Correspondence Between Self-Report and Interview-Based Assessments of Antisocial Personality Disorder

Laura S. Guy  
Simon Fraser University

Kevin S. Douglas  
Simon Fraser University and Mid-Sweden University

Norman G. Poythress  
University of South Florida

Jennifer L. Skeem  
University of California, Irvine

John F. Edens  
Texas A&M University

Antisocial personality disorder (ASPD) is associated with suicide, violence, and risk-taking behavior and can slow response to first-line treatment for Axis I disorders. ASPD may be assessed infrequently because few efficient diagnostic tools are available. This study evaluated 2 promising self-report measures for assessing ASPD—the ASPD scale of the Personality Diagnostic Questionnaire-4 (PDQ-4; S. E. Hyler, 1994) and the Personality Assessment Inventory (PAI; L. Morey, 1991, 2007)—as well as the ASPD module of the Structured Clinical Interview for DSM–IV Axis II (SCID-II; M. B. First, R. L. Spitzer, M. Gibbon, J. B. W. Williams, & L. S. Benjamin, 1997). The measures were administered to 1,345 offenders in court-mandated residential substance abuse treatment programs and prisons. PDQ-4 and PAI scores related strongly to SCID-II symptom counts ($r = .67$ and $.51$, respectively), indicating these measures convey useful clinical information about the severity of offenders’ ASPD pathology. The dimensional association between the measures was relatively invariant across gender, race, and site, although differences in mean scores were observed. Levels of agreement of the SCID-II with the PDQ-4 ($\kappa = .31$) and PAI ($\kappa = .32$) in classifying participants as ASPD was limited. Alternative thresholds for both self-report measures were identified and cross-validated.

Keywords: antisocial personality disorder, Personality Diagnostic Questionnaire-4, Personality Assessment Inventory, Structured Diagnostic Clinical Interview-II, prisoners

In the current diagnostic nomenclature, individuals who commit repeated criminal and/or other antisocial acts most often fall under the diagnostic label of antisocial personality disorder (ASPD). Establishing efficient and accurate procedures for diagnosing ASPD is important because the diagnosis is associated with several significant outcomes such as future involvement in antisocial behavior and engaging in risk-taking behaviors that have intentional (e.g., suicide attempts; Hills, Cox, McWilliams, & Sareen, 2005) and unintentional (e.g., motor vehicle crashes; see Nordstrom, Zwerling, Stromquist, Burmeister, & Merchant, 2001) consequences. A diagnosis of ASPD also has implications for intervention efforts, as personality disorders can complicate the first-line treatment of major mental disorders, thereby slowing rates of response (e.g., Kopta, Howard, Lowry, & Beutler, 1994; Shea, Widiger, & Klein, 1992). Moreover, reliable and valid diagnostic procedures are critical to informing decisions in the management of people with ASPD in both clinical (e.g., hospital release decisions) and correctional (e.g., sentencing, parole) settings.

Consistent with the lack of consensus regarding the construct validity of personality disorders (PDs; e.g., Clark, Livesley, & Morey, 1997), no single assessment tool or approach for diagnosing PDs is regarded as “universally accepted.” Both self-report questionnaires and clinical interviews present advantages and dis-
advantages. Structured interviews typically are resource intensive, often requiring interviewers with advanced clinical training to administer a lengthy interview to one individual at a time. The comparative efficiency of self-report measures, which may be relatively brief and, in some contexts, administered to groups of individuals by a technician (although subsequently interpreted by a trained clinician), has encouraged a number of investigators to develop measures of this type. On the other hand, shortcomings associated with self-report measures include respondents' idiosyncratic understanding of items, inability to clarify state versus trait issues, and inability to appraise the validity of the basis upon which respondents endorsed criteria (see Smith, Klein, & Benjamin, 2003). It would seem that best clinical practice would draw on the advantages of both assessment approaches such that a reliable and valid self-report tool could be used to augment diagnostic and symptom information gathered during an interview. However, following such an approach requires demonstrated concordance between measures. In addition, it is important to examine the concordance of self-report questionnaires with clinical interviews because, historically, they have tended to overdiagnose PDs (Guthrie & Mobley, 1994).

The present study investigated the convergence of two subscales from self-report measures with differing conceptual and empirical bases of development—the ASPD scale of the Personality Diagnostic Questionnaire-4 (PDQ-4; Hyler, 1994) and the Antisocial Features (ANT) scale of the Personality Assessment Inventory (PAI; Morey, 1991, 1996)—with the ASPD module of one of the most well known and often-used semistructured interviews: the Structured Clinical Interview for DSM-IV Axis II (SCID-II; First et al., 1997). The PDQ-4 is a forced choice questionnaire whose items closely track the DSM-IV diagnostic criteria. The PDQ-4 has been used in studies of psychopathology in clinical samples (e.g., borderline personality; see van den Bosch, Verheul, Schippers, & van den Brink, 2002) and in studies of the relation between PD and measures of general personality constructs in nonreferred samples (e.g., Fossati et al., 2004). Several studies have focused on the ASPD scale of the PDQ-4 with offenders and individuals with other problem behaviors, including forensic hospital inpatients (e.g., Whyte, Fox, & Coxell, 2006) and prisoners (e.g., Tye & Mullen, 2006). Of the studies cited that investigated justice-involved samples, only Blackburn, Donnelly, Logan, and Renwick (2004) reported the internal consistency of the ASPD scale (α = .88). Chance-corrected categorical agreement between the PDQ-4 ASPD scale and criterion diagnoses based on semistructured clinical interview and records review has been fair to poor in the few studies published thus far (e.g., Davison, Leese, & Taylor, 2001).

The PAI is a personality inventory that was designed to assess critical clinical variables. It has been used in studies covering a wide array of populations (see Morey, 1991, 2007). The ANT scale was developed to measure symptoms of antisocial personality disorder and psychopathy. ANT consists of three conceptually distinct subscales whose item content taps disparate facets of antisociality: ANT-A (Antisocial Behaviors), which assesses a history of conduct problems and criminality; ANT-S (Stimulus Seeking), which reflects a tendency toward thrill-seeking and low boredom tolerance; and ANT-E (Egocentricity), which taps a self-centered, callous, remorseless interpersonal style. Compared with that for the PDQ-4, the body of research that supports the validity of the PAI for use in forensic and correctional settings is more developed (for reviews, see Edens & Ruiz, 2005; Morey, 1991, 2007). Some research has been completed on the diagnostic concordance between clinical interviews and some PAI scales, but little research to date has investigated the concordance of the ANT scale with interview-based measures of ASPD.

Although several clinical inventories are available for assessing ASPD, and commentators have noted that “there is little empirical justification for preferring one interview for ASPD over others” (Lilienfeld, Purcell, & Jones-Alexander, 1997, p. 66), the SCID-II was used in the present study as the criterion because it is widely used and is among the most well researched semistructured interview tools for use by clinicians. In a review of the convergence between structured interviews and questionnaires assessing PDs, median kappas across PD categories for measures from 19 studies indicated that the SCID-II evidenced slightly higher rates of convergence with self-report questionnaires relative to other interview guides (Clark et al., 1997).

The present study fills an important gap in research on the assessment of ASPD with offenders. Using two large samples of offenders drawn from prisons and residential substance abuse treatment programs, we report on the psychometric properties of the PDQ-4 ASPD and PAI ANT scales and their diagnostic efficiency against the SCID-II as a criterion measure. Building a solid body of empirical knowledge about the degree of concordance between self-report measures and clinical interviews is important because it may inform decision making regarding effective allocation of typically scarce human and financial resources. Another important aim of the present project was to evaluate whether there are differences in reliability and diagnostic efficiency across race and gender, issues which to our knowledge have not been explored using the PAI with offenders. Although no systematic investigation of these issues has occurred with the PDQ-4 either, researchers (Cale & Lilienfeld, 2002) have reported a large gender difference on scores on the PDQ-4 + ASPD scale (Cohen’s d = 0.86). To ensure the fair use and unbiased application of psychological measures, one must evaluate whether they operate differentially across groups of people, especially given that minorities are overrepresented in the criminal justice system (Federal Bureau of Prisons, 2007) and comparatively less research has been conducted with female offenders (Blanchette & Brown, 2006).

Method

Participants

Participants were 1,345 offenders either incarcerated in prisons (n = 678) or court-ordered to participate in substance-related residential treatment (n = 667). They were recruited from state prisons and residential drug treatment sites in Florida, Nevada, Oregon, Texas, and Utah. Eligibility criteria for study inclusion were: (1) Black or White, (2) English speaking, (3) estimated screening IQ of at least 70, and (4) not receiving psychotropic medication for active symptoms of psychosis. The primary target age range was 21–40 years, although 109 participants over 40 years were recruited and retained in analyses. The mean age was 1 The PDQ-4+ includes the same items that assess the established PDs as the PDQ-4 but differs from the PDQ-4 by having additional items that assess two experimental personality disorders.
31.40 years ($SD = 6.69$). Most participants were men (83%) and White (66%; 34% Black; 10% additionally identified themselves as of Hispanic ethnicity).

**Measures**

**Structured Clinical Interview for DSM–IV Axis II (SCID-II) Antisocial Personality Disorder (ASPD) Scale (First et al., 1997).** The ASPD module assesses the 22 possible symptom indicators of **DSM–IV** criteria for ASPD. It yields both dimensional and categorical scores. SCID-II scores in the present study were based on information obtained during the interview and from a detailed review of participants’ institutional files. SCID-II assigns a diagnosis of ASPD if at least two items from the conduct disorder criteria and at least three items from the adult criteria are endorsed. High levels of interrater reliability (e.g., Maffei et al., 1997) and high concurrent validity for consensus diagnoses of ASPD (e.g., Skodol, Rosnick, Kellman, Oldham, & Hyler, 1988) have been demonstrated for the ASPD module.

For the present sample, interrater reliability of the SCID-II was determined through observation of SCID-II interviews of study participants by research assistants (RAs), who were advanced-level clinical psychology graduate students, in addition to the requisite file reviews. All such observations were done by one of the current authors (Kevin S. Douglas), whose SCID-II scores were treated as the “criterion” against which RA scores were measured. Kevin S. Douglas traveled to each site approximately every 6 months and observed two cases per visit, for a maximum of six visits. Given some minor variations in this general procedure, there were a total of 51 live interrater reliability cases (3.8% of the sample). Concordance was excellent for ASPD diagnoses ($k = .74$; $n = 50$). Intercorater reliability for total symptom count also was high: ICC, which is a measure of agreement that is acceptable for noncategorical data (Bartko & Carpenter, 1976; Cicchetti & Sparrow, 1981), was .86 ($n = 46$). For the 22 items comprising the SCID-II ASPD module for the present sample, internal consistency ($\alpha$) was .83 and the mean interitem correlation (MIC) was .18.

**Personality Diagnostic Questionnaire–4 (PDQ–4) ASPD Scale (Hyler, 1994).** The PDQ–4 ASPD scale consists of 22 true–false self-report items (one for each **DSM–IV** ASPD criterion). To meet diagnostic threshold for ASPD on the PDQ–4, at least three items from the conduct disorder criteria and at least four items from the adult criteria need to be endorsed (personal communication, S. Hyler, November 9, 2006). Participants were classified into ASPD positive and negative groups based on these parameters, which were used in the analyses reported below. Internal consistency for the PDQ–4 was good ($\alpha = .85$); MIC was .20.

**Personality Assessment Inventory: Antisocial Features Scale (PAI ANT, Morey, 1991, 2007).** Noted above, ANT was designed to tap the core affective, interpersonal, and behavioral features that traditionally have been associated with psychopathy and antisocial personality. Internal consistency for ANT among offender samples typically has been good (e.g., $\alpha = .82$; Edens & Ruiz, 2005). In terms of validity data, ANT typically correlates moderately to highly with other measures of antisocial and psychopathic traits. ANT has also been shown to predict various types of theoretically relevant forms of social deviance (e.g., institutional adjustment difficulties) among offenders (for recent reviews, see Edens and Ruiz, 2005; Morey, 2007). Although there is no specific diagnostic cutoff that is recommended in the PAI manual, scores greater than or equal to 70 $T$ frequently have been examined in group-level analyses (Edens & Ruiz, 2005) and are indicative of individuals who should be “impulsive and hostile, perhaps with a history of reckless and/or antisocial acts” (Morey, 2007, p. 42). Accordingly, 70 $T$ was used as the primary cutoff for classification into the ANT-defined ASPD group in the present study. For the 24 items comprising ANT, internal consistency ($\alpha$) was .75 and MIC was .11 (based on a subset of approximately 700 participants who were available for these item-level analyses).

**Procedure**

Data were collected as part of a larger study in which RAs were trained in the administration of the study protocol prior to data collection. At each site, RAs randomly selected potential participants from lists of individuals who met inclusion criteria for the study. Enrollment interviews were conducted in a private room, and informed consent was obtained through procedures approved by a university institutional review board. Next, an IQ screening test was administered; participants whose estimated IQ was below 70 were excused from the study ($n = 6$). A reading ability test was administered to participants who did not meet certain educational history requirements and who had difficulty reading the first few items of the PAI. The larger study protocol included several additional measures (not described here), took on average 4.5 hr to complete, and typically was administered in two sessions. Also, a detailed review of each participant’s institutional files was completed. At the end of protocol administration, $20$ was deposited into the agency account of each participant, unless reimbursement was prohibited by the agency’s policies (one site).

The participants included in analyses ($N = 1,345$) are those for whom there were complete data on the PDQ–4, SCID-II, PAI ANT, and PAI Inconsistency and Infrequency scales. These two PAI validity scales were used to evaluate participants’ vulnerability to careless or random responding (data for the PDQ–4 validity indices were not available because only the ASPD scale was administered). Consistent with other research, participants were retained for analyses if they obtained Inconsistent and Infrequency scores below 80 $T$ (see Edens & Ruiz, 2005). Individuals were also excluded if they did not complete or were not administered the PDQ–4, SCID-II, or PAI ($n = 225$). Of the remaining 1,516 individuals, 171 were screened out because data were missing for one or more PDQ–4 or SCID-II items or the PAI scores were invalid, resulting in the final sample size of 1,345 participants. There were no statistically significant differences in age or gender between the 171 excluded participants and those retained. However, significantly more of the excluded participants were recruited from prisons ($n = 112$) than from drug treatment facilities ($n = 59$; $\chi^2 = 13.84, p < .001$).

**Results**

The primary aim of the present study was to evaluate the psychometric properties of the PDQ–4 ASPD and PAI ANT scales and their diagnostic efficiency relative to the SCID-II ASPD module. We examined concordance at dimensional and categorical levels and calculated traditional indices of efficiency, including sensitivity, specificity, positive predictive power (PPP), negative predictive power (NPP), overall hit rate, and receiver operating characteristic analyses.
First, in terms of descriptive information and bivariate correlations between measures, mean total symptom counts for the 22 items on the PDQ-4 ($M = 7.98, SD = 4.65$) and SCID-II ($M = 7.42, SD = 4.37$) ASPD scales were significantly different, $t(1344) = 5.57, p \leq .01$; Cohen’s $d = 0.12$. The mean ANT score was $70.93$ ($SD = 11.90$). The zero-order correlations between the measures’ total scores were large and all statistically significant ($p < .01$): SCID-II/PDQ-4 (.67); SCID-II/PAI (.51); and PDQ-4/PAI (.66). The SCID-II and PAI classified similar proportions of the sample with ASPD (SCID-II: $n = 744, 55$%; PAI: $n = 721, 54$%), whereas the PDQ-4 classified fewer participants into this category ($n = 456, 34$%). Rates of diagnostic agreement (kappa) between the measures were: SCID-II/PDQ-4 (.48); SCID-II/PAI (.39); and PDQ-4/PAI (.40). Generally, kappa values of $.75$ and greater are considered to reflect excellent agreement; $.60$–$.74$, good agreement; $.40$–$.59$, fair agreement; and $.00$–$.39$, poor agreement (Cicchetti & Sparrow, 1981).

Diagnostic efficiency statistics are presented in Table 1 (for the PDQ-4) and Table 2 (for the PAI). In these analyses, classifications of PDQ-4 ASPD and PAI ANT were compared with classifications of SCID-II ASPD, which was treated as the criterion. We used tests of the equality between proportions to gauge whether the classification indices (sensitivity, specificity) of the measures differed from one another. These tests revealed that whereas the PAI was significantly more sensitive than the PDQ-4 ($\chi^2 = 18.71, p < .01$), the PDQ-4 was significantly more specific than ANT ($\chi^2 = 11.99, p < .01$).

PPP for the PDQ-4 (.79) was higher than for ANT (.70), and NPP for ANT (.62) was higher than for the PDQ-4 (.57). The hit rates, which are the proportion of correct decisions, for the PDQ-4 and ANT were comparable (.64 and .66, respectively). Area under the curve (AUC) values in the tables represent the likelihood that a randomly selected individual with SCID-II-defined ASPD will have a higher PDQ-4 score (i.e., symptom count) or ANT T score than that of a randomly selected individual without ASPD. The AUC for the PDQ-4 (.80, $SE = .01$) was significantly larger than the AUC for ANT (.72, $SE = .01$); $z = 4.83, p < .01$. Given the rather modest rates of diagnostic concordance, we investigated alternative PDQ-4 decision thresholds for ASPD diagnostic classification. The PDQ-4, as mentioned, requires the identification of at least three conduct disorder symptoms and four adult symptoms for ASPD to be considered present. Given that

Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>PDQ-4*</th>
<th>SCID-II*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
<td>$%$</td>
</tr>
<tr>
<td>Total sample ($N = 1,345$)</td>
<td>456</td>
<td>34</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women ($n = 223$)</td>
<td>74</td>
<td>33</td>
</tr>
<tr>
<td>Men ($n = 1,122$)</td>
<td>382</td>
<td>34</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black ($n = 454$)</td>
<td>140</td>
<td>31</td>
</tr>
<tr>
<td>White ($n = 869$)</td>
<td>311</td>
<td>36</td>
</tr>
<tr>
<td>Site</td>
<td>Prison ($n = 678$)</td>
<td>185</td>
</tr>
<tr>
<td>Drug treatment ($n = 667$)</td>
<td>271</td>
<td>41</td>
</tr>
</tbody>
</table>

Note. PDQ-4 = Personality Diagnostic Questionnaire-4; SCID-II = Structured Clinical Interview for DSM-IV Axis II; ASPD = antisocial personality disorder; PPP = positive predictive power; NPP = negative predictive power; HR = hit rate; AUC = area under the curve. * Values represent participants who met the measure’s diagnostic threshold (for PDQ-4, at least four adult items and at least three conduct disorder items; for SCID-II, at least three adult items and at least two conduct disorder items). ** $p < .05$. *** $p < .001$.

2 According to Cohen (1992), small, moderate, and large $d$ values are .20, .50, and .80, respectively.

3 Correlations between the ANT subscales were as follows: ANT-A/ANT-E (.47); ANT-A/ANT-S (.52); ANT-E/ANT-S (.51), all $ps \geq .001$.

4 Because the PAI manual discusses scores of 60 $T$ and 82 $T$ as meaningful breaking points, we also investigated the diagnostic efficiency within the total sample using these cut scores in addition to the cut score of 70 $T$, which was used in all analyses. Diagnostic agreement was lower using 60 $T$ and 82 $T$ ($\kappa = .21$ and .18, respectively). Results of utility analyses for 60 $T$ and 82 $T$, respectively, were as follows: sensitivity (.93, .29); specificity (.27, .92); PPP (.61, .80); NPP (.76, .51); HR (.64, .56).

5 Aside from ANT, the PAI provides an additional method of measuring antisocial features that has not been the focus of much research to date. Using data from the clinical normative group (Morey, 1991), Morey (1996) developed a logit function for estimating ASPD: 0.044(Antisocial Behaviors) + .017(Aggression–Physical Aggression) – .008(Antisocial Egocentricty) – .002(Antisocial Stimulus Seeking) – .028(Anger–Affective) + 1.85. The PAI clinical interpretive software uses a cut of 2 probability value on this function to flag ASPD as a diagnostic consideration (personal communication, Leslie C. Morey, August 1, 2007). With this probability as a cut score in our derivation sample, 36% ($n = 241$) of participants were classified as having ASPD. The categorical rate of agreement with SCID-II diagnosis was $\kappa = .34$. The correlations between the continuous variable created by the logit function and total scores on SCID-II, PDQ-4, and ANT were .46, .49, and .52, respectively (all $ps < .01$). When this cut score was applied in the holdout sample described later in the text, 38% ($n = 254$) of participants were classified as having ASPD and kappa decreased to .25. The correlations between the continuous variable created by the logit function and total scores on SCID-II, PDQ-4, and ANT were .48, .52, and .55, respectively (all $ps < .01$). Compared with the ANT cut score of 70 $T$ (see Table 2), the cross-validated logit function yielded a similar overall hit rate (.66 for 70 $T$ vs. .61 for the function). Other diagnostic efficiency results for the logit function in the holdout sample were: sensitivity (.49), specificity (.77), PPP (.72), NPP (.55). Given that the logit function at best performed comparably to ANT in isolation, the subsequent analyses reported below focus solely on the performance of ANT.
there was a stronger association in the present sample between SCID-II total scores and the number of conduct disorder symptoms endorsed on the PDQ-4 \( (r = .67, p < .01) \) compared with the number of adult symptoms \( (r = .45, p < .01) \), it is possible that different combinations of symptom counts (e.g., four conduct disorder and two adult symptoms) would provide increased concordance between the SCID-II and PDQ-4. To examine the impact on diagnostic concordance between the two measures when different numbers of adult (Criterion A) and conduct disorder (Criterion C) symptoms are endorsed on the PDQ-4, we calculated the chance-corrected agreement (kappa) between the measures’ classifications for all 70 permutations of adult and conduct disorder symptoms.

We first split the sample randomly to create derivation \( (n = 668) \) and “holdout,” or cross-validation \( (n = 677) \), subgroups. The best item combination in the derivation sample yielded a poor to fair level of agreement \( (\kappa = .40) \), based on endorsement of four or more conduct disorder items and three or more adult items on the PDQ-4. When this symptom combination was tested in the holdout sample, agreement continued to be fair \( (\kappa = .43) \). As a point of comparison, kappa values in the derivation and holdout samples using the recommended PDQ-4 criteria (i.e., three or more conduct disorder and four or more adult items) were poor \( (\kappa = .25 \text{ and } \kappa = .36, \text{ respectively}) \). Compared with these PDQ-4 criteria, the alternative criteria tested (i.e., four or more conduct disorder and three or more adult items) also yielded higher values for the overall hit rate \( (.70 \text{ vs. } .64) \), sensitivity \( (.63 \text{ vs. } .48) \), and NPP \( (.63 \text{ vs. } .57) \). Specificity was lower \( (.80 \text{ vs. } .85) \), and PPP remained unchanged \( (.79) \).

An alternative basis for deriving cut scores that did not take into account the number of conduct disorder and adult symptoms on the PDQ-4 also was investigated. Using the total set of 22 PDQ-4 items, we formed dichotomous groups of ASPD/no ASPD for cut scores ranging from 5 to 10. Rates of agreement were then calculated between the SCID-II and PDQ-4 using these six cut scores. Kappa values for the PDQ-4 cut scores of 5 through 10 were .36, .41, .46, .42, .39, and .42, respectively. The cut score of 7, which yielded the largest kappa and best overall hit rate \( (.73) \) in the derivation sample, produced the highest rate of diagnostic concordance in the holdout sample \( (\kappa = .49) \). Thus, irrespective of whether the PDQ-4 cut score is attained by endorsement of at least three adult and four conduct disorder items (Hyler’s criteria; personal communication) or vice versa (the criteria proposed herein), there is consistency in that at least seven items should be endorsed.

We also examined, as an additional approach to investigating cut scores that did not take into account the number of conduct disorder and adult symptoms on the PDQ-4, the utility of using either only conduct disorder or only adult symptoms. The mean number of conduct disorder and adult symptoms endorsed in the entire sample was 4.91 \( (SD = 3.50) \) and 3.07 \( (SD = 1.75) \), respectively. First, by examining only the 15 PDQ-4 conduct disorder items and using the total SCID-II score as the criterion, we obtained the largest kappa value in both the total \( (\kappa = .45) \) and derivation \( (\kappa = .47) \) samples for the cut score of 4 (overall HR in the total and derivation samples were .73 and .74, respectively). In the holdout sample, this cut score yielded a kappa of .43 \( (HR = .72) \). Second, by examining only the seven PDQ-4 adult symptom items, we obtained the largest kappa value in both the total \( (\kappa = .27) \) and derivation \( (\kappa = .23) \) samples for the cut score of 3 (overall HR in the total and derivation samples were .64 and .62, respectively). In the holdout sample, this cut score yielded a kappa of .31 \( (HR = .66) \).

The second aim of this project involved “testing the boundaries” of generalization of the main findings. Specifically, we evaluated whether there were differences in reliability and diagnostic efficiency across race and gender. We also investigated the potential impact of the site from which participants were drawn (substance abuse treatment facilities or prisons). As a first step in evaluating performance of the measures across these groups, we compared the measures’ total scores across these domains. Gender differences were observed, with men scoring significantly higher than women on the PDQ-4 \( (\text{men: } M = 8.10, SD = 4.65; \text{women: } M = 7.38, SD = 4.61) \), \( t(1343) = 2.12, p = .03; d = 0.16 \), and SCID-II \( (\text{men: } M = 7.69, SD = 4.37; \text{women: } M = 6.06, SD = 4.15) \), \( t(1343) = 5.14, p < .01; d = 0.38 \), but not on ANT \( (\text{men: } M = 71.15, SD = .05 \text{ and } .01). \)

### Table 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>PAI ANT</th>
<th>SCID-II</th>
<th>( \kappa )</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPP</th>
<th>NPP</th>
<th>HR</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sample ( (N = 1,345) )</td>
<td>721</td>
<td>54</td>
<td>744</td>
<td>55</td>
<td>.32**</td>
<td>.51*</td>
<td>.68</td>
<td>.64</td>
<td>.70</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women ( (n = 223) )</td>
<td>106</td>
<td>48</td>
<td>97</td>
<td>44</td>
<td>.31**</td>
<td>.48*</td>
<td>.65</td>
<td>.66</td>
<td>.59</td>
</tr>
<tr>
<td>Men ( (n = 1,122) )</td>
<td>615</td>
<td>55</td>
<td>647</td>
<td>58</td>
<td>.32**</td>
<td>.52*</td>
<td>.68</td>
<td>.64</td>
<td>.72</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black ( (n = 454) )</td>
<td>203</td>
<td>45</td>
<td>254</td>
<td>56</td>
<td>.28**</td>
<td>.51*</td>
<td>.57</td>
<td>.72</td>
<td>.72</td>
</tr>
<tr>
<td>White ( (n = 869) )</td>
<td>508</td>
<td>59</td>
<td>478</td>
<td>55</td>
<td>.34**</td>
<td>.54*</td>
<td>.74</td>
<td>.60</td>
<td>.69</td>
</tr>
<tr>
<td>Site</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prison ( (n = 678) )</td>
<td>300</td>
<td>44</td>
<td>376</td>
<td>56</td>
<td>.31**</td>
<td>.52*</td>
<td>.58</td>
<td>.73</td>
<td>.73</td>
</tr>
<tr>
<td>Drug treatment ( (n = 667) )</td>
<td>421</td>
<td>63</td>
<td>368</td>
<td>55</td>
<td>.33**</td>
<td>.54*</td>
<td>.78</td>
<td>.55</td>
<td>.68</td>
</tr>
</tbody>
</table>

Note. PAI ANT = Antisocial Features scale of the Personality Assessment Inventory; SCID-II = Structured Clinical Interview for DSM-IV Axis II; ASPD = antisocial personality disorder; PPP = positive predictive power; NPP = negative predictive power; HR = hit rate; AUC = area under the curve.

* Values represent participants who met diagnostic threshold (for PAI ANT, \( T \geq 70 \); for SCID-II, at least three adult items and at least two conduct disorder items).

\( p < .05 \). ** \( p < .001 \).
11.89; women: \( M = 69.81, SD = 11.94 \), \( t(1343) = 1.54, p = .13 \); 
\( d = 0.11 \). Significant differences across race also were observed: 
Whereas Whites had higher ANT scores (\( M = 72.03, SD = 12.10 \)) 
than did Blacks (\( M = 68.89, SD = 11.24 \), \( t(1321) = 4.60, p < .01 \); 
\( d = 0.27 \), Blacks had higher SCID-II scores (\( M = 7.79, SD = 4.42 \)) 
than did Whites (\( M = 7.22, SD = 4.31 \), \( t(1321) = 2.27, p = .02 \); 
\( d = 0.13 \). PDQ-4 scores did not differ significantly between 
the two groups (Whites: \( M = 8.17, SD = 4.70 \); Blacks: \( M = 7.65, 
SD = 4.55 \), \( t(1321) = 1.93, p = .05 ; d = 0.11 \). Scores on the 
ASPD scales of the two self-report measures were significantly 
higher in the treatment sample compared with the prison sample. 
Mean scores on the PDQ-4 were: treatment (\( M = 8.49, SD = 4.74 \)); 
prison (\( M = 7.48, SD = 4.51 \), \( t(1343) = 4.00, p < .01 ; d = 0.22 \). 
Mean scores on ANT were: treatment (\( M = 73.30, SD = 11.70 \)); 
prison (\( M = 68.60, SD = 11.65 \), \( t(1343) = 7.38, p < .01 \); 
\( d = 0.40 \). However, no significant differences between site 
samples were observed for the interview-based SCID-II (treatment: 
\( M = 7.32, SD = 4.30 \); prison: \( M = 7.52, SD = 4.44 \), \( t(1343) = 
0.83, p = .41 ; d = 0.05 \). We note that all significant differences 
were small in magnitude, according to Cohen’s recommended 
interpretations of standardized mean differences (Cohen, 1992).

Tables 1 (PDQ-4) and 2 (PAI) present indicators of the corre-
respondence between the SCID-II and the two self-report measures 
among the various groups. Based on tests of the differences be-
tween independent correlations, there were no significant differ-
ences in diagnostic agreement between the SCID-II and either 
the PDQ-4 or ANT as a function of gender, race, or referral sample 
(pri son vs. substance abuse; see Tables 1 and 2). Similar results 
were obtained when AUC values were examined. For the PDQ-4, 
AUCs across groups ranged from .78 (for Blacks) to .83 (for 
women and participants at drug treatment sites), and all were 
statistically significant (\( SEs \) ranged from .01 to .03). AUC values 
for ANT were comparatively smaller and ranged from .70 (Blacks) 
to .73 (Whites and prisoners); all were statistically significant as 
well (\( SEs \) ranged from .01 to .04). There were no significant 
differences in AUC values between the subgroups (i.e., gender, 
race, or referral) for either self-report questionnaire.

Results for the PDQ-4 (see Table 1) revealed good specificity 
but limited sensitivity. To test whether sensitivity and specificity 
differed significantly across groups, we computed the equality of 
the proportions of ASPD classifications for each group. There 
were no significant differences in sensitivity between women 
and men or between Blacks and Whites. However, sensitivity was 
significantly higher among drug treatment participants compared 
with prisoners (\( \chi^2 = 11.39, p < .01 \)). Analyses of specificity 
indicated that although no significant differences were observed 
across race or site groups, specificity was significantly higher for 
women compared with men (\( \chi^2 = 5.86, p = .02 \)). Variation within 
groups was observed for both overall hit rates and AUC values.

For ANT (see Table 2), the expected tradeoff between specifici-
ty and sensitivity was smaller compared with that observed for 
the PDQ-4. Tests of the equality of proportions for sensitivity revealed 
significant differences across all the groups. Compared with their 
counterpart group, ANT more frequently correctly classified the 
following groups as having ASPD: men (\( \chi^2 = 4.73, p = .03 \)), 
Whites (\( \chi^2 = 4.09, p = .04 \)), and drug treatment participants (\( \chi^2 = 
7.29, p = .01 \)). Analyses for specificity yielded significant differ-
ences for gender, with women more often than men being correctly 
classified by ANT as not having ASPD (\( \chi^2 = 5.10, p = .02 \)), and 
for site, with prisoners more often than treatment participants 
being correctly classified by ANT as not having ASPD (\( \chi^2 = 5.87, 
p = .02 \)). Specificity rates did not differ significantly for racial 
subgroups. Table 2 also indicates substantial invariance across 
groups for overall hit rates and AUC values.

Discussion

Our first objective was to evaluate the psychometric properties of 
the ASPD scales of two self-report measures—the PDQ-4 and 
PAI—and their diagnostic efficiency compared with the SCID-II 
ASPD scale. Results indicated that all three measures were related 
strongly at the dimensional (i.e., symptom count) level, particu-
larly when one considers that the strength of the relationship 
between the two is constrained by the reliability of the scales 
themselves. At the diagnostic level, however, our results were less 
supportive of the concordance between measures. Rates of cate-
gorical agreement with the SCID-II were comparably poor for the 
PDQ-4 (\( \kappa = .31 \)) and ANT (\( \kappa = .32 \)), and results did not support 
the interchangeability of the interview-based SCID-II with self-
report measures. Introducing method variance clearly had an im-
 pact, as the rate of categorical agreement between the two self-
report measures was still limited, but higher (\( \kappa = .40 \)) than the rate 
for either measure with the SCID-II. The diagnostic discordance 
between the PDQ-4 and SCID-II is not surprising given the dif-
ferential symptom requirements to make diagnoses of ASPD 
across the two instruments.\(^6\) Although self-reports tend to overdi-
agnose personality disorders (Guthrie & Mobley, 1994), the 
PDQ-4 criteria are more stringent than SCID-II criteria in that they 
require two additional symptoms (one child, one adult) to diagnose 
ASPD. Nevertheless, despite the higher thresholds used by the 
PDQ-4, our finding that the SCID-II yielded a higher proportion of 
ASPD cases than did the PDQ-4 was not expected in light of 
research demonstrating that self-report measures tend to overdiag-
nose. Similarly, the finding that ANT (using 70 \( T \) as a cutoff) and 
the SCID-II classified similar proportions of participants as having 
ASPD was unexpected.

From a psychometric perspective, a lack of association in di-
ensional scores would have been more problematic than the 
observed discordance in diagnostic classifications. That is, al-
though dichotomous diagnoses are attractive from a clinical per-
spective, the practice of assigning diagnoses assumes the presence 
of taxonicity (i.e., that there is a genuine demarcation between 
the presence versus absence of the disorder), and this is not supported 
by empirical evidence (Marcus, Lilienfeld, Edens, & Poythress, 
2006). Additional challenges to attaining high rates of diagnostic 
concordance for personality disorders have been described, such as 
the failure to demonstrate that personality disorders correspond to 
discrete psychiatric entities and the lack of consistent, robust 
evidence of the longitudinal stability of diagnostic status (see 
McCrae et al., 2001, for an informed discussion). In addition, and 
of particular importance to the present study, the use of clinical

\(^6\) Indeed, supplementary analyses indicated that diagnostic concordance was increased when the same threshold was used for both measures. When the PDQ-4 criteria (i.e., four or more adult symptoms and three or more child symptoms) were applied to the SCID-II, \( \kappa = .33 \). When the SCID-II criteria (i.e., three or more adult symptoms and two or more child symp-
toms) were applied to the PDQ-4, \( \kappa = .37 \).
ratings as the criterion against which the accuracy of self-report measures is compared may be problematic in light of research (Clark et al., 1997) indicating low agreement among psychiatrists’ ratings of personality disorder when using different interviews (although the “clinical” ratings in the current study were collected for research purposes and evidenced high reliability).

Traditionally, screening tools are designed to “screen in” for further assessment individuals who may have a given disorder. Applying the PDQ-4 criteria yielded poor sensitivity and low negative predictive power. These are undesirable features in a screening tool designed to identify offenders with potential ASPD. However, the high specificity and positive predictive power indicate that a favorable application of the PDQ-4 using the recommended cutoff value might be to “screen out” individuals who do not have ASPD. This might be a desirable attribute in certain clinical contexts. For example, one might wish to use the PDQ-4 to identify persons who are unlikely to have ASPD and therefore are appropriate for inclusion in interpersonal psychotherapy groups (Yalom & Leszcz, 2005). In contrast to the PDQ-4, the PAI was significantly more sensitive and less specific, which suggests that it would be a preferable measure compared with the PDQ-4 for identifying persons in need of more detailed assessment.

Should one’s goal be to use the PDQ-4 in the more traditional manner to “screen in” offenders who may have ASPD, however, one may wish to use the alternative criteria we identified to better distinguish between SCID-II-defined ASPD and non-ASPD cases. When the number of conduct disorder and adult symptoms was considered, optimal classification was achieved at four or more child and three or more adult symptoms, which also cross-validated well in the holdout sample. Unfortunately, however, the level of diagnostic agreement was raised only from “poor” (recommended criteria) to “fair” (alternative criteria)—even when no constraints were imposed on the specific number of conduct disorder and adult symptoms required to meet ASPD diagnostic criteria. As such, it seems that the PDQ-4 performs better as a dimensional tool than as a categorical tool. This finding is consistent with reviews (e.g., Lilienfeld et al., 1997) concluding that although diagnostic concordance between interview and self-report measures of ASPD generally is poor, the association between measures appears more robust when the total symptom count is considered. Knowledge about the severity of an individual’s personality pathology could be informative in several clinical decision-making contexts—and would entail a dimensional application of the PDQ-4. To the extent that dimensional models of personality disorder may be considered in revisions of the DSM (Trull, Tragesser, Solhan, & Schwartz-Mette, 2007; Widiger & Trull, 2007), our findings suggest that the PDQ-4 may be a useful clinical tool.

Our second objective was to examine the degree to which our findings using the total sample would hold up across referral site, gender, and race. When mean scores on each of the three tests were examined, several significant differences were observed. However, it is important to note that statistical significance generally is easily attainable when studying samples as large as was investigated in the present study. Moreover, all of the effect sizes obtained would be characterized as small in magnitude (Cohen, 1992). Also, and importantly, although we observed some mean differences in scores, the degree of dimensional association between the self-report measures and the SCID-II appeared to be relatively invariable across the three subgroups. We must therefore consider the possibility that, in addition to one interpretation of our results as indicating “genuine” differences in mean scores, the results are due to bias of some sort (see Widiger, 1998).

Bias regarding the application of the diagnostic criteria can result either from clinicians failing to adhere to DSM criteria or failing to apply the criteria equally to different members of a group (e.g., Blacks and Whites). Our finding that men had higher mean scores on the PDQ-4 and SCID-II than did women is consistent with research indicating that prevalence rates of ASPD are relatively higher among men (American Psychiatric Association, 2000). However, this finding—especially given that the largest mean difference for gender was observed for the interview-based SCID-II—also is consistent with research documenting the presence of interviewer bias. For example, Garb (1997) reviewed empirical literature indicating that clinicians may be less likely to perceive women as violent or antisocial. Our finding that Blacks were given higher ratings than Whites on an interview-based measure also is consistent with Garb’s conclusions that clinicians may be more likely to perceive Blacks as violent or antisocial.

Although we are unable to offer a definitive explanation for the differential findings across groups, it is important to document and pursue them in future work that focuses specifically on different forms of potential gender and racial bias. Work in this area is critical, especially given that much more research on psychological assessment in correctional and forensic settings has been conducted with men than with women (Blanchette & Brown, 2006) and minorities are overrepresented in U.S. prisons (Federal Bureau of Prisons, 2007).

Strengths of this study are the use of large and diverse samples across several settings, comparisons across heretofore unexamined subgroups of practical and theoretical interest, and use of trained examiners with evidence of high interrater reliability. Various limitations should be noted, however. For example, we examined only two primary racial groups (Blacks and Whites), which limits the generalizability of our findings to other groups (e.g., Asians, Native Americans). Another limitation of the present study pertains to our use of “offender” samples, which raises questions as to how well the self-report measures would work as a diagnostic tool in samples where the base rates of ASPD would be appreciably lower (e.g., epidemiological studies of community samples) and the potential for false positives (versus false negatives) would be much more likely. These caveats notwithstanding, the current results showed strong concordance of the PDQ-4 and ANT with the SCID-II at the dimensional level but limited correspondence diagnostically.

References


Received March 14, 2007
Revision received November 7, 2007
Accepted November 8, 2007