Psychological Medicine (2012), **42**, 657–667. © Cambridge University Press 2011
doi:10.1017/S0033291711001632

ORIGINAL ARTICLE

Concordance between personality disorder assessment methods

G. Nestadt[1]*, C. Di[2], J. F. Samuels[1], Y.-J. Cheng[3], O. J. Bienvenu[1], I. M. Reti[1], P. Costa[4], W. W. Eaton[5] and K. Bandeen-Roche[6]

[1] *Department of Psychiatry and Behavioral Sciences, The Johns Hopkins University School of Medicine, Baltimore, MD, USA*
[2] *Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA*
[3] *Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan*
[4] *Gerontology Research Center, National Institute on Aging, Baltimore, MD, USA*
[5] *Department of Mental Hygiene, The Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA*
[6] *Department of Biostatistics, The Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA*

Background. Studies have criticized the low level of agreement between the various methods of personality disorder (PD) assessment. This is an important issue for research and clinical purposes.

Method. Seven hundred and forty-two participants in the Hopkins Epidemiology of Personality Disorders Study (HEPS) were assessed on two occasions using the Personality Disorder Schedule (PDS) and the International Personality Disorder Examination (IPDE). The concordance between the two diagnostic methods for all DSM-IV PDs was assessed using standard methods and also two item response analytic approaches designed to take account of measurement error: a latent trait-based approach and a generalized estimating equations (GEE)-based approach, with *post-hoc* adjustment.

Results. Raw criteria counts, using the intraclass correlation coefficient (ICC), κ and odds ratio (OR), showed poor concordance. The more refined statistical methods showed a moderate to moderately high level of concordance between the methods for most PDs studied. Overall, the PDS produced lower prevalences of traits but higher precision of measurement than the IPDE. Specific criteria within each PD showed varying endorsement thresholds and precision for ascertaining the disorder.

Conclusions. Concordance in the raw measurement of the individual PD criteria between the two clinical methods is lacking. However, based on two statistical methods that adjust for differential endorsement thresholds and measurement error in the assessments, we deduce that the PD constructs themselves can be measured with a moderate degree of confidence regardless of the clinical approach used. This may suggest that the individual criteria for each PD are, in and of themselves, less specific for diagnosis, but as a group the criteria for each PD usefully identify specific PD constructs.

Received 3 November 2010; Revised 29 July 2011; Accepted 29 July 2011; First published online 24 August 2011

Key words: Concordance (measurement), personality, personality disorders.

Introduction

A major methodological issue in studying the epidemiology of personality disorders (PDs) is choice of assessment approach. There are two major clinical approaches for evaluating these disorders in the community. One approach is to evaluate PD criteria in the context of a systematic clinical examination; this is the typical clinical method of patient evaluation extended to include the comprehensive list of all PD criteria, each of which are defined and required to be assessed. The alternative approach is to use a semi-structured interview instrument that poses specific questions that target each PD trait (criterion) to be rated; this is the typical method in research studies. There is also a questionnaire approach (self-report), which is not discussed in this paper.

In a review of comparability studies, Perry (1992) reported poor concordance between direct-interview methods for diagnosing PDs in patients. He suggested that methods differ in their ability to adequately distinguish between long-standing patterns of affect, behavior and cognition that are characteristic of PDs, as opposed to sporadic occurrences of these traits. However, little is known about the agreement between

* Address for correspondence: Dr G. Nestadt, Department of Psychiatry and Behavioral Sciences, Johns Hopkins Hospital, Meyer 113, 600 N. Wolfe Street, Baltimore, MD 21287, USA.
(Email: gnestadt@jhmi.edu)

https://doi.org/10.1017/S0033291711001632 Published online by Cambridge University Press

different interview-based methods for evaluating PDs in the community. Therefore, in the current study, we evaluated the concordance between the above-mentioned two clinical assessment methods for diagnosing PDs in the community using item response analytic approaches designed to account for measurement error.

Method

Participants

As described previously (Samuels *et al*. 2002), individuals participating in the Hopkins Epidemiology of Personality Disorder Study (HEPS) were sampled from the Baltimore Epidemiologic Catchment Area Follow-up (Eaton *et al*. 1998). In brief, 3481 adult household residents of East Baltimore were sampled probabilistically in 1981 and interviewed using the Diagnostic Interview Schedule (DIS); 810 of these individuals were also examined by psychiatrists in the Clinical Reappraisal (Anthony *et al*. 1985) in the same year. Between 1993 and 1996, 1920 (73%; 1920/3481) of the available participants were reinterviewed (many had died, as the original cohort was oversampled for elderly participants).

A stratified random sample of 1258 participants (of the aforementioned 1920 participants) were selected to participate in the HEPS, which was conducted between 1997 and 1999. These participants included: all available participants examined by psychiatrists in the Clinical Reappraisal in 1981; all participants identified by the DIS as having a lifetime diagnosis of mania, depression, panic disorder, obsessive–compulsive disorder, alcohol use disorders, or drug use disorders at follow-up, between 1993 and 1996; and a 25% random sample of the remaining participants who did not meet either of the two previous criteria ($n=222$). Of the 1258 subjects selected with these criteria, 516 could not be interviewed because they could not be traced, refused participation, had died, or were too ill or elderly to participate. A total of 742 subjects were included in the study, of whom 713 completed both personality assessments.

The gender and ethnic distributions of these 713 individuals were similar to those of the 516 who were not interviewed; however, the interviewed subjects were younger (mean age 47 years) than the non-interviewed subjects (mean age 61 years). The gender and ethnic distributions of the participants were similar to those of the 3481 subjects examined in 1981, although the study subjects were younger.

All participants provided signed, informed consent and the study was approved by the Institutional Review Board (IRB) at Johns Hopkins University.

Personality assessment

A total of 713 participants completed two independent, direct, in-person examinations, using the two above-mentioned methods for the assessment of PDs. One method involved a thorough examination of psychiatric symptoms and PD criteria conducted by psychiatrists, using the Personality Disorder Schedule (PDS) of the Standardized Psychiatric Examination (SPE) to assess DSM-IV PDs (Romanoski *et al*. 1998). Primarily an unstructured but systematic clinical method, the assessment of PD criteria relies on information from direct questioning about specific personality criteria, in addition to information emerging during the participant–examiner interaction. These examinations were conducted by five board-eligible psychiatrists. The psychiatrists collected biographical information including schooling, employment and relationship history, with an emphasis on interpersonal relationships, as an ingredient for PD assessment. The assessment is designed to ensure that these domains are completed in the examination. The aim was to identify emotional responses across various situations and circumstances. A glossary definition for each PD criterion was provided and, for certain criteria, questions to use as diagnostic prompts were provided. Each PD criterion was rated '0' (absent), '1' (accentuated or exaggerated), '2' (severe) or '9' (missing). The psychiatrists did not make PD diagnoses; these were computed based on the criteria endorsed using a DSM- IV algorithm.

The other method was an interview conducted by clinical psychologists, using the International Personality Disorder Examination (IPDE; Loranger *et al*. 1994), a semi-structured method that asks participants directly about each specific personality criterion, and queries informants about a subset of the personality criteria. Four masters-level clinical psychologists conducted these interviews. The psychologists evaluated PD criteria manifest over the participant's entire adult life. For each criterion, the definition was provided, in addition to one or more questions to be used as diagnostic prompts; the psychologists were encouraged to ask clarifying questions. Criteria were rated '0', '1', '2' or '9', as defined for the PDS. Following the interview, the psychologists interviewed at least one informant by telephone, using questions from the IPDE. The psychologists were required to evaluate more than half of the criteria for each of the PDs. Discrepancies between informants and participant ratings were resolved by the psychologists' judgment on the evidence available for each criterion when recording a final score. In 40 interviews of participants rated jointly by the psychologists, the intraclass correlation coefficients (ICCs) for the

https://doi.org/10.1017/S0033291711001632 Published online by Cambridge University Press

number of PD criteria rated as present were schizoid (0.81), schizotypal (0.58), paranoid (0.63), antisocial (0.80), borderline (0.76), histrionic (0.62), narcissistic (0.62), avoidant (0.89), dependent (0.76), and obsessive–compulsive (0.70) (Samuels *et al*. 1994).

The PDS and IPDE were conducted independently, and the psychiatrists and psychologists were blind to each others' assessments. Both methods rate each of the DSM-IV PD criteria and allow diagnoses for each of the 10 DSM-IV PDs (APA, 2000).

Statistical approach

The analytic approach evaluating concordance is based on, and parallels, that described in detail in Nestadt *et al*. (2010). Conceptually, we construe PD to be dimensional; that is, envision a given disorder as a manifestation of an underlying severity ranging along a continuum from low (no disorder) to high. The IPDE or PDS uses a set of criteria that are endorsed if the underlying severity exceeds a threshold level. We consider such thresholds to be potentially person specific, and hence subject to random uncertainty. The effect of such uncertainty is to attenuate associations and hence standard measures of concordance; the item response approach we propose here explicitly addresses and attempts to correct for such attenuation. Thresholds triggering criterion endorsement are almost certainly criterion and instrument specific. Both analytic approaches proposed here use measures of association, and hence concordance, that are robust to this sort of specificity. Finally, instruments may target fundamentally different constructs of disorder, so we conceptualize parallel severity scores for individuals, one for the IPDE construct and one for the PDS. We make concordance operational as the correlation between the two available representations of a person's disorder severity, one inferred through the IPDE and the other inferred through the PDS. Concordance as so defined should be high if the two representations rank individuals similarly as to their disorder severity.

In addition to the framework described above, there were also practical considerations. The two assessment methods differed considerably with respect to the distributions of the number of endorsed criteria. This suggested that there were systematic differences between the levels of severity that 'triggered' a positive endorsement of criteria. Moreover, the criteria count distributions were extremely skewed with a preponderance of 0 counts. These features may invalidate the use of standard analyses of concordance between criteria; the approach we now outline accommodates both of these practical complexities.

Our analytic approach was grounded in item response analysis (IRA; Lord, 1980); to optimize our confidence in its findings, we implemented parallel versions of it. Take Y_{ijk} to represent the presence (0 or 1) of the kth criterion ($k=1, \ldots, K$) using the jth method ($j=1$ for IPDE, $j=2$ for PDS) for some ith person, for a given disorder. Let d_{ij} be the associated severities of disorder for IPDE ($j=1$) and PDS ($j=2$) constructs; these are envisioned as z scores (normally distributed variables with mean 0 and variance 1). The model relating criteria to disorder severity is a random effects item response theory (IRT) model:

$$\log \text{odds}(Y_{ijk}=1)=\gamma_j+\beta_{jk}+\lambda_{jk}d_{ij}. \quad (1)$$

In IRT terminology, $-(\gamma_j+\beta_{jk})$ gives the item k 'difficulty' parameter for each method j, λ_{jk} is the discrimination parameter, and d_{ij} is the 'ability'. In our context $-(\gamma_j+\beta_{jk})$ distinguishes criteria by their severity, such that higher 'difficulty' identifies a more rarely exhibited criterion. In our context λ_{jk} distinguishes criteria by how precisely they measure disorder severity. In the standard diagnostic terminology of sensitivity and specificity, higher 'discrimination' identifies a criterion whose reporting is more sensitive and specific to the underlying disorder among two criteria of equal difficulty, lower difficulty identifies a criterion that is more sensitive among two equally discriminating criteria, and higher difficulty identifies a criterion that is more specific among two equally discriminating criteria. Our statement of (1) separates the criterion difficulty into a disorder-wide difficulty ($-\gamma_j$) and terms differentiating difficulty of each criterion from the disorder-wide difficulty ($-\beta_{jk}$). This interpretation is achieved by constraining $\beta_{j1}+\ldots+\beta_{jK}=0$, as in classical ANOVA. The resultant measure of concordance in an individual's assessments is the correlation between a person's two representations of disorder severity as inferred from the IPDE and PDS, d_{i1} and d_{i2}.

Model (1) describes systematic differences in the assessment of disorder severity between the methods through $\gamma_2-\gamma_1$, $\beta_{2k}-\beta_{1k}$ and $\lambda_{2k}-\lambda_{1k}$. We shall say that a method with higher γ, or a criterion with a higher β, tends to elicit disorder at a lower severity threshold than that with a lower such value, and a criterion with a higher λ has a higher precision for detecting the disorder. $\text{Exp}(\gamma_j)$ can be interpreted as the average odds of exhibiting disorder for a person with mean underlying disorder severity, where the (geometric) average is taken over all the criteria defining the disorder. Similarly, $\exp(\beta_{jk})$ gives the factor by which the odds of exhibiting the kth trait is multiplied compared to the average, that is by which the kth criterion is relatively more or less frequently exhibited among the criteria defining the disorder.

Findings from an analysis based on equation (1) are potentially influenced by the assumption of normally

https://doi.org/10.1017/S0033291711001632 Published online by Cambridge University Press

distributed severity scores. To assure robustness of our findings to this assumption, we implemented a second item response model focused entirely on describing criterion prevalence, without explicitly building in an underlying disorder severity:

$$\log \text{odds}(Y_{ijk}=1)=\gamma_j+\beta_{jk}. \quad (2)$$

Here, γ_j and β_{jk} once again denote disorder and criterion difficulties using method j, except that they represent population-wide averages rather than values for persons with the mean disorder severity. This second approach is a generalized estimating equations (GEE) model (Zeger & Liang, 1986), in which the concordance in an individual's assessments is estimated by a summary of the pairwise associations between a person's criterion ratings from the different methods (see equation 3 and associated text in Nestadt *et al*. 2010), across all possible pairs of criteria representing the disorder. Similarly as with the criterion prevalence models (1) and (2), we decomposed this measure of concordance into an overall measure (averaged over traits, within disorder) and amounts by which concordance for different traits differs from the disorder-wide average. Importantly, concordance estimated by the approach just described is for measured disorder, not accounting for differences in the measurement error with which traits assess the underlying disorder. By contrast, model (1) estimates across-method concordance in assessing underlying disorder, correcting for random error in criterion measurement. We previously developed a method to estimate underlying disorder concordance using the approach described in this paragraph (see equations 4–7 in Nestadt *et al*. 2010); we use the same method here.

Statistical analyses

All analyses were conducted for each of the 10 DSM-IV PDs, individually by disorder. Criteria ratings of '2' were rare; therefore we counted a criterion as 'present' if rated at level '1' or higher and absent otherwise. To permit comparability to standard analyses, counts of present criteria were cross-tabulated by the two assessment methods (IPDE and PDS) and ICCs between the IPDE and PDS counts were computed. κ agreement statistics were computed after merging criteria counts into categories of 0, 1–2 and 3. Pairwise odds ratios (ORs) were used to assess the association between a criterion assessed by one method and a criterion assessed using the other method for all possible criteria pairs (same and different); these assess the relative likelihood of the criterion assessed by PDS being present if the trait assessed by IPDE was, or was not, present. For several criteria spanning several disorders, no participants reported the criterion in either assessment, and these were eliminated from further analysis.

Table 1. *Demographic characteristics of participants examined with both personality disorder (PD) assessment methods (PDS and IPDE; n = 713)*

Sex	
Men	262 (37)
Women	451 (63)
Race/ethnicity	
White	427 (60)
Non-white	286 (40)
Age (years)	
34–44	256 (36)
45–54	237 (33)
55–64	102 (14)
65–94	118 (17)
Marital status	
Married	361 (51)
Widowed	80 (11)
Separated/divorced	178 (25)
Never married	94 (13)
Highest grade completed	
<9th grade	70 (10)
9th–11th grade	171 (24)
12th grade	203 (29)
>12th grade	264 (37)
Income (household, annual)	
<US$20 000	214 (35)
US$20 000–49 999	237 (38)
⩾US$50 000	168 (27)

PDS, Personality Disorder Schedule; IPDE, International Personality Disorder Examination.
Values given as *n* (%).

Latent trait analyses corresponding to equation (1) were fit using software written in Winbugs (www.mrc-bsu.cam.ac.uk/bugs/) to support the work of Nestadt *et al*. (2010). Analyses corresponding to equation (2) were conducted using GEEs (Huang *et al*. 2002) implemented in the R statistical computing language (www.r-project.org/); one to estimate (2) and a second to estimate the concordance (Heagerty & Zeger, 2000). Each analysis defined criteria as present or absent as described in the previous paragraph. For three disorders the algorithm for fit of the latent trait analyses was aborted by the software absent a successful fit, indicative of too low a frequency of reporting multiple criteria together to obtain a stable estimate of the latent trait model parameters for association; for these, we report the GEE analyses only.

Results

Demographic characteristics

The demographic characteristics of the 713 participants are shown in Table 1. About 60% of the

https://doi.org/10.1017/S0033291711001632 Published online by Cambridge University Press

Table 2. *Concordance between diagnostic methods for counts of personality disorder (PD) criteria*[a]

	Proportion with any criterion, by method		Concordance statistics		
PD	IPDE	PDS	ICC	κ[b]	OR[c]
Paranoid	0.58	0.20	0.07	0.07	2.2
Schizoid	0.38	0.19	0.23	0.14	3.2
Schizotypal	0.58	0.33	0.22	0.16	2.5
Antisocial	0.63	0.20	0.20	0.13	6.3
Borderline	0.71	0.37	0.10	0.09	2.8
Histrionic	0.42	0.33	0.27	0.17	2.5
Narcissistic	0.42	0.10	0.06	0.09	3.9
Avoidant	0.42	0.12	0.06	0.05	3.4
Dependent	0.36	0.08	0.14	0.08	4.1
Obsessive–compulsive	0.70	0.54	0.25	0.08	1.7

PDS, Personality Disorder Schedule; IPDE, International Personality Disorder Examination; ICC, intraclass correlation coefficient; OR, odds ratio.

[a] Criteria were counted as positive if they were 'present' (rating of 1 or 2).

[b] κ was computed from 3×3 tables, with the number of endorsed criteria categorized into 0, 1–2 or $\geqslant 3$.

[c] Median cross-method pairwise OR.

participants were women, and 60% were white. The participants ranged in age from 34 to 94 years; most were younger than 55 years old, and the mean age was 51 years. Most participants were married; 25% were separated or divorced. About one-third of the participants were high-school graduates, and another third had post-high-school education. The annual household income was less than US$20 000 for approximately one-third of the participants, and more than US$50 000 for another third.

Concordance between methods for counts of PD traits

There was substantial variation between PDs in the degree of concordance of criteria counts as assessed by IPDE and PDS (Table 2). For all disorders, participants were considerably more likely to be rated as having one or more criteria by the IPDE than by the PDS method. ICCs ranged from 0.06 to 0.27, and κ statistics ranged from 0.05 to 0.17; thus concordance between raw trait counts was poor. The best agreement occurred for histrionic personality, and the poorest agreement for avoidant PD. Median (over traits) pairwise ORs for association of trait presence between methods ranged from 1.7 (obsessive–compulsive) to 6.3 (antisocial), the latter indicating a more than sixfold higher odds for rating an antisocial trait as present in the PDS if a participant was rated, rather than if they were not rated, as having the trait in the IPDE. These can be roughly compared to the ICCs using the Digby transformation from ORs to correlation coefficients (Digby, 1983); the corresponding range on a correlation scale is 0.20–0.60. That the ORs indicate a higher degree of concordance than the ICCs can be explained by the fact that the former are invariant to differences in trait prevalence across methods, whereas the latter penalize for such.

Also provided in Table 2 is the proportion with any positive criterion, for each disorder. Taken as a measure of prevalence, this measure construes PD to be dimensional with no obvious cutoff, such that endorsement of any criterion constitutes a 'hit' reflecting the first detectable indication of disorder with the available data. However, in the present paper listing of these proportions mainly aims to convey the frequency of positive responses, as we have defined them, and a comparison of these frequencies across the two instruments. The data available for analysis, for most of the disorders, were not unduly sparse.

Systematic between-method differences in PD prevalence

Table 3 summarizes systematic differences in disorder prevalence as assessed by IPDE and PDS, for each disorder, as analyzed by the GEE method [equation (2)]. In that method, the ORs estimate the factors $\exp(\gamma_2)/\exp(\gamma_1)$ by which the geometric average odds of trait presence across traits were higher or lower using the PDS as compared to the IPDE. For histrionic disorder there was no significant difference between the methods [OR 1.03, 95% confidence interval (CI) 0.85–1.26]. However, for all other PDs, the estimated disorder prevalence was significantly lower using the PDS than the IPDE. ORs ranged from 0.68 for obsessive–compulsive disorder, indicating a 32% decrease in odds of trait presence, to 0.11 for narcissism, indicating a 89% decrease in the odds; the odds for all other PDs were more than 60% lower using the PDS compared to the IPDE. However, within disorders there was considerable heterogeneity in individual traits about this global trend. For example, for obsessive–compulsive disorder, the trend to lower PDS trait prevalence was statistically significantly mitigated, and even reversed, for traits of perfectionism, excessive devotion to work and overconscientiousness, and statistically significantly exacerbated for traits of inability to discard worthless objects, reluctance to delegate and miserliness. Table 3 provides data to identify items that oppose the

https://doi.org/10.1017/S0033291711001632 Published online by Cambridge University Press

https://doi.org/10.1017/S0033291711001632 Published online by Cambridge University Press

Table 3. *Item response analysis (IRA) of specific personality disorder (PD) criteria using a GEE approach; evaluation of systematic differences in prevalence rates between clinical methods*

Disorder	Relative odds of exhibiting criteria: averaged over all criteria[a]	Individual criteria	Percentage reporting criteria		Percentage difference: each criterion compared to the average of all criteria[b]
			IPDE	PDS	
Obsessive–compulsive	0.68 (0.58–0.79)	Preoccupied with details, etc.	19.5	26.5	92.0
		Perfectionism	26.0	39.5	177.6*
		Excessively devoted to work	15.8	24.5	156.2*
		Overconscientious, scrupulous	18.5	21.0	73.1*
		Unable to discard worthless objects	27.4	2.8	−88.8*
		Reluctant to delegate	28.9	13.2	−44.4*
		Miserly	8.1	3.3	−43.9*
		Rigidity and stubbornness	42.8	34.5	4.3
Paranoid	0.29 (0.23–0.36)	Suspects exploiting, harming or deceiving	14.8	7.8	47.5*
		Doubts loyalty or trustworthiness	19.5	5.0	−23.7
		Reluctant to confide	23.8	4.7	−44.1*
		Reads hidden meaning	17.7	3.6	−40.7*
		Bears grudges	33.9	8.3	−39.5*
		Perceives attacks	4.9	9.1	565.5*
		Suspects fidelity of spouse or partner	15.3	2.9	−42.6*
Schizoid	0.41 (0.32–0.53)	No desire for close relationships	12.9	3.4	−29.4*
		Chooses solitary activities	21.2	6.0	−42.4*
		Little sexual interest	9.9	3.6	−14.8
		Pleasure in few activities	4.8	3.4	93.2*
		Lacks close friends or confidants	21.9	14.5	49.3*
		Indifferent to praise or criticism	8.7	2.0	−48.4*
		Emotional coldness or flatness	7.2	6.8	128.7*
Schizotypal	0.38 (0.31–0.46)	Ideas of reference	14.0	7.8	33.4
		Odds beliefs or magical thinking	29.4	10.8	−23.2*
		Unusual perceptual experiences	25.2	5.7	−53.3*
		Odd thinking and speech	9.6	2.6	−34.3*
		Suspiciousness	24.9	10.8	−2.5
		Inappropriate or constricted affect	8.0	3.3	5.9
		Odd behavior or appearance	5.8	2.1	−6.3
		Lack of close friends or confidants	21.9	14.5	60.2*
		Excessive social anxiety	8.8	6.8	100.0*
Antisocial	0.18 (0.14–0.22)	Repeated unlawful behaviors	37.9	10.1	89.5*
		Deceitfulness	15.1	2.0	−35.5*
		Impulsivity or failure to plan ahead	6.5	4.0	197.0*
		Irritability and aggressiveness	32.6	6.3	−32.0*
		Reckless disregard for others' safety	34.5	11.8	25.8*
		Consistent irresponsibility	34.6	8.7	−14.2
		Lack of remorse	28.2	3.4	−20.2

https://doi.org/10.1017/S0033291711001632 Published online by Cambridge University Press

Borderline	0.18 (0.14–0.22)	Frantic efforts to avoid abandonment	9.2	4.0	−23.9
		Unstable, intense interpersonal relationships	25.7	4.0	−32.7*
		Identity disturbance	27.7	1.7	−73.6*
		Impulsivity in more than two areas	44.0	31.3	223.7*
		Recurrent suicidality	19.2	3.7	−6.2
		Affective instability	17.8	6.4	78.7*
		Chronic feelings of emptiness	12.3	1.7	−27.6
		Inappropriate, intense anger	45.8	8.8	−37.2*
Histrionic	1.03 (0.85–1.26)	Uncomfortable if not center of attention	10.7	19.8	106.8*
		Sexually seductive or provocative	6.8	2.3	−68.8*
		Rapidly shifting, shallow emotions	7.8	15.3	101.9*
		Style of speech is impressionistic, lacks detail	11.7	8.6	−31.7*
		Self-dramatization	17.5	20.4	15.5
Narcissistic	0.11 (0.07–0.15)	Grandiose sense of self-importance	14.5	2.6	86.0*
		Preoccupied with fantasies of success, etc.	9.6	0.4	−63.8
		Believes self special or unique	10.2	0.4	−67.0*
		Requires excessive admiration	9.7	3.5	211.7*
		Sense of entitlement	10.8	3.3	154.6*
		Interpersonally exploitative	11.1	2.1	54.6
		Lacks empathy	11.4	1.3	−11.4
		Envious of others	17.5	1.3	−45.6
Avoidant	0.31 (0.23–0.42)	Avoids jobs with interpersonal contact	3.7	2.6	−28.1
		Unwilling to get involved with others	12.1	1.6	−62.3*
		Preoccupied with being criticized or rejected	11.3	5.3	34.7
		Views self as socially inept, inferior	21.7	7.2	−11.0
Dependent	0.24 (0.17–0.34)	Difficulty making decisions without reassurance	5.1	1.3	−37.8*
		Needs others to assume responsibility for life	14.1	2.7	−29.9
		Difficulty expressing disagreement	6.7	1.8	14.2
		Difficulty doing things on own because unconfident	5.6	1.1	−11.7
		Excessive lengths for nurturance and support	14.6	2.8	−26.7
		Uncomfortable or helpless when alone	6.9	3.8	116.9*
		Difficulty making decisions without reassurance	6.6	1.4	−11.3

GEE, Generalized estimating equation; CI, confidence interval; IPDE, International Personality Disorder Examination; PDS, Personality Disorder Schedule.

[a] Log odds averaged and exponentiated. 95% Confidence interval in parentheses.

[b] Factor by which odds of reporting criterion in PDS *versus* IPDE.

* Statistical significance at or below the $\alpha = 0.05$ level.

prevailing distinction between the two methods (as in 'perceives attacks' for paranoid PD) and to characterize systematic differences in the modes of assessment.

Primary concordance findings

Both latent trait and GEE analyses indicated moderate concordance between IPDE and PDS assessments of persons' underlying disorder severities (Table 4). Latent trait-based estimates of correlation between the IPDE and PDS representations of disorder severity ranged from 0.33 to 0.78; those for schizotypal, antisocial and borderline disorder exceeded 0.60. These directly estimate concordance between the underlying severity constructs targeted by the two methods and thus adjust for random, normally distributed measurement error, in addition to systematic error associated with different thresholds for registering the presence or absence of a trait; as a result they assess the degree to which the methods rank persons similarly in their degrees of PD. To a great extent, the corresponding models found the precision of PDS traits for assessing disorder to be higher than IPDE traits (higher λ estimates; data not shown). GEE-based estimates for concordance in traits as assessed by the two methods, assessing similarity of the methods' rankings of persons but not accounting for measurement error, ranged from 0.17 to 0.46. Their counterparts upon applying adjustment for measurement error so as to estimate correlations between the two available representations of underlying disorder increased to the same range as for latent trait analyses. These closely approximated the latent trait estimates of the same quantity, agreeing to within five hundredths for all but two disorders and falling within the 95% CI for the latent trait-based concordance for all disorders where a latent trait estimate was possible.

For each disorder and method, some PD criteria were identified that seemed to reflect the underlying disorder more or less precisely than others (Table 4). For instance in obsessive–compulsive PD, preoccupation with details, perfectionism and reluctance to delegate, as measured with the IPDE, exhibited enhanced precision; whereas inability to discard worthless objects had the least precision; the other criteria, whether measured by the PDS or the IPDE, had an intermediate performance.

Discussion

In this study the level of concordance between two PD clinical assessment methods was investigated in a population-based sample of 713 individuals. In a direct comparison of the methods, using traditional statistical approaches that compare methods (ICC, κ, and OR), the level of concordance is relatively poor. Based upon those analyses we should conclude that there is poor agreement between PD assessment methods, thus limiting their application in research studies and clinical practice. However, we recognize that there are inherent limitations in clinical measurement of personality attributes that do not necessarily invalidate their utility. First, there is no 'gold standard' to calibrate PD measurement. Second, the assessment requires the distillation of clinical judgments over time, circumstance, and with regard to related sets of emotion and behavior; this complicates the assessment beyond that required for the evaluation of a contemporaneous symptom. These aspects of assessment may lead to differing thresholds for criterion endorsement; this is prominently observed in this study.

Substantial differences in the two methods of assessment of trait prevalence indicate substantial systematic differences in the severity thresholds at which IPDE and PDS criteria detect disorder. Overall, PDS-based criteria prevalence was considerably lower than IPDE-based criteria prevalence for all disorders except histrionic PD. This indicates that IPDE-based criteria tend to be endorsed at a lower threshold of disorder severity than PDS-based criteria. However, there was also considerable heterogeneity in the degree to which individual criteria paralleled the overall trend.

These limitations need not be construed as necessarily an absence of validity of the measurement; and it is precisely for this reason that we adopted two IRT methodologies that take these factors into consideration. Both analytic strategies address measurement error; the latent IRT method considers that the manifest measures (scores endorsed in the diagnostic instrument) are approximations to an underlying or latent trait whose measurement represents the 'true' trait; the commonality-adjusted GEE requires fewer analytic assumptions in its procedure to adjust for measurement error. The findings that emerge after applying both statistical methods to the data are encouraging.

The concordance between the PDS and IPDE ranged between 0.32 and 0.78 for the PD scale measures, as estimated using IRT methods. For seven of the 10 PDs tested, the ICC was greater than 0.54. In practice, this means that the methods infer disorder severities that rank individuals moderately to highly similarly on their PD, provided that differences in endorsement thresholds and precision of criteria for detecting disorder are accounted for within and across methods. The PD scales with the lowest concordance estimates were histrionic, obsessive–compulsive and paranoid (and according to the latent IRT, avoidant). This suggests that these PDs are more difficult to assess.

https://doi.org/10.1017/S0033291711001632 Published online by Cambridge University Press

https://doi.org/10.1017/S0033291711001632 Published online by Cambridge University Press

Table 4. *Item response analysis (IRA) of personality disorder (PD) concordance between clinical methods*

Disorder	Estimated concordance (95% CI)			Traits evidencing differential precision (GEE)I = IPDE, S = PDS
	Latent trait	GEE, unadjusted	GEE, *post-hoc* adjustment	
Obsessive–compulsive	0.33 (0.22–0.41)	0.17 (0.1–0.24)	0.32	+ Preoccupied with details, etc. (I) + Perfectionism (I) − Unable to discard worthless objects (I) + Reluctant to delegate (I)
Paranoid	0.41 (0.27 –0.54)	0.24 (0.15–0.33)	0.42	+ Suspects exploiting, harming or deceiving (I) + Doubts loyalty or trustworthiness (I) − Bears grudges (I) − Suspects fidelity of spouse or partner (I)
Schizoid	–	0.37 (0.27–0.47)	0.54	+ No desire for close relationships (I, S) − Little sexual interest (I) + Lacks close friends or confidants (I) + Indifferent to praise or criticism (I)
Schizotypal	0.62 (0.52–0.71)	0.35 (0.28–0.41)	0.58	− Odds beliefs or magical thinking (I, S) + Odd thinking and speech (I, S) − Lack of close friends or confidants (I, S) + Inappropriate or constricted affect (S)
Antisocial	0.78 (0.71–0.86)	0.5 (0.41–0.58)	0.75	+ Repeated unlawful behaviors (I) − Reckless disregard for others' safety (I, S) + Impulsivity or failure to plan ahead (S)
Borderline	0.67 (0.57–0.76)	0.38 (0.31–0.45)	0.66	+ Recurrent suicidality (I, S) + Affective instability (I) + Identity disturbance (S) − Impulsivity in more than two areas (S)
Histrionic	0.55 (0.44–0.66)	0.3 (0.2–0.4)	0.49	− Uncomfortable if not center of attention (I) − Sexually seductive or provocative (I) + Style of speck is impressionistic, lacks detail (I) + Self-dramatization (I, S)
Narcissistic	–	0.37 (0.26–0.47)	0.61	+ Grandiose sense of self-importance (I, S) + Believes self special or unique (I) − Lacks empathy (I)
Avoidant dependent	0.42 (0.24–0.59) –	0.4 (0.24–0.53) 0.46 (0.32–0.58)	0.55 0.67	+ Avoids jobs with interpersonal contact (I) − Excessive lengths for nurturance and support (I) + Difficulty making decisions without reassurance (S)

GEE, Generalized estimating equation; CI, confidence interval; IPDE, International Personality Disorder Examination; PDS, Personality Disorder Schedule.

The findings from this study are mixed. It is evident that the methods cannot be substituted directly, although they do correspond moderately once imprecision and differential endorsement thresholds are accounted for. Moreover, both approaches seem to be tapping the same (or similar) latent PD construct. This suggests that the two assessments are successfully identifying the construct based upon the individual units (criteria) of measurement, but that the individual units themselves are less reliable. It could be concluded that better assessment methods that account for measurement differences are needed. This could be accomplished by the identification of criteria that correspond more uniformly between assessment methods.

Nevertheless, the findings in this study suggest that endorsement, or lack thereof, of individual criteria is not necessarily reliable, but that endorsement of members of the group of criteria underlying each PD provides a viable diagnostic approach.

Limitations

The methods examined in this study are not the only available approaches. Other diagnostic interviews, such as the Structured Interview for DSM-IV Personality (SIDP-IV) or Structured Clinical Interview for DSM-IV Axis II Personality Disorders (SCID-II), may have given different results. The PDS was designed for this study in 1981, as there were no alternative instruments available at that time. Additional items were subsequently added to include changes in the criteria required by DSM-III-R and DSM-IV, although the original items were maintained for consistency. Psychiatrists used the PDS whereas masters-level psychologist conducted the IPDE. Both groups were trained extensively and ratings were reviewed by the respective group to maintain consensus across the study period. The examinations were not conducted simultaneously; however, the majority occurred with a 3-week period of each other. The PDS assessment was conducted as part of a complete diagnostic examination for Axis I psychopathology; this may have led to certain personality traits being construed as symptoms of an Axis I disorder and not endorsed for the PD assessment.

The sample, selected from the community-residing population of eastern Baltimore, was overselected for individuals with Axis I disorders, and therefore the results may differ from a true community survey population. Nevertheless, these participants were not selected from a treated population, and therefore are more representative of the performance of these measures without the influence of 'training' of patients that occurs in a treated sample. This ascertainment strategy identifies fewer individuals with frank PD resulting in a more strenuous test of the instrument performance.

This is an older cohort because of the long follow-up. The prevalence of certain PDs declines with age and the stability of the PDs is lower in early adult life. It is plausible that the findings may have been influenced by age.

This study focused on the estimation of concordance between IPDE and PDS assessment of specific DSM-IV PDs. It makes no attempt to establish that PD as so defined is well discriminated between disorders. Indeed, our prior work indicates that the criteria may be considerably associated across different disorders (Nestadt *et al*. 2006). Moreover, the methodology applied here may well heighten the estimates of association across PD relative to an analysis not endeavoring to correct for measurement error in assessment or differential thresholding by criterion or method. Ultimately, we suspect that the current 10-PD schema lacks strong discriminant validity and a different classification than the 10-PD schema currently used by DSM up to the present may lead to superior diagnoses for the identification and tracking of PD for clinical and epidemiological purposes. This important topic is of sufficient scope to merit a study in its own right.

Implications

The statistical approaches for the comparison of PD data assessed by different raters require careful consideration, given that the aforementioned estimates of concordance and methods applying standard ICC- and κ-based approaches to symptom counts differ greatly. Careful identification of the target of concordance assessment is also required. Concordance of underlying severity has to do with the extent to which inferred rankings of individuals by their severity agree, after accounting for differences in measurement properties; concordance of clinically assessed traits has to do with the extent to which rankings of individuals by their severity agree, as if all symptoms were equally reliably assessed; and the standard analyses measure agreement of symptom reports *per se*. Which of these is the most relevant will vary by situation and potentially will depend on the adjudication of the individual reader.

We interpret the findings of this study to suggest that concordance in the raw measurement of the individual PD criteria between the two clinical methods is lacking. However, based on two statistical methods that adjust for differential endorsement thresholds and measurement error in the assessment, we can deduce that the PD constructs themselves can be measured

https://doi.org/10.1017/S0033291711001632 Published online by Cambridge University Press

with a moderate degree of confidence regardless of the clinical approach used. It may be possible to strengthen this confidence in the constructs by refining the constructs themselves; for example by taking advantage of alternative PD constructs than those defined in the DSM-IV. Several studies have posited four or five PD dimensions that encompass the entire domain (e.g. Nestadt *et al*. 2006). The constructs can also be strengthened by identifying the specific traits that are most useful in the measurement of the PD constructs. The findings of this study suggest that the individual criteria for each PD are, in and of themselves, less specific for diagnosis, but as a group the criteria for each PD usefully identify specific PD constructs.

Acknowledgments

This study received support from National Institutes of Health grants MH50616, MH64543 and DA026652.

Declaration of Interest

None.

References

Anthony JC, Folstein M, Romanoski AJ, Von Korff MR, Nestadt GR, Chahal R, Merchant A, Brown CH, Shapiro S, Kramer M, Gruenberg EM (1985). Comparison of the lay Diagnostic Interview Schedule and a standardized psychiatric diagnosis. Experience in eastern Baltimore. *Archives of General Psychiatry* **42**, 667–675.

APA (2000). *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn (DSM-IV). American Psychiatric Association: Washington, DC.

Digby PGN (1983). Approximating the tetrachoric correlation coefficient. *Biometrics* **39**, 753–757.

Eaton WW, Anthony JC, Romanoski A, Tien A, Gallo J, Cai G, Neufeld K, Schlaepfer T, Laugharne J, Chen LS (1998). Onset and recovery from panic disorder in the Baltimore ECA follow-up. *British Journal of Psychiatry* **173**, 501–507.

Heagerty PJ, Zeger SL (2000). Multivariate continuation ratio models: connections and caveats. *Biometrics* **56**, 719–732.

Huang GH, Bandeen-Roche K, Rubin GS (2002). Building marginal models for multiple ordinal measurements. *Applied Statistics* **51**, 37–57.

Loranger AW, Sartorius N, Andreoli A, Berger P, Buchheim P, Channabasavanna SM, Coid B, Dahl A, Diekstra RFW, Ferguson B, Jacobsberg LB, Mombour W, Pull C, Ono Y, Regier DA (1994). The International Personality Disorder Examination. *Archives of General Psychiatry* **51**, 215–224.

Lord FM (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum: Hillsdale, NJ.

Nestadt G, Di C, Samuels JF, Bienvenu OJ, Reti IM, Costa P, Eaton WW, Bandeen-Roche K (2010). The stability of DSM personality disorders over twelve to eighteen years. *Journal of Psychiatric Research* **44**, 1–7.

Nestadt G, Hsu FC, Samuels JF, Bienvenu OJ, Reti I, Costa Jr. PT, Eaton WW (2006). Latent structure of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition personality disorder criteria. *Comprehensive Psychiatry* **47**, 54–62.

Perry JC (1992). Problems and considerations in the valid assessment of personality disorders. *American Journal of Psychiatry* **149**, 1645–1653.

Romanoski AJ, Nestadt G, Chahal R, Merchant A, Folstein MF, Gruenberg EM, McHugh PR (1998). Interobserver reliability of a 'Standardized Psychiatric Examination' (SPE) for case ascertainment (DSM-III). *Journal of Nervous and Mental Disorders* **176**, 63–71.

Samuels J, Eaton WW, Bienvenu 3rd OJ, Brown CH, Costa Jr. PT, Nestadt G (2002). Prevalence and correlates of personality disorders in a community sample. *British Journal of Psychiatry* **180**, 536–542.

Samuels JF, Nestadt G, Romanoski AJ, Folstein MF, McHugh PR (1994). DSM-III personality disorders in the community. *American Journal of Psychiatry* **151**, 1055–1062.

Zeger SL, Liang KY (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.

https://doi.org/10.1017/S0033291711001632 Published online by Cambridge University Press