walters P, Thornicroft G, Mann A. Standardised Assessment of Personality - Abbreviated Scale (SAPAS): preliminary validation of a brief screen for personality disorder. Br J Psychiatry. 2003 Sep;183(3):228–32.
49 Derogatis LR. Symptom Checklist 90, R-Version manual I: scoring, administration, and procedures for the SCL-90. Baltimore (MD): Johns Hopkins Press; 1977.
50 Chalmers RP. mirt: a multidimensional item response theory package for the R environment. J Stat Softw. 2012;48(6):1–29.
51 R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.
52 Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monogr Suppl. 1969;34 S1:100–14.
53 Cai L, Hansen M. Limited-information goodness-of-fit testing of hierarchical item factor models. Br J Math Stat Psychol. 2013 May;66(2):245–76.
54 Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling. 1999;6(1):1–55.
55 Reise SP, Scheines R, Widaman KF, Haviland MG. Multidimensionality and structural coefficient bias in structural equation modeling. Educ Psychol Meas. 2013;73(1):5–26.
56 Thissen D, Pommerich M, Billeaud K, Williams VS. Item response theory for scores on tests including polytomous items with ordered responses. Appl Psychol Meas. 1995;19(1):39–49.
57 Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet. 1995 Oct;346(8982):1085–7.
58 Sleep CE, Lynam DR, Widiger TA, Crowe ML, Miller JD. An evaluation of DSM-5 section III personality disorder criterion A (impairment) in accounting for psychopathology. Psychol Assess. 2019 Oct;31(10):1181–91.
59 Bach B, Anderson JL. Patient-reported ICD-11 personality disorder severity and DSM-5 level of personality functioning. J Pers Disord. Epub 2018 Sep.
60 Rek K, Thielmann I, Henkel M, Crawford M, Piccirilli L, Graff A, et al. A psychometric evaluation of the Standardized Assessment of Severity of Personality Disorder (SASPD) in clinical and non-clinical German samples. Submitted for publication. 2019. DOI: 10.31234/osf.io/unqhm.
61 Ebert DD, Buntrock C, Reins JA, Zimmermann J, Cuijpers P. Efficacy and moderators of psychological interventions in treating subclinical symptoms of depression and preventing major depressive disorder onsets: protocol for an individual patient data meta-analysis of randomised controlled trials. BMJ Open. 2018 Mar;8(3):e018582.
62 Morey LC, Good EW, Hopwood CJ. Global personality dysfunction and the relationship of pathological and normal trait domains in the DSM-5 Alternative Model for Personality Disorders. Submitted for publication. 2020.
63 Schroeders U, Wilhelm O, Olaru G. Meta-heuristics in short scale construction: ant colony optimization and genetic algorithm. PLoS One. 2016 Nov;11(11):e0167110.
64 Zimmermann J, Böhnke JR, Eschstruth R, Mathews A, Wenzel K, Leising D. The latent structure of personality functioning: investigating criterion A from the alternative model for personality disorders in DSM-5. J Abnorm Psychol. 2015 Aug;124(3):532–48.
65 Wendt LP, Wright AG, Pilkonis PA, Nolte T, Fonagy P, Montague PR, et al. The latent structure of interpersonal problems: validity of dimensional, categorical, and hybrid models. J Abnorm Psychol. 2019 Nov;128(8):823–39.
66 Ophir Y, Sisso I, Asterhan CS, Tikochinski R, Reichart R. The turker blues: hidden factors behind increased depression rates among Amazon's Mechanical Turkers. Clin Psychol Sci. 2020;8(1):65–83.
67 Shapiro DN, Chandler J, Mueller PA. Using Mechanical Turk to study clinical populations. Clin Psychol Sci. 2013;1(2):213–20.
68 O'Connor BP. The search for dimensional structure differences between normality and abnormality: a statistical review of published data on personality and psychopathology. J Pers Soc Psychol. 2002 Oct;83(4):962–82.
69 Bach B, Sellbom M, Simonsen E. Personality inventory for DSM-5 (PID-5) in clinical versus nonclinical individuals: generalizability of psychometric features. Assessment. 2018 Oct;25(7):815–25.
70 Verheul R, Andrea H, Berghout CC, Dolan C, Busschbach JJ, van der Kroft PJ, et al. Severity Indices of Personality Problems (SIPP-118): development, factor structure, reliability, and validity. Psychol Assess. 2008 Mar;20(1):23–34.
71 Livesley WJ. General Assessment of Personality Disorder (GAPD). Vancouver (BC): Department of Psychiatry, University of British Columbia; 2006.
72 Huprich SK, Nelson SM, Meehan KB, Siefert CJ, Haggerty G, Sexton J, et al. Introduction of the DSM-5 levels of Personality Functioning Questionnaire. Pers Disord. 2018 Nov;9(6):553–63.
73 Guenole N, Brown AA, Cooper AJ. Forced-choice assessment of work-related maladaptive personality traits: preliminary evidence from an application of Thurstonian item response modeling. Assessment. 2018 Jun;25(4):513–26.
74 Westen D, Weinberger J. When clinical description becomes statistical prediction. Am Psychol. 2004 Oct;59(7):595–613.
75 Oltmanns TF, Turkheimer E. Person perception and personality pathology. Curr Dir Psychol Sci. 2009 Jan;18(1):32–6.
76 Bornstein RF. Toward a process-focused model of test score validity: improving psychological assessment in science and practice. Psychol Assess. 2011 Jun;23(2):532–44.