Available online at www.sciencedirect.com
SCIENCE DIRECT
Comprehensive Psychiatry 47 (2006) 368–375
Comprehensive PSYCHIATRY
www.elsevier.com/locate/comppsych

The interrater reliability of the Structured Interview for DSM-IV Personality

J. Serrita Jane[a], Jason L. Pagan[b], Eric Turkheimer[c], Edna R. Fiedler[d], Thomas F. Oltmanns[c,*]

[a]Department of Psychiatry, Yale University New Haven, CT 06510, USA
[b]Department of Psychology, Washington University in St Louis, St. Louis, MO 63130, USA
[c]Department of Psychology, University of Virginia, Charlottesville, VA 22903, USA
[d]National Space Biomedical Research Institute, Baylor University School of Medicine, Houston, TX 77289, USA

Abstract

We examined the joint interview interrater reliability of the Structured Interview for *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*, Personality Disorders (SIDP-IV) in 433 non–treatment-seeking military recruits. Reliability was computed for the diagnosis of a specific personality disorder (PD) and for the number of PD criteria present, and computed using a dimensional score. Reliability increased when PDs were computed using dimensional scores rather than categorical scores. Avoidant and dependent PDs demonstrated the highest interrater reliability, whereas schizoid and schizotypal showed the lowest. This large sample allowed us to perform item-level analyses of the SIDP-IV. Interrater reliability for each of the PD criteria was generally more than 0.70, with the notable exception of criteria scored through observation only. Overall, the SIDP-IV demonstrated good reliability in a non–treatment-seeking population.

1. Introduction

Historically, the reliable diagnosis of personality pathology has been hampered by, among other factors, variations in the conceptualization of personality, difficulty in defining the course of Axis II disorders, disagreements over inclusion/exclusion criteria, and lack of systematic research across the field of psychology [1-4]. Most of the diagnostic variability is the result of the idiosyncrasy in diagnostic evaluations and not the patients or changes in their symptoms [5]. Most researchers and clinicians now view the use and examination of semistructured interviews as vital to sound clinical science.

Several prominent interviews based on the Diagnostic and Statistical Manual (DSM) PD criteria have been developed and extensively examined since the publication of the *DSM, Third Edition* (*DSM-III*) [6]: the Personality Disorder Examination ([7]), the Personality Disorder Interview–IV (PDI-IV [8]), the Structured Clinical Interview for *DSM, Revised Third Edition* (*DSM-III-R*) ([9]), and the Structured Interview for *DSM-III-R* (and *DSM, Fourth Edition* [*DSM-IV*]) Personality (SIDP-R [10], SIDP-IV [11]). The creators of these interviews have attempted to control, through standardization, the variance often found in traditional clinical questioning and the methods of rating the presence/severity of personality disorder (PD) symptoms.

In a review of the reliability of semistructured interviews using *DSM-III* and *DSM-III-R* [12] PD criteria, the κ reliability coefficients across 15 studies where 2 interviewers concurrently conducted an interview (joint interview designs) were quite high, with 60% of the κ values 0.70 or higher [13]. In 4 studies assessing short-interval test-retest reliability, mean κ values ranged from 0.35 for obsessive-compulsive PD to 0.77 for antisocial PD. κ values reported in 5 studies using long-interval test-retest designs were all below 0.60 with the exception of antisocial PD, which is primarily diagnosed based on directly observable and unambiguous behaviors.

Relatively little information is available regarding the reliability of semistructured interviews based on *DSM-IV* [14] PD criteria. However, the interrater reliability of previous versions of the structured interviews may be

* Corresponding author. Department of Psychology, Campus Box 1125, One Brookings Dr, St Louis, MO 63130, USA.
E-mail address: tolmann@artsci.wustl.edu (T.F. Oltmanns).

0010-440X/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.comppsych.2006.01.009

generalized to the updated interviews [15]. Reliability assessments for semistructured interviews have typically been conducted by the creators of each instrument. The ability of independent groups to reliably diagnose PDs with these instruments should be assessed because the interrater reliability of highly trained vs newly trained interviewers may vary widely. The present study hopes to add to the few studies that have assessed the reliability of interviews based on *DSM-IV* PD criteria. Specifically, we assessed the interrater reliability of ratings for more than 400 military recruits given the SIDP-IV.

For the purpose of estimating reliability in this study, a second-rater design was used. This design is similar to the joint interview designs described above. In second-rater designs, 1 individual conducts the semistructured interview and rates the patient on each trait. Then, a second interviewer, blind to previous ratings, watches a video recording of the original interview and rerates the same patient.

We are aware of only 1 study using the SIDP, which was conducted by the authors, that used a sample of more than 50 interviewees [16]. Among the 104 nonpatients assessed, primarily by telephone, 7 PDs (including schizotypal, histrionic, antisocial, dependent, obsessive-compulsive, and passive-aggressive PD) were present at levels permitting reliability calculations. The reliability for the presence or absence of any PD using a joint interview design with this moderate sample size was 0.93, whereas κ's ranged from 0.66 for avoidant to 1.00 for antisocial, with a mean of 0.86. These results suggest good reliability of the SIDP.

The present study extends the research described above concerning the reliability of the SIDP in several ways. First, we reported reliability coefficients using 3 types of scores: (1) those based on a categorical diagnosis, (2) symptom/criteria counts, and (3) dimensional/continuous scores for each PD based on summing severity scores across PD traits. Second, the reliability estimates reported here were based on a large sample of more than 400 recruits given the SIDP. This large sample provided the first opportunity to conduct an item-level reliability analysis of the SIDP interview. Finally, interviews were conducted by a group independent of the authors of the SIDP. Finding comparably strong reliability estimates would indicate that people other than the authors could reliably use the SIDP-IV.

2. Method

2.1. Participants

The data for these analyses were collected from 433 (59% male) United States Air Force recruits who voluntarily participated in a larger study of personality pathology. The study was approved by the human subject committees at both the university and Wilford Hall Medical Center (Lackland Air Force Base, San Antonio, TX). Additional details about the study and our assessment procedures have been reported elsewhere [17-19]. All recruits signed informed consent statements before participation. We initially screened 2033 recruits (62% male) for self- and peer-rated personality pathology who were identified and tested in groups that were nearing completion of 6 weeks of basic military training. From the screened sample, we selected the 433 recruits to participate in a semistructured diagnostic interview. Most of the people who participated in the interview were selected on the basis of evidence suggesting that they might display evidence of some type of PD. Roughly one third of the participants were selected because their peers had nominated them as exhibiting pathologic personality traits, one third were selected because they produced high scores on a self-report measure of personality pathology, and the final third were selected randomly from the remaining participants as a control group. Interviewers were kept blind to information regarding scores on all of the screening measures.

Comorbid Axis I disorders were unlikely among this sample of Air Force recruits interviewed for PDs. Screening for mental health problems such as substance abuse, major mood disorders, and psychosis is conducted during enlistment and again as training begins. In addition, these recruits undergo rigorous training that would be difficult for people with severe problems to complete.

2.2. Measures

The SIDP-IV [11] is a semistructured interview designed to assess the diagnostic criteria for the 10 PDs listed in *DSM-IV*. Questions are arranged by themes rather than by disorders (eg, work style, interpersonal relationships, emotions, interests, and activities), and each criterion is rated on a scale from 0 to 3. This less transparent organization provides fewer indications that the interview is designed to assess personality pathology and may also reduce interviewer bias [20]. For each PD, criterion scores were summed, and these summed scores were used as an index of PDs. Interviews typically took between 45 and 90 minutes to administer.

2.3. Procedure

Recruits first provided informed consent and completed the screening procedure. The interviews were conducted on the same day, within a few hours of the initial screening. Twelve interviewers conducted the 433 interviews: 3 doctoral level clinical psychologists and 9 graduate students in clinical psychology. Five of the graduate students had master's level clinical experience. Ten of the interviewers were trained by one of the developers of the SIDP, Nancee Blum, before the start of interviews. Training included watching and rating Ms Blum's videotaped interviews, followed by discussion of how to rate the criteria. In addition, Ms Blum supervised 1 live group interview, which also was followed by a discussion of interview style and of how to rate the criteria. All interviewers also watched a number of videotapes of interviews together before beginning the study to improve reliability through discussion of

Table 1
The number of participants who qualified for 1 or more PD diagnoses based on the SIDP-IV

No. of PD diagnoses	No. of participants diagnosed
0	351
1	52
2	26
3	3
4	1
Total	433

the interviews and ratings. After the study began, 20 videotaped interviews, 2 from each interviewer, were sent to Ms Blum for her ratings and comments on interview style. This was done in an effort to maintain the reliability of our interviewers.

At the request of Air Force administrators, a question pertaining to interest or importance of sexual experiences (schizoid PD item) and questions concerning drug use (antisocial PD item) and sexual orientation (borderline PD item) were not included. These item omissions limited the number of items assessing social interests in schizoid PD, conduct disorder in antisocial PD, and identity disturbance in borderline PD. Items relevant to the optional research categories (depressive, negativistic, and self-defeating personalities) were not asked.

All of the interviews were recorded on videotape, and each was rerated independently by a second judge. Each interviewer was assigned an equal number of recorded interviews from each of the other interviewers to second rate. Second rating interviews also allowed each interviewer to extend their training and knowledge of the SIDP-IV interview by watching the styles and techniques used by other interviewers in the group. When confusion or disagreement arose over the original intent of PD criteria or how they were to be rated, we referred to Widiger et al [8] PDI-IV manual.

Intraclass correlations were computed according to the guidelines presented by Shrout and Fleiss [21]. One-way analyses of variance were run for each diagnosis and each criterion. The analyses of variance yield a between-targets mean square (BMS) and a within-target mean square (WMS) for each diagnosis and criteria. The formula for calculating the intraclass correlation (ICC) is as follows:

$$ICC = \frac{BMS - WMS}{BMS + (k-1)WMS}$$

where k is the number of judges rating each target.

3. Results

Overall, 19% of the participants who were interviewed qualified for at least 1 PD diagnosis on the SIDP. Another 10% of the sample qualified for a probable PD, defined as falling 1 criterion short of the threshold for a diagnosis without qualifying for any other definite diagnosis. The disorder most frequently diagnosed in this sample was obsessive-compulsive PD, and the disorders least often diagnosed were schizoid, schizotypal, histrionic, and dependent PDs. Table 1 provides the number of comorbid PD diagnoses found, and Table 2 lists the frequency of each PD diagnosis and subthreshold diagnosis in the sample. The SIDP-IV does not assess for PD not otherwise specified (PDNOS) because no empirical criteria or agreed upon assessment method currently exists in the *DSM-IV*. However, a recent set of analyses conducted in our laboratory suggested that individuals who meet at least 10 *DSM-IV* PD criteria without meeting criteria for any 1 PD demonstrate comparable distress and impairment as those who meet criteria for at least 1 PD [22]. We applied this standard to the interview data for the 433 Air Force recruits and found that additional 24 recruits (5.5% of the interviewed sample) qualified for a diagnosis of PDNOS.

Reliability values were computed using intraclass correlations, which are equivalent to κ coefficients [23]. Table 2 also provides the interrater reliabilities for each PD using data on the presence or absence of a PD, the number of criteria met (sum of the number of criteria with a score of 2 or 3), and the dimensional scores (sum of scores assigned to each criterion). The κ values using the binary diagnosis variables ranged from −0.01 for schizoid PD to 0.85 for avoidant PD (mean = 0.50). When reliability estimates were calculated using criteria counts, κ values ranged from 0.65

Table 2
Frequency of PD diagnoses and interrater reliabilities

Diagnosis	Definite	Probable	Interrater reliability		
	Freq (%)	Freq (%)	Categorical	No. of criteria	Continuous
Paranoid	12 (2.8)	31 (7.2)	0.57	0.75	0.84
Schizoid	1 (0.6)	9 (2.1)	−0.01	0.77	0.81
Schizotypal	0 (0.0)	8 (1.8)	0.03	0.65	0.79
Antisocial	13 (3.0)	20 (4.6)	0.62	0.79	0.84
Borderline	12 (2.8)	17 (3.9)	0.60	0.79	0.85
Histrionic	4 (0.9)	8 (1.8)	0.55	0.72	0.77
Narcissistic	8 (1.8)	18 (4.2)	0.35	0.77	0.82
Avoidant	17 (3.9)	25 (5.8)	0.85	0.90	0.93
Dependent	4 (0.9)	8 (1.8)	0.84	0.85	0.88
Obsessive-compulsive	43 (9.9)	80 (18.5)	0.55	0.75	0.84

N = 433. Freq indicates number of people with the diagnosis; (%), percentage of people with the diagnosis; Probable, 1 criterion short of diagnosis.

Table 3
Interrater reliability of each of the criteria for PDs assessed by the SIDP

PD criteria	Interrater reliability (ICC)
Paranoid	
1. Suspects that others are exploiting or deceiving him/her	0.51
2. Doubts the loyalty/trustworthiness of friends/associates	0.72
3. Reluctant to confide because info might be used against him/her	0.77
4. Reads demeaning/threatening meanings into benign remarks/events	0.76
5. Persistently bears grudges when insulted/slighted	0.77
6. Believes others are attacking his/her reputation and reacts with anger	0.66
7. Suspects the fidelity of his/her spouse/sexual partner	0.73
Schizoid	
1. Neither desires nor enjoys close relationships, including family	0.75
2. Almost always chooses solitary activities	0.86
3. Has little interest in having sexual experiences with another person	[a]
4. Takes pleasure in few, if any, activities	0.62
5. Lacks close friends or confidants other than first-degree relatives	0.79
6. Appears indifferent to praise or criticism	0.61
7. Shows emotional coldness, detachment, or flattened affectivity[b]	0.42
Schizotypal	
1. Ideas of reference	0.69
2. Odd beliefs/magical thinking influencing behavior	0.76
3. Unusual perceptual experiences, including bodily illusions	0.73
4. Odd thinking and speech[b]	0.22
5. Suspiciousness or paranoid ideation	0.72
6. Inappropriate or constricted affect[b]	0.17
7. Behavior or appearance that is odd, eccentric, or peculiar[b]	0.32
8. Lack of close friends or confidants other than first-degree relatives	0.79
9. Excessive social anxiety associated with paranoid fears	0.46
Antisocial	
1. Repeatedly performing acts that are grounds for arrest	0.76
2. Deceitfulness for personal profit or pleasure	0.71
3. Impulsivity or failure to plan ahead	0.66
4. Repeated physical fights or assaults	0.74
5. Reckless disregard for safety of self or others	0.78
6. Inconsistent work behavior or unable to honor financial obligations	0.73
7. Lack of remorse, after having hurt, mistreated or stolen from another	0.75
8. Evidence of conduct disorder before 15 years old	0.74
Borderline	
1. Frantic efforts to avoid real or imagined abandonment	0.65
2. Has unstable/intense relationships that alternate between loving/hating	0.76

Table 3 (*continued*)

PD criteria	Interrater reliability (ICC)
Borderline	
3. Identity disturbance; persistently unstable self-image or sense of self	0.72
4. Impulsivity in at least 2 areas that are potentially self-damaging	0.54
5. Recurrent suicidal or self-mutilating behavior	0.80
6. Affective instability/marked reactivity of mood	0.74
7. Chronic feelings of emptiness	0.84
8. Inappropriate, intense anger, or difficulty controlling anger	0.69
9. Transient stress-related paranoid ideation/severe dissociative symptoms	0.80
Histrionic	
1. Uncomfortable in situations in which he/she is not the center of attention	0.81
2. Inappropriate sexually seductive/provocative behavior with others	0.69
3. Displays rapidly shifting and shallow expressions of emotion	0.59
4. Consistently uses physical appearance to draw attention to self	0.71
5. Style of speech that is excessively impressionistic and lacking in detail[b]	0.14
6. Shows self-dramatization/theatricality/exaggerated expression of emotion	0.64
7. Is suggestible, ie, easily influenced by others or circumstances	0.70
8. Considers relationships more intimate than they actually are	0.66
Narcissistic	
1. Has a grandiose sense of self-importance	0.55
2. Fantasies of unlimited success/power/ brilliance/beauty/ideal love	0.42
3. Believes he/she is "special" and can only be understood by "special" people	0.62
4. Requires excessive admiration	0.68
5. Has a sense of entitlement	0.69
6. Interpersonally exploitative/takes advantage of others to achieve own ends	0.80
7. Lacks empathy with the feelings and needs of others	0.69
8. Is often envious of others or believes that others are envious of him/her	0.63
9. Shows arrogant haughty behaviors or attitudes	0.71
Avoidant	
1. Avoids occupational activities with others for fear of criticism/rejection	0.72
2. Is unwilling to get involved with people unless certain of being liked	0.92
3. Shows restraint within intimate relationships for fear of being shamed	0.86
4. Is preoccupied with being criticized or rejected in social situations	0.75
5. Inhibited in new interpersonal situations because of feelings of inadequacy	0.74
6. Views self as socially inept, personally unappealing, or inferior to others	0.86
7. Reluctant to engage in new activities because they may prove embarrassing	0.90

(*continued on next page*)

Table 3 (*continued*)

PD criteria	Interrater reliability (ICC)
Dependent	
1. Difficulty making everyday decisions without excessive advice from others	0.71
2. Needs others to assume responsibility for most major areas of his/her life	0.78
3. Difficulty disagreeing with others for fear of loss of approval/rejection	0.78
4. Difficulty doing things on his/her own because of lack of self-confidence	0.84
5. Goes to excessive lengths to obtain nurturance and support from others	0.73
6. Feels helpless when alone for fear of being unable to care for self	0.67
7. Urgently seeks another relationship for care when a relationship ends	0.84
8. Unrealistically preoccupied with fears of being left to care for self	0.80
Obsessive-compulsive	
1. Preoccupied with details, rules, lists, order, organization, or schedules	0.75
2. Shows perfectionism that interferes with task completion	0.75
3. Excessively devoted to work to the exclusion of leisure activities	0.78
4. Inflexible about matters of morality, ethics, or values	0.60
5. Unable to discard worthless objects	0.78
6. Reluctant to delegate unless others submit to his/her way of doing things	0.75
7. Adopts a miserly spending style toward self and others	0.77
8. Shows rigidity and stubbornness	0.85

[a] Items not asked during the interview.
[b] Items rated by observation only.

for schizotypal PD to 0.90 for avoidant PD (mean = 0.77). Reliability estimates were consistently the highest when dimensional scores were used; κ values ranged from 0.77 for histrionic PD to 0.93 for avoidant PD (mean = 0.84) using dimensional scores. Avoidant and dependent PDs generally had the highest interrater reliability values across the 3 types of scores, whereas κ values were lower for schizoid and schizotypal PDs.

Reliability estimates for the individual criteria for each of the PDs assessed by the SIDP-IV are presented in Table 3. Item-level κ coefficients are generally good for each criterion with 64% of the κ values 0.70 or above. Item-level κ values were 0.85 or above for the second schizoid PD item (0.86); the second (0.92), third (0.86), sixth (0.86), and seventh (0.90) avoidant items; and the eighth obsessive-compulsive PD item (0.85). Reliability estimates were especially low for the 5 items rated by observation only. The κ values for these observation items ranged from 0.14 for the fifth histrionic item to 0.42 for the seventh schizoid item, with a mean κ across the 5 items of only 0.25. In addition to the observational items, reliability estimates were below 0.60 for the first paranoid PD item (0.51), the ninth schizotypal PD item (0.46), the fourth borderline PD item (0.54), the third histrionic PD item (0.59), and the first and second narcissistic PD items (0.55 and 0.42, respectively).

4. Discussion

The SIDP has been found to be reliable across different types of samples (for a review, see Ref. [13]), including the non–treatment-seeking military recruits assessed in the present study. Reliability estimates found using categorical diagnostic scores were somewhat low but within the range of those reviewed by Zimmerman [13]. We found that reliability estimates based on dimensional scores were more reliable than those based on categorical diagnosis scores, which are consistent with previous studies [16,24,25]. Raters were also more likely to agree on the number of criteria met for a particular disorder than whether a participant does or does not meet threshold for a PD diagnosis. The gains in reliability were much smaller when comparing reliability estimates based on criteria counts and continuous scores than when comparing estimates based on categorical diagnosis scores and criteria counts. This suggests that acceptable reliability estimates may be obtained using criterion counts without the need to calculate dimensional scores.

Reliability estimates based on categorical diagnoses fell into 3 categories: poor reliability estimates for schizoid (−.01), schizotypal (0.03), and narcissistic (0.35) PDs; moderate reliability estimates for histrionic (0.55), obsessive-compulsive (0.55), paranoid (0.57), borderline (0.60), and antisocial (0.62) PDs; and excellent reliabilities estimates for dependent (0.84) and avoidant (0.85) PDs. κ values based on categorical diagnoses in the present study were lower than those reported in a study of treatment-seeking patients [26], which may be a function of the samples used in each study. The base rate for PDs is higher in patient populations, so more variability in pathologic personality symptoms was probably found in the sample of Arntz et al [26]. Researchers have described this issue as the "base rate problem" of κ: "Kappa is influenced by the illness base rate, such that a few diagnostic disagreements have a more pronounced effect on reliability when the base rate is either very low or high" ([13], p 226). The modest to low prevalence rates among our military recruits would cause disagreements in ratings to have a larger impact on reliability estimates.

We found relatively few recruits meeting full criteria for schizotypal, schizoid, histrionic, and dependent PDs, whereas obsessive-compulsive PD was the most commonly diagnosed PD, with 43 (9.9%) recruits meeting the criteria. This high prevalence rate is not surprising. In studies of *DSM-III* and *DSM-III-R* PD criteria, obsessive-compulsive PD was consistently the most frequently diagnosed PD (see Ref. [27]). Furthermore, many obsessive-compulsive traits are likely to be adaptive in a military setting because basic training includes inspection of such things as how they fold

clothes and make their beds. Perfectionism was encouraged in this setting, so military recruits probably thought that it was acceptable to admit to possessing these traits. Because the SIDP-IV does not assess for PDNOS, we applied an empirically derived threshold of 10 PD criteria for diagnosing this disorder in our sample [22]. This method identified 24 (5.5%) recruits as potentially deserving a diagnosis of PDNOS, which makes PDNOS the second most common PD diagnosis behind obsessive-compulsive PD.

Reliability estimates varied widely across PDs in the present study. What might account for this broad range of κ scores? Those PDs with criteria that are more heavily weighted with objective and easily identifiable behaviors (eg, antisocial PD) are arguably more reliably assessed than those PDs based largely on subjective judgments or observations (eg, schizotypal PD). We also found that the reliability of individual criteria within each disorder varied. Because of the variation in reliability estimates across as well as within PDs, we thought it useful to examine each disorder and its specific criteria to determine which are more reliably assessed. In doing so, the PDI-IV manual was referred to frequently for information regarding the original intent of each *DSM-IV* PD criteria as well as any issues surrounding the assessment of these criteria [7].

4.1. Paranoid personality disorder

The range of κ's for paranoid PD was rather narrow with the exception of "suspects others are exploiting him" ($\kappa = 0.51$), which had a much lower reliability than the other criteria. It may have been difficult to determine accurately whether any suspiciousness present had a valid basis [7]. Corroborating evidence not available to our interviewers may have helped determine the validity of reported suspiciousness and thereby improved assessment of this item.

4.2. Schizoid personality disorder

Reliability estimates for the schizoid PD criteria varied widely. The item "almost always chooses solitary activities" ($\kappa = 0.86$) had the highest κ value. Participants generally knew whether or not they preferred doing things alone and could articulate this clearly. The criterion "shows emotional coldness" ($\kappa = 0.42$) was, by far, the least reliably assessed schizoid item. This is an observational criterion with no behavioral anchors and could have been confused with depressive symptoms. Determining emotional coldness was also subject to rater interpretation, and participants may have appeared differently when observed live or via videotape.

4.3. Schizotypal personality disorder

The κ values for the schizotypal criteria also ranged widely. The criterion most reliably rated was "lacks close friends or confidants other than first-degree relatives" ($\kappa = 0.79$). This criterion is straightforward and requires little conjecture on the part of the rater. The observational criteria for this PD have significantly lower reliability estimates than the rest of the criteria. These include "odd thinking and speech" ($\kappa = 0.22$) and "inappropriate or constricted affect" ($\kappa = 0.17$). As noted above, these observational criteria did not have behavioral anchors to assist in rating them reliably. Rating whether "inappropriate or constricted affect" was present during the interview may have been especially difficult to do reliably because those with obsessive-compulsive, paranoid, schizoid, or depressive traits might appear to have this schizotypal feature [7].

4.4. Antisocial personality disorder

For antisocial PD, the criterion most reliably assessed was "reckless disregard for safety of self or others" ($\kappa = 0.78$), which was ascertained by determining the presence of a specific set of behaviors. The least reliable criterion was "impulsivity or failure to plan ahead" ($\kappa = 0.66$), which, according to Widiger et al [8], is meant to capture the motivation behind the impulsive changes. This criterion does cover a series of impulsive behaviors, and perhaps, the lower reliability was due to rater disagreement on what qualifies for impulsive behavior and the level at which these behaviors must be present to satisfy this criterion.

4.5. Borderline personality disorder

The criteria that performed well were "chronic feelings of emptiness" ($\kappa = 0.84$) and "recurrent suicidal or self-mutilating behavior" ($\kappa = 0.80$). Although it may be difficult for some participants to discern whether or not they feel empty inside, this criterion was typically answered with a "yes" or a "no" followed, when present, by the percentage of time that empty feelings were evident. It is not surprising that an item loading heavily on observable behaviors such as recurrent suicidal threats or self-mutilation was assessed reliably. We did not expect that the criterion assessing identity disturbances would have a reasonably high κ value (0.72), considering its poor performance in the *DSM-III-R* field trials when it was considered for deletion [7]. Perhaps those who joined the military had or developed a clear sense of identity or were able to articulate that they joined the military because they lacked a clear sense of who they were or what they wanted in life.

The borderline criteria that performed more poorly in terms of reliability were "frantic efforts to avoid real or imagined abandonment" ($\kappa = 0.65$) and "impulsivity in at least 2 areas that are potentially self-damaging" ($\kappa = 0.54$). The criterion "frantic efforts to avoid abandonment" does not include suicidal threats or gestures, so it may have led to disagreements on whether the characteristic was actually more indicative of a dependent PD feature. We did not expect that the "impulsivity" criterion would have the lowest κ value among the borderline items. This item demands specific behaviors for its endorsement, which theoretically should make it easier to assess reliably. In this case, it may be that the raters disagreed on such things as how often an individual would have to speed or drive recklessly for it to count as one area of impulsivity or when spending was enough to be considered self-damaging.

4.6. Histrionic personality disorder

Most of the histrionic PD criteria had similar reliabilities with the exception of 2 items. The criterion "uncomfortable in situations in which he/she is not the center of attention" (κ = 0.81) was considerably higher than the others and is considered the most diagnostic symptom of histrionic PD [7]. In our sample, individuals appeared willing to say, without much ambiguity, whether or not they liked being at the center or attention in social situations. The second item that deviated from the group was "style of speech that is excessively impressionistic and lacking in detail" (κ = 0.14). This criterion is assessed through observations made during the interview. Not only does this criterion require a subjective judgment, but raters may also have different impressions of a participant's style of speech.

4.7. Narcissistic personality disorder

Only 1 narcissistic PD criterion performed very well: "interpersonally exploitive/takes advantage of others to achieve own ends" (κ = 0.80). When this feature was present, our participants were usually willing and happy to describe how they got others to do things for them, thereby making endorsement of the criterion an easier one for raters. Two criteria performed poorly: "has a grandiose sense of self-importance" (κ = 0.55) and "fantasies of unlimited success, power, brilliance" (κ = 0.42). The grandiosity criterion was difficult to reliably assess during a 1-hour interview because it required a great deal of subjectivity and lacked consistent behavioral anchors. A therapist may be able to rate this item more reliably after having multiple opportunities to interact with and observe the behaviors of a client. The fantasies criterion was difficult to assess reliably because determining the amount of time spent daydreaming about success, power, and brilliance required for someone to meet this criterion was arbitrary and left up to the rater [7].

4.8. Avoidant personality disorder

The reliability estimates for avoidant criteria were generally good. The criteria that performed the best were "is unwilling to get involved with others unless certain of being liked" (κ = 0.92) and "reluctant to engage in new activities because they might prove embarrassing" (κ = 0.90). These criteria were perhaps more reliably assessed because interviewees were able to provide examples that exemplify how these characteristics affected their lives. The least reliable criteria were "avoids occupational activities with others for fear of rejection or criticism" (κ = 0.72) and "inhibited in new interpersonal situations due to feelings of inadequacy" (κ = 0.74). The former criterion may be difficult to reliably assess because the reasons for not enjoying occupational activities are often nebulous. As such, it is difficult to infer whether the avoidance is because of fear of rejection and criticism or for other reasons. In addition, it is difficult at times to get a clear sense of whether the endorsed avoidance generalizes to most settings and circumstances or is more specific to a particular event in the interviewees' past. The latter criterion may have had lower reliability because it was difficult to differentiate from other avoidant criteria [7].

4.9. Dependent personality disorder

All the criteria for dependent PD had high reliability estimates with the exception of "feels helpless when left alone for fear of being unable to care for oneself" (κ = 0.67). This appears to be another example of a criterion not anchored by any directly observable behaviors where the rater must make a subjective judgment. An interviewer may have difficulty knowing whether a participant is endorsing this item because he/she needs others to fill an empty void (a borderline feature) or because he/she actually feels incapable of taking care of him or herself [7].

4.10. Obsessive-compulsive personality disorder

Most of the criteria for obsessive-compulsive PD have similar reliability. The criterion "shows rigidity and stubbornness" (κ = 0.85) had the highest reliability estimate. Participants did not seem to have trouble giving concrete behavioral examples when this feature was present. The criterion "is over conscientious about morality" (κ = 0.60) had the lowest κ among the obsessive-compulsive items. Raters may have had difficulty differentiating between conscientiousness and closed mindedness and confused a conservative value system with being closed-minded [8].

4.11. Summary

With only a few exceptions, the items most reliably assessed were those that had clear behavioral anchors, those that people do not mind disclosing about themselves, and those items that easily elicit obvious examples from interviewees. The criteria rated more poorly were those without clear behavioral anchors, those that were more likely to be influenced by social desirability drives, and those requiring a great deal of insight on the part of the participant and on the part of the rater. The observational criteria, which rely on the interviewer's subjective ratings of behaviors observed during the interview and the ability to identify blunted affect, had the lowest reliability. Finding such poor reliability for these items that relied, in part, on direct behavioral observation seemed counterintuitive given that having behavioral anchors typically improved the reliability. The low reliability for these observational criteria may have partly been a function of the fact that the second rater viewed the interview via videotape and may have interpreted behaviors and affect differently than during a live interview. However, this finding is consistent with past research that has reported poor reliability for these observational criteria [25].

4.12. Limitations

The second-rater design used in this study meant that the second rater was not present and in the room during the original interview. Some behavioral nuances such as

gestures, eye contact, warmth or coldness of the participant, and interest in the interview may have been lost or distorted, in part, when viewing an interview on a videotape vs in person. On the other hand, raters viewing an interview via videotape have the luxury to rewind and watch again parts of the interview for clarification and analysis, something obviously not possible when doing an interview live. The differences inherent in being in the room during an interview vs viewing it on a videotape may have further reduced the reliability of difficult to rate observational criteria. The training that our interviewers went through did not focus a lot of time on how to rate these observational criteria, which may have also contributed to the poor reliability estimates found for these items. Future studies would benefit from extensive discussions and training on how to identify and rate these items.

One aspect unique to this sample was that the dress and appearance of all recruits were controlled. The Air Force supplied uniforms, dictated how they were to be worn, and even supplied standard eyeglasses for those who needed them. The standardization of appearance may have limited our ability to accurately diagnose histrionic PD, which is often expressed with extravagant, odd, or exaggerated apparel and schizotypal PD, which is often associated with an odd or unusual appearance. Our ability to accurately rate schizoid and schizotypal PD criteria that assess the presence of odd, eccentric, or peculiar appearances may have also been affected by the standardized appearances.

4.13. Conclusion

Overall, the SIDP-IV demonstrated good reliability in a non–treatment-seeking population. As in the study of Arntz et al [25], only the observational criteria had very poor reliability. Most other criteria have interrater reliabilities more than 0.70, with only a few exceptions where κ values were 0.6 or below. Our understanding of the reliability of structured interviews for PDs would be enhanced with further psychometric data from large samples of both patients and nonpatients of varying ages. The second-rater design used here appeared to work well and resulted in good reliability estimates. It is rather simple and inexpensive to record interviews on videotape and have them second rated.

References

[1] Cloninger CR, Martin RL, Guze SB, Clayton PJ. Diagnosis and prognosis in schizophrenia. Arch Gen Psychiatry 1985;42:15-25.

[2] Panzetta AF. Toward a scientific psychiatric nosology: conceptual and pragmatic issues. Arch Gen Psychiatry 1974;30:154-61.

[3] Rogers R. Handbook of diagnostic and structured interviewing. New York: The Guildford Press; 2001.

[4] Rogers R, Dion KL. Rethinking the DSM-III-R diagnosis of antisocial personality disorder. Bull Am Acad Psych Law 1991;19:21-31.

[5] Ward CH, Beck AT, Mendelson M, Mock JE, Erbaugh JK. The psychiatric nomenclature. Arch Gen Psychiatry 1962;7:198-205.

[6] American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 3rd ed. Washington (DC): Author; 1980.

[7] Loranger AW. Personality disorder examination (PDE) manual. Yonkers (NY): DV Communications; 1988.

[8] Widiger TA, Mangine S, Corbitt EM, Ellis CG, Thomas GV. Personality disorder interview–IV: a semistructured interview for the assessment of personality disorders. Odessa (Fla): Psychological Assessment Resources, Inc; 1995.

[9] First MB, Spitzer RL, Gibbon M, Williams JBW. Structured clinical interview for DSM-IV personality disorders (SCID-II). Washington (DC): American Psychiatric Press, Inc; 1997.

[10] Pfhol B, Blum N, Zimmerman M, Stangl D. Structured interview for DSM-III-R personality (SIDP-R). Iowa City, Iowa: Department of Psychiatry, University of Iowa; 1989.

[11] Pfohl B, Blum N, Zimmerman M. Structured interview for DSM-IV personality (SIDP-IV). Washington (DC): American Psychiatric Association; 1997.

[12] American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 3rd ed. rev. Washington (DC): Author; 1987.

[13] Zimmerman M. Diagnosing personality disorders: a review of issues and research methods. Arch Gen Psychiatry 1994;51:225-45.

[14] American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed Washington(DC): Author; 1994.

[15] Clark LA, Harrison JA. Assessment instruments. Handbook of personality disorders: theory, research and treatment. New York: Guilford Press; 2001. p. 277-306.

[16] Zimmerman M, Coryell W. The Reliability of personality disorder diagnoses in a non-patient sample. J Pers Disord 1989;3:53-7.

[17] Oltmanns TF, Turkheimer E. Perceptions of self and others regarding pathological personality traits. In: Krueger R, Tackett J, editors. Personality and psychopathology: building bridges. New York: Guilford Press [in press].

[18] Thomas RC, Turkheimer EN, Oltmanns TF. Factorial structure of pathological personality as evaluated by peers. J Abnorm Psychology 2003;112:1-12.

[19] Clifton A, Turkheimer E, Oltmanns TF. Contrasting perspectives on personality problems: descriptions from the self and others. Pers Individ Differ 2004;36:1499-514.

[20] Ford MR, Widiger TA. Sex bias in the diagnosis of histrionic and antisocial personality disorders. J Consult Clin Psychol 1989;57:301-5.

[21] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.

[22] Pagan JL, Oltmanns TF, Whittmore MJ, Turkheimer E. Personality disorder not otherwise specified: searching for an empirically based diagnostic threshold. Journal of Personality Disorders [in press].

[23] Fleiss JL. The measurement of interrater agreement. Statistical methods for rates and proportions. 2nd ed. New York: John Wiley & Sons; 1981. p. 212-36.

[24] Heumann KA, Morey LC. Reliability of categorical and dimensional judgments of personality disorder. Clin Res Rep 1990;147:498-500.

[25] Pilkonis PA, Heape CL, Proietti JM, Clark SW, McDavid JD, Pitts TE. The reliability and validity of two structured diagnostic interviews for personality disorders. Arch Gen Psychiatry 1995;52:1025-33.

[26] Arntz A, van Beijsterveldt B, Hoekstra R, Hofman A, Eussen M, Sallaerts S. The interrater reliability of a Dutch version of the structured clinical interview for DSM-III-R personality disorders. Acta Psychiatr Scand 1992;85:394-400.

[27] Mattia JI, Zimmerman M. Epidemiology. In: Livesley WJ, editor. Handbook of personality disorders: theory, research and treatment. New York: Guilford Press; 2001. p. 277-306.