A Psychometric Evaluation of the DSM-IV Criteria for Antisocial Personality Disorder

Paap, Muirne C. S.; Braeken, Johan; Pedersen, Geir; Urnes, Øyvind; Karterud, Sigmund; Wilberg, Theresa; Hummelen, Benjamin

Published in:
Assessment

DOI:
10.1177/1073191117745126

Article

A Psychometric Evaluation of the *DSM-IV* Criteria for Antisocial Personality Disorder: Dimensionality, Local Reliability, and Differential Item Functioning Across Gender

Assessment
1–13

DOI: 10.1177/1073191117745126
journals.sagepub.com/home/asm
SAGE

Muirne C. S. Paap[1], Johan Braeken[2], Geir Pedersen[3,4], Øyvind Urnes[4], Sigmund Karterud[4], Theresa Wilberg[4], and Benjamin Hummelen[4]

Abstract

This study aims at evaluating the psychometric properties of the antisocial personality disorder (ASPD) criteria in a large sample of patients, most of whom had one or more personality disorders (PD). PD diagnoses were assessed by experienced clinicians using the Structured Clinical Interview for *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition, Axis II PDs. Analyses were performed within an item response theory framework. Results of the analyses indicated that ASPD is a unidimensional construct that can be measured reliably at the upper range of the latent trait scale. Differential item functioning across gender was restricted to two criteria and had little impact on the latent ASPD trait level. Patients fulfilling both the adult ASPD criteria and the conduct disorder criteria had similar latent trait distributions as patients fulfilling only the adult ASPD criteria. Overall, the ASPD items fit the purpose of a diagnostic instrument well, that is, distinguishing patients with moderate from those with high antisocial personality scores.

Keywords

antisocial personality disorder, psychopathy, item response theory, conduct disorder, gender bias

Antisocial personality disorder (ASPD), as described by the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*; American Psychiatric Association [APA], 2013), is defined by a set of seven criteria of which at least three must be fulfilled in order to establish the diagnosis. In addition, there should be evidence of conduct disorder (CD) with onset before age 15 years. The term *antisocial personality* was introduced in the *DSM* system in 1968 with the publication of the second edition (*DSM-II*; APA, 1968). According to this manual, a person with antisocial personality is grossly selfish, callous, irresponsible, impulsive, unable to feel guilt or to learn from experience and punishment, and has low frustration tolerance. In the third edition of *DSM* and its revision (APA, 1980, 1987), more emphasis was placed on overt behavior in defining the ASPD criteria, with the intention to obtain greater diagnostic reliability (Widiger et al., 1996). In *DSM-IV* (APA, 1994), ASPD is conceptualized using a "hybrid approach," including criteria that are more personality-oriented and criteria that are more behavior-focused (Widiger et al., 1996; see also Table 1). From *DSM-IV* to *DSM-5* (APA, 2013), the ASPD criteria have not been changed.

Prevalence rates for ASPD in community samples range from 0.2% to 3.6% (Grant et al., 2005; Torgersen, Kringlen, & Cramer, 2001). This broad range in prevalence rates may partly be due to differences in assessment procedures. For instance, Trull, Jahng, Tomko, Wood, and Sher (2010) demonstrated significant reductions of personality disorder (PD) prevalence rates in the study of Grant et al. (2005) by requiring that each PD criterion be associated with significant distress or impairment. In clinical situations, prevalence rates are highly influenced by sample characteristics. For instance, Zimmerman, Rothschild, and Chelminski (2005) found a prevalence of 3.1% in a general clinical outpatient practice, whereas Mariani et al. (2008) found a prevalence of 17.3% in a sample of treatment-seeking cocaine- and cannabis-dependent individuals.

[1]University of Groningen, Groningen, The Netherlands
[2]Centre for Educational Measurement (CEMO), University of Oslo, Oslo, Norway
[3]Norwegian Centre for Mental Disorders Research (NORMENT), University of Oslo, Oslo, Norway
[4]Oslo University Hospital, Oslo, Norway

Corresponding Author:
Muirne C. S. Paap, Department of Special Needs, Education, and Youth Care, Faculty of Behavioural and Social Sciences, University of Groningen, Grote Rozenstraat 38, 9712 TJ Groningen, The Netherlands.
Email: m.c.s.paap@rug.nl

Table 1. ASPD Criteria According to *DSM-IV* and *DSM-5*.

1. Failure to conform to social norms with respect to lawful behaviors as indicated by repeatedly performing acts that are grounds for arrest
2. Deception, as indicated by repeatedly lying, use of aliases, or conning others for personal profit or pleasure
3. Impulsivity or failure to plan ahead
4. Irritability and aggressiveness, as indicated by repeated physical fights or assaults
5. Reckless disregard for safety of self or others
6. Consistent irresponsibility, as indicated by repeated failure to sustain consistent work behavior or honor financial obligations
7. Lack of remorse, as indicated by being indifferent to or rationalizing having hurt, mistreated, or stolen from another

Note. ASPD = antisocial personality disorder. At least three criteria are required for an ASPD diagnosis, in addition to evidence of childhood conduct disorder.

Although the frequency might be relatively low in general outpatient clinics, it is important to assess ASPD reliably and effectively as the presence of ASPD may have important consequences for clinical decision making. ASPD is typically assessed using a subscale of a broader instrument encompassing multiple PDs, like the Structured Clinical Interview for DSM-IV Axis II PDs (SCID-II; First, 1994). It is of yet unclear whether the SCID-II ASPD subscale, which is explicitly based on the *DSM-IV* ASPD criteria, taps into one or multiple underlying factors (e.g., a personality-oriented factor and a behavior-oriented factor). Studies focusing on ASPD or psychopathy (which is a construct closely related to ASPD) have not been consistent with respect to the factorial structure: while some studies found evidence for a one-dimensional structure (Harford et al., 2013; Jane, Oltmanns, South, & Turkheimer, 2007; Rosenström et al., 2017), others found support for two or more factors (Hare & Neumann, 2008; Kendler, Aggen, & Patrick, 2012; Marcus, Lilienfeld, Edens, & Poythress, 2006).

In the assessment of ASPD, another important point of discussion has been whether ASPD should be scored along a continuum or as a categorical diagnosis. Early taxometric studies mostly suggested that ASPD has a latent categorical structure (Haslam, 2003). However, most subsequent taxometric studies found support for a continuum approach (Edens, Marcus, Lilienfeld, & Poythress, 2006; Guay, Ruscio, Knight, & Hare, 2007; Marcus et al., 2006). Such a continuum approach may be helpful from a clinical point of view. Some treatment programs for PDs may tolerate patients with low-grade ASPD but not those who are severely disturbed (Bateman, O'Connell, Lorenzini, Gardner, & Fonagy, 2016). In the PD field, the overall number of PD criteria is often taken as a measure of the general PD severity (Hopwood et al., 2011). This is, however, not an optimal approach since certain criteria may be stronger indicators of PD severity than others. The same would apply to obtaining a score reflecting ASPD severity. An alternative, more suitable approach, would be to estimate latent ASPD severity scores using item response theory models (IRT; Reise & Revicki, 2014); these models have the advantage that they can be used to evaluate item (criterion) properties, and take these properties into account when estimating a latent severity score.

Several authors have suggested that measurement bias across gender might be present in items measuring ASPD, as some items seem to describe more male-specific behavior, for example, "Irritability and aggressiveness, as indicated by repeated physical fights or assaults" (Dolan & Vollm, 2009; Widiger, 1998). Measurement bias across gender can be investigated by testing whether items show differential item functioning (DIF; e.g., Holland & Wainer, 1993) for gender. DIF is present if the item parameters in one group differ from those in the other group (discrimination parameter and/or threshold parameter). In other words, gender-based DIF would imply that men would be more likely (or less likely) to obtain a given item score compared with women who exhibit a similar trait level. Jane et al. (2007) conducted a DIF analysis on the Structured Interview for *DSM-IV* Personality items (Pfohl, Blum, & Zimmerman, 1997), using a nonclinical sample (United States Air Force recruits and undergraduate college students). The respondents were assessed by doctoral-level clinical psychologists and graduate students in clinical psychology. Jane et al. (2007) found DIF for three ASPD items, all focused on behavior: Item 1 (failure to conform), Item 4 (aggressiveness), and Item 5 (reckless disregard). These items were more likely to be endorsed by men than by women with comparable trait levels. The authors concluded that their results "reinforce the possibility that the current ASPD criteria do not adequately reflect how the construct is expressed in women." DIF for behavioral items was also found in a study using the Psychopathy Checklist–Revised, conducted by Bolt, Hare, Vitale, and Newman (2004). In this study, items that belonged to the antisocial/lifestyle domain (Factor 2) were more prone to display DIF than the affective/interpersonal items (Factor 1). This study was based on a sample of criminal offenders, in which female participants may exhibit more male-like antisocial behavior. The generalizability of the results obtained by Jane et al. (2007) and Bolt et al. (2004) has not been sufficiently tested. In this study we aim to extend the literature by carefully assessing whether gender-related DIF is found in the SCID-II ASPD subscale in a large clinical sample; in contrast to the study by Jane et al. (2007), the ASPD criteria were assessed by experienced clinicians.

Among the specific PDs in *DSM-IV* and *DSM-5*, ASPD is the only one that requires the presence of childhood precursors, that is, CD, with onset before age 15 years. Although there seems to be good empirical evidence for the

continuity between CD and ASPD (Gelhorn, Sakai, Price, & Crowley, 2007; Moffitt et al., 2008; Robins, 1978), a substantial number of individuals fulfilling the adult ASPD criteria do not meet criteria for a prior CD diagnosis (Kim-Cohen et al., 2003) and most comparison studies so far have not found clinically significant differences between antisocial individuals with CD and antisocial individuals without CD (Black & Braun, 1998; Perdikouri, Rathbone, Huband, & Duggan, 2007). However, in a study of 327 male prisoners who were assessed by the SCID-II, Walters and Knight (2010) reported that antisocial individuals with evidence of prior CD, showed more severe adult antisocial features, that is, higher levels of criminal thinking, antisocial attitudes, and behavioral adjustment difficulties. Moreover, CD symptom count appeared to have moderate utility in forecasting institutional misconduct in a study of 353 inmates, of whom 185 had ASPD (Edens, Kelley, Lilienfeld, Skeem, & Douglas, 2015). Since the severity of antisocial features is relevant in clinical decision making, it is of special importance to know whether assessing CD symptoms retrospectively may help in determining the severity of ASPD. Since this question appears to be as yet unresolved, more studies are needed, preferably using large clinical samples and a modern psychometric approach.

Aims of the Study

The aim of this study is to perform a psychometric evaluation of the adult *DSM-IV* ASPD diagnostic criteria, as assessed by experienced clinicians using the SCID-II (First, 1994), in a large sample of personality-disordered patients. More specifically, we will examine whether the SCID-II ASPD items are tapping into a common underlying trait, whether the SCID-II ASPD items can be used for reliable measurement, and whether the items are free of measurement bias across gender. Moreover, we will investigate the diagnostic relevance of CD by comparing latent ASPD severity levels obtained by IRT across four diagnostic groups: (1) patients with ASPD according to *DSM-IV* (i.e., ASPD with CD), (2) patients with three or more ASPD criteria without CD (late-onset ASPD), (3) patients without ASPD but with evidence of prior CD, and (4) patients without ASPD and without evidence of prior CD.

IRT (Embretson & Reise, 2000) provides a great framework and toolbox for psychometric evaluation. IRT encompasses a family of measurement models that focuses on explaining the dependencies between item responses within a person and between persons. IRT models are especially suitable for dichotomous or polytomous (e.g., Likert-type scale) item response data, where the items are expected to measure a common latent trait. The reliability of a measurement instrument is usually represented by a single fixed number such as Cronbach's alpha; yet, this in conflict with the fact that a test cannot be expected to measure each person equally efficiently along the latent trait dimension. In IRT, this problem is solved by using (Fisher) information as an estimate of measurement precision/reliability conditional on the latent trait value. This function, showing information for different latent trait values, is known as the *test information function*. Since the goal of the instrument under study is diagnosis, we are interested in having sufficient information for relatively high latent trait values: the focus is on distinguishing patients with moderate levels of antisocial personality from those with high levels (i.e., fulfilling the criteria). Since there has been some debate as to whether the ASPD criteria may focus on behavior more typical for men, we also wanted to check for gender-related item bias or—in IRT terminology—*differential item functioning*. DIF can potentially lead to measurement artefacts by masking or even inflating group differences, because the relationship between an item showing DIF and the latent trait is not identical for individuals belonging to different subgroups.

Method

Sample

The original sample consisted of 3,391 patients from the Norwegian Network of Personality Focused Treatments Programs (Karterud et al., 2003), admitted to treatment from 1996 to 2008 and diagnosed according to *DSM-IV*. Among these patients, 75 had missing criteria sets for the adult ASPD criteria (i.e., the ASPD criteria were not assessed or registered), and two patients had missing criteria sets for childhood CD. Moreover, one patient had a mismatch between ASPD diagnosis and the number of ASPD criteria. All these patients ($N = 78$) were excluded from the analyses, resulting in a sample of 3,313 individuals, of whom 924 were men (28%) and 2,389 were women (72%). Mean age was 37 ($SD = 9.3$) and 35 ($SD = 9.3$) years for men and women, respectively.

All units in the network adhered to the same treatment model, consisting of short-term day treatment followed by long-term outpatient group therapy. All patients in the sample were admitted to day treatment, including those with ASPD. Most patients had a PD diagnosis (77%, $N = 2,595$). Fifty-six percent had one PD diagnosis, 15% had two PD diagnoses, and 6.5% had three or more PD diagnoses. Avoidant PD was the most frequent PD (37%), followed by borderline PD (22%) and PD not otherwise specified (17%). The majority (97%) of patients had one or more symptom diagnoses, mostly an affective disorder (74%) or an anxiety disorder (64%). Other frequent symptom disorders were eating disorder (12%) and substance use disorder (9%).

Chi-square analyses revealed that ASPD was significantly associated with schizotypal PD ($\varphi = .079$, $p < .001$), paranoid PD ($\varphi = .088$, $p < .001$), narcissistic PD ($\varphi = .122$, $p < .001$), and borderline PD ($\varphi = .171$, $p < .001$). The

prevalence of these disorders in the subgroup of patients with ASPD was 7% for schizotypal PD, 29% for paranoid PD, 9% for narcissistic PD, and 76% for borderline PD.

Measures

The SCID-II (First, 1994) is a semistructured clinical interview that covers the 11 *DSM-IV* Personality Disorders, including Personality Disorder not otherwise specified. The SCID-II follows a modular approach, where PDs are assessed one at a time. The initial question for each SCID-II item closely follows the content of the corresponding *DSM-IV* criterion. The SCID-II items are accompanied by open-ended prompts that can be used to encourage patients to elaborate freely about their symptoms. At times, open-ended prompts can be followed by closed-ended questions to further clarify a specific PD symptom. In the current study, the focus is on the ASPD subscale, which consists of 7 items. The SCID-II items are rated within one of three response categories: 1 = *absent or false*; 2 = *subthreshold* (i.e., the threshold for the criterion is almost but not quite, met); and 3 = *threshold or true*. In order to establish a *DSM-IV* ASPD diagnosis, it is required that the patient is also (retrospectively) diagnosed with childhood CD. The diagnosis of CD was made when at least three CD criteria were met. The SCID-II does not require that these criteria are confirmed by early caregivers or other sources of information. Interrater reliability studies have shown that adequate interrater reliability can be obtained by using the SCID-II (Maffei et al., 1997; Weertman, Arntz, Dreessen, van Velzen, & Vertommen, 2003).

Procedures

All units in this study complied with the diagnostic and data collection procedures required for membership of the Norwegian Network. The SCID-II was administered by experienced clinicians, that is, health care professionals (mental health nurses, psychologists, or medical doctors) working at clinical units specialized in the assessment and treatment of PDs. Clinicians were trained in PD diagnostics through attendance at local courses and Network conferences. Final PD diagnoses were established by way of the *longitudinal expert evaluation using all data* (LEAD) standard (Spitzer, 1983). Tentative diagnoses were made at the time of admission, on the basis of referral letters, self-reported history and complaints, as well as two structured clinical diagnostic interviews: (1) Mini-International Neuropsychiatric Interview for Axis I diagnoses (Sheehan et al., 1994) and (2) SCID-II for PDs (First, 1994). During the 18 weeks of day treatment, therapists could affirm or review diagnoses based on information gathered in a variety of clinical situations. A final PD diagnosis required that the criteria from the original SCID-II protocol were confirmed by clinical observations. It is assumed that the LEAD procedure resulted in more valid diagnoses (Pedersen, Karterud, Hummelen, & Wilberg, 2013).

Psychometric Analyses

Dimensionality Analyses. To ascertain whether the SCID-II ASPD items form a scale and thus measure one underlying trait, we assessed the dimensionality of the SCID-II ASPD items using two complementary methods: confirmatory Mokken Scale Analysis (MSA), which is a nonparametric method; and the Empirical Kaiser Criterion (EKC), which is an eigenvalue-based method. The dimensionality analyses were run for the total sample first, followed by separate analyses by gender.

In recent years, MSA has increased in popularity in the fields of psychological and health assessment (e.g., Chou, Lee, Liu, & Hung, 2017; Lenferink et al., 2016; Murray, McKenzie, Murray, & Richelieu, 2014; Stewart, Allison, Baron-Cohen, & Watson, 2015; van den Berg, Paap, & Derks, 2013; Watson et al., 2012). MSA identifies scales that allow an ordering of individuals on an underlying scale using unweighted sum scores. In order to ascertain which items covary and form a scale, scalability coefficients are calculated on three levels: item-pairs (H_{ij}), items (H_i), and scale (H). H is based on H_i and reflects the degree to which the scale can be used to reliably order persons on the latent trait using their sum score. A scale is considered acceptable if $0.3 \leq H < 0.4$, good if $0.4 \leq H < 0.5$, and strong if $H \geq 0.5$ (Mokken, 1971; Sijtsma & Molenaar, 2002).

Eigenvalue-based methods are among the most popular and common methods for dimensionality assessment. Unfortunately, possibly due to historical and/or ease-of-access reasons, many applied researchers still rely on flawed criteria. In particular, the eigenvalue-greater-than-1 rule, also known as the Kaiser criterion (Kaiser, 1960), has repeatedly been shown to have low accuracy (observe that this is not a recent finding; see, e.g., Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986). Braeken and van Assen (2017) clarify that the reason why the Kaiser criterion fails is that it does not account for sampling variation in eigenvalues. To remedy this shortcoming, they proposed a modification based on the asymptotical sampling distribution of eigenvalues. Instead of comparing the observed sample eigenvalues to a fixed reference value of 1, the EKC establishes reference eigenvalues that can be expected for a data set of specified size (i.e., persons by items), if no factor structure would be present. The number of dimensions to retain then corresponds to the length of the series of first-ranked eigenvalues that are all greater than these null-reference eigenvalues. Graphically, this simply means finding the point where the line formed by the reference eigenvalues crosses the screeplot of observed sample eigenvalues (for an easy-to-use webapplet, see https://cemo.shinyapps.

io/EKCapp). The EKC is a non–simulation-based relative of parallel analysis (which simulates null reference eigenvalues), the current gold standard in the field (Garrido, Abad, & Ponsoda, 2013; Timmerman & Lorenzo-Seva, 2011). Simulation studies show that the EKC performs at par with parallel analysis for uncorrelated scales, and even better than parallel analysis for short correlated scales.

IRT Model. The graded response model (GRM; Samejima, 1996) was used to scale and evaluate the seven SCID-II ASPD items. The GRM applies to ordered categorical item scores. Let the variable Y_{pi} represent the score of a patient p on an item i, where the observed response Y_{pi} can range from $j = 1$ over 2 to 3. The GRM directly models the cumulative conditional probability of scoring greater than or equal to each of the response options

$$\Pr\left(Y_{pi} \geq j \mid \theta_p\right) = \frac{1}{1 + exp\left(-a_i\left[\theta_p - b_{ij}\right]\right)}$$

where θ_p is the position of the person on the latent trait scale and where a_i and b_{ij} are item parameters describing how the item is linked to the latent trait scale. The item parameter a_i is a discrimination parameter expressing the degree to which the item i can differentiate between patients on the latent trait scale (i.e., higher values for a_i indicate that small differences in position on the latent trait can lead to large changes in probability). Item parameter b_{ij} is a threshold parameter for item i indicating the position on the latent trait scale for which a patient would have 50% probability of being assigned a score greater than or equal to j on the item i. The regular response probabilities can then simply be derived by taking differences between the cumulative probabilities:

$$\Pr\left(Y_{pi} = j \mid \theta_p\right) = \Pr\left(Y_{pi} \geq j \mid \theta_p\right) - \Pr\left(Y_{pi} \geq j+1 \mid \theta_p\right)$$

Written out in full, this implies the following set of three category response curves:

$$\Pr\left(Y_{pi} = 1 \mid \theta_p\right) = 1 - \Pr\left(Y_{pi} \geq 2 \mid \theta_p\right)$$

$$\Pr\left(Y_{pi} = 2 \mid \theta_p\right) = \Pr\left(Y_{pi} \geq 2 \mid \theta_p\right) - \Pr\left(Y_{pi} \geq 3 \mid \theta_p\right)$$

$$\Pr\left(Y_{pi} = 3 \mid \theta_p\right) = \Pr\left(Y_{pi} \geq 3 \mid \theta_p\right) - 0$$

Note that $\Pr\left(Y_{pi} \geq 1 \mid \theta_p\right) = 1$, because everyone will at least get $j = 1$, which is the lowest score that can be assigned to a patient. Hence, similar to the dummy coding principle for a categorical predictor, the number of threshold parameters for an item is always one less than the item's number of response categories. The item score range stops at 3, so by definition $\Pr\left(Y_{pi} \geq 3+1 \mid \theta_p\right) = 0$.

Local Reliability: Test Information Function and Targeting. In IRT, measurement error is conceptualized in terms of information: More information means more precision, meaning less error of measurement. The information a test provides on the scale-position of a patient varies across the latent trait scale and is a direct function of the psychometric properties and scale-position of the items in the test. Given that the squared standard error of measurement $SE\left(\theta_p\right)^2$ is equal to the reciprocal of the test information $\mathrm{I}\left(\theta_p\right)$, an estimate of local reliability can be computed as

$$r\left(\theta_p\right) = 1 - \frac{SE\left(\theta_p\right)^2}{VAR\left(\theta_p\right)} = 1 - \frac{1}{I\left(\theta_p\right)}$$

The first equality stems from the traditional formulation of reliability as a ratio of variances, true variance divided by total variance, or equivalently, 1 minus error variance divided by total variance. The second equality stems from the reciprocal information-error relation and the fact that our scale metric in a GRM is standardized such that $VAR\left(\theta_p\right) = 1$.

Measurement Bias Across Gender: Differential Item Functioning. We used a DIF model comparison approach[1] to screen for gender-related item bias in the seven SCID-II ASPD items. For more detailed information about this procedure as well as other ways to assess DIF, we refer the reader to Thissen, Steinberg, and Wainer (1993) and Millsap (2011). Two reference models were estimated: a gender equivalent model and a gender nonequivalent model. The gender-equivalent model allows for scale-level differences in means and standard deviations of the latent trait between male and female patients, while constraining the item parameters to be equal across groups; in contrast, the gender-nonequivalent model allows for differences in both item discrimination and item thresholds for all items between male and female patients, while constraining the means and standard deviations to be equal across groups. If the gender-equivalent model shows better fit compared with the nonequivalent model, this would imply that there are only overall scale-level group differences between males and females, whereas if the opposite is true, it would imply that the scales for males and females are to some extent incomparable and that ASPD criteria may function differently for males and females. If the gender-nonequivalent model shows better fit, a set of model comparisons are performed with the goal to establish which items cause the nonequivalence. This is done by taking the gender-equivalent model as a starting point, and relaxing the equivalence constraints, one item at a time. When DIF items have been identified in this manner, Wald tests are used to assess whether the DIF is uniform across the scale (i.e., whether it only affects the thresholds) or also varies across the scale (i.e., also affects the discrimination parameters; nonuniform DIF).

Software. All statistical analyses were coded and performed in the open source software program R version 3.2.3 (R Development Core Team, 2012). The GRM was estimated using a full information maximum likelihood approach in the R package mirt version 1.16 (Chalmers, 2012).

Results

Descriptive Statistics: Sample Prevalence and Gender Distribution of ASPD

Of the total sample of 3,313 patients (72% women), 108 patients scored a "3" on three or more ASPD items (48% women). Fifty-four of these patients (42% women) also fulfilled the criteria for childhood CD and were therefore diagnosed as having ASPD according to the *DSM-IV* (labeled as "ASPD-*DSM-IV*"). The 54 patients (55% women) who did not fulfill the CD criteria were tentatively labeled as *ASPD-late onset*.

Since there were 3 answering categories per item and 7 items, the total number of possible scoring patterns equaled 3^7 = 2,187. There were only 415 unique ASPD symptom endorsement patterns (19% of 2,187), which is typical when studying a clinical diagnosis. One pattern had the highest frequency of occurrence by far: the pattern 1111111 (i.e., absence of all ASPD-related symptoms) occurred 1,845 times. This indicates that SCID-II ASPD items are not too commonly endorsed and can be expected to differentiate well between patients with and without ASPD. Furthermore, among the 108 patients scoring "3" on three or more ASPD items, 90 different endorsement patterns occurred of which 74 were reported by a single patient only. Hence, there is no prototypical ASPD endorsement pattern.

Dimensionality of the SCID-II ASPD Items

The *H* values exceeded the threshold of .3 for all MSA analyses (.371 and .336 for women and men, respectively, and .303 for the total sample). For the total sample, all but one H_i value exceeded .3; for the remaining item an H_i value of .295 was found. Taken together, these findings provide support for a weak to acceptable unidimensional scale.

Figure 1 shows the screeplot accompanying the EKC results for the total sample. A sharp drop can be observed between the first and second component. Furthermore, the observed eigenvalue λ was higher than the reference value only for the first component (total sample: $\lambda_1 = 2.71 > EKC_1 = 1.03$, $\lambda_2 = .88 < EKC_2 = 1.00$; men: $\lambda_1 = 2.84 > EKC_1 = 1.18$, $\lambda_2 = .96 < EKC_2 = 1.00$; women: $\lambda_1 = 2.55 > EKC_1 = 1.11$, $\lambda_2 = .86 < EKC_2 = 1.00$). The EKC findings show very clear support for a unidimensional solution.

Since both the MSA and EKC results provided support for a unidimensional scale, the IRT analyses were performed using the unidimensional GRM, which showed good model fit (root mean square error of approximation = .04; Tucker-Lewis index = .98; comparative fit index = .99).

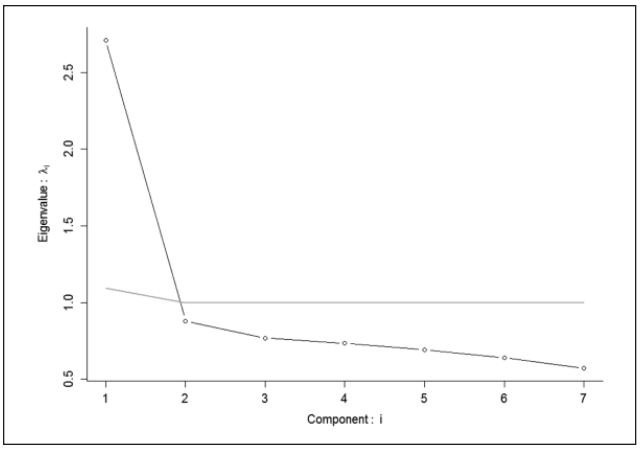

Figure 1. Scree plot of eigenvalues. Empirical Kaiser criterion reference line depicted in gray.

Table 2. Item Parameters Based on the Graded Response Model for the SCID-II ASPD Items.

Item *i*	Item content	*a*	b_{i2}	b_{i3}
#1	Failure to conform	2.33	1.22	1.77
#2	Deceitfulness	2.06	1.83	2.59
#3	Impulsivity	1.87	1.35	2.14
#4	Aggressiveness	1.53	1.50	2.62
#5	Reckless disregard	1.60	1.13	1.95
#6	Irresponsibility	1.82	1.65	2.64
#7	Lack of remorse	2.35	1.92	2.62

Note. SCID-II = Structured Clinical Interview for *Diagnostic and Statistical Manual of Mental Disorders*, 4th Edition, Axis II PD; ASPD = antisocial personality disorder. *a* = estimated discrimination parameter; b_{ij} = estimated threshold parameter for item *i* indicating the position on the latent trait scale for which a patient would have 50% probability of being assigned a score greater than or equal to *j*. Following the SCID-II manual, responses were coded as 1, 2, or 3. Since the probability of scoring in Category 1 or higher equals 1, only b_{i2} and b_{i3} are reported.

Item Parameters and Local Reliability

The item parameters estimated with the GRM are reported in Table 2. Characteristic of clinical settings (Reise & Waller, 2009), the discrimination parameters are fairly high, and so are the threshold values; in other words, the items discriminate well but mostly at the upper range of the latent trait scale. This finding is also illustrated by the test information function, which shows that the highest information (local reliability) is found for latent trait values between 1 and 3 (see Figure 2). Hence, this is the zone best targeted by the test where we can differentiate between patients with a high degree of precision. This matches well with the

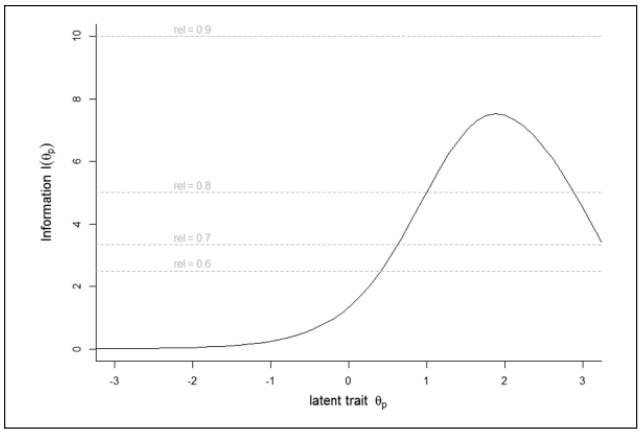

Figure 2. Test information function based on the graded response model, with estimated latent trait values (θ_p) on the *x*-axis, and information conditional on θ_p on the *y*-axis. Four lines have been drawn horizontally to indicate which information scores correspond to a reliability estimate of 0.6, 0.7, 0.8, and 0.9, respectively.

purpose of a diagnostic instrument: distinguishing patients with moderate from those with high antisocial personality scores.

The category response curves, which depict the probability of choosing a particular response category as a function of the latent trait (here: antisocial personality), clearly indicate that the middle category (Category 2) does not get endorsed often; it hardly ever has a higher probability of being chosen compared with Category 1 or 3. This is the case for all items. Figure 3 contrasts one of the empirical category response curve sets from our study to a hypothetical ideal set where all categories contribute information. The corresponding item information curves in Figure 4 illustrate that hypothetically, polytomous items have the potential to provide information across a wider range of the latent trait (i.e., multiple thresholds imply multiple peaks) when compared with dichotomous items (i.e., only one threshold = maximally one peak). Our data illustrate, however, that polytomous items where one of the categories is hardly ever the dominant category may not have much added value as compared with dichotomous items in estimating a patient's position on the latent trait scale.

To compare the *DSM-IV* scoring rule with IRT-based scoring, we calculated the latent trait score distributions for patients scoring below and above the *DSM-IV* ASPD cutoff rule (a "3" on at least three items). Figure 5 shows the latent trait distributions for all possible number of 3s: from zero to seven. As expected, the means of the latent trait increase as the number of items on which a "3" is scored increases. However, the figure also indicates that there is still some variability in IRT-based scores within most of the groups. If we focus on the groups near the *DSM-IV* cutoff, it can be seen that there is still quite some overlap in score distributions. More specifically, for persons with exactly three 3s, the specific items on which they score these 3s matter when it comes to calculating their IRT-based scores. Looking at Table 2, we can see that the b_{i3} -values for Items 1, 5, and 3 are markedly lower than those for Items 4, 6, and 7; this means that scoring a "3" on Items 1, 5 and 3 would result in a substantially lower latent trait score than scoring a "3" on Items 4, 6, and 7.

Differential Item Functioning Across Gender

The first step in examining whether there was DIF for any of the items was comparing the gender-equivalent model (item parameters constrained to be equal) with the gender-nonequivalent model (unconstrained item parameters, equal means and standard deviations). Table 3 shows all the models that were estimated, and which model comparisons were made. The gender-equivalent model is used as the reference model in most cases and was therefore labeled as Model 0. The gender-nonequivalent model showed a significantly better fit compared with the equivalent model. Further model comparisons indicated that models in which the item parameters of Items 3 (impulsivity) and 5 (reckless disregard) were unconstrained (free to vary over groups) showed a significantly better fit than the equivalent model. This indicates that there was DIF for these items. The model where the item parameters of both these items were free to vary across groups was not significantly different from the nonequivalent model. This indicates that it was sufficient to relax the equivalence constraints on the item parameters for these two items (and constrain the other item parameters to be equal across groups).

The Wald tests showed that there was no evidence for nonuniform but only for uniform DIF. In other words, the DIF only affected the thresholds and not the discrimination parameters (Item 3: Δa = .24, p = .189; Item 5: Δa = .13, p = .246). The thresholds for Item 3 were higher for male patients (Δb = .33/.16, p = <.001/.006), whereas the thresholds for Item 5 were lower for male patients (Δb = −.53/−.55, p = <.001/<.001). To facilitate understanding of the effect size of these parameter differences we calculated them in terms of response probabilities as well (the difference between the category response curves). Females had on average a probability that was .07 higher than that of males (with similar θ_p scores) to score in Category 3 on Item 3, with a maximum probability difference of .17. DIF on Item 5 was associated with an average probability difference in favor of males of .14 to score in Category 3, with a maximum probability difference of .21. Summarizing, for a given latent trait level, female patients were more likely to be perceived as being impulsive (Item 3), while male patients were more likely to be perceived as being reckless (Item 5).

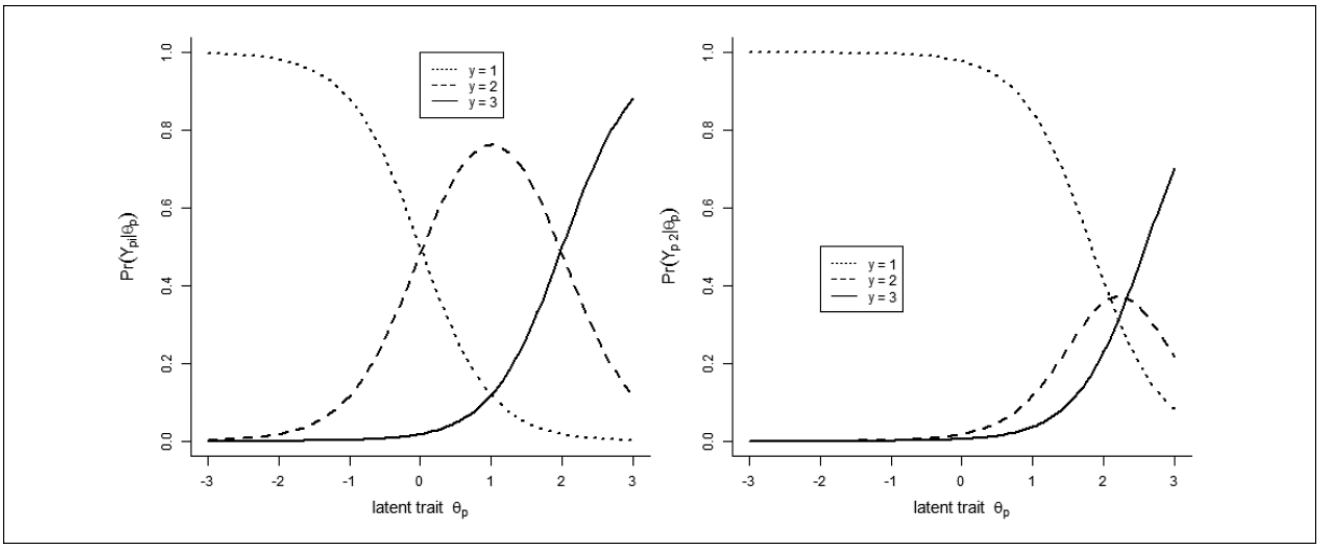

Figure 3. Category characteristic curves for (1) a hypothetical item where all categories contribute information (Left) and (2) Item 2 (deceitfulness), which shows that Category 2 hardly contributes any information (Right). In the left plot, each category is the most dominant one (highest probability of being selected) for a range of latent trait values; in the right plot Categories 1 and 3 clearly dominate Category 2.

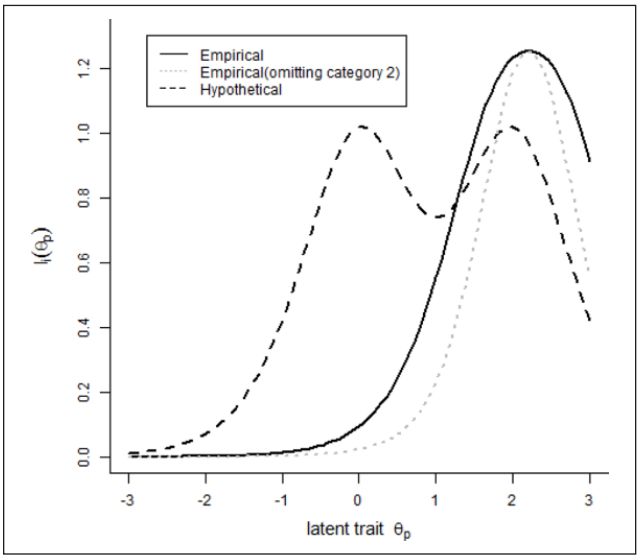

Figure 4. Item information functions for (1) Item #2 (deceitfulness) using all three categories (solid line), (2) the same item but now ignoring Category 2 (gray dotted line), and (3) a hypothetical item where all categories contribute information (dashed black line).

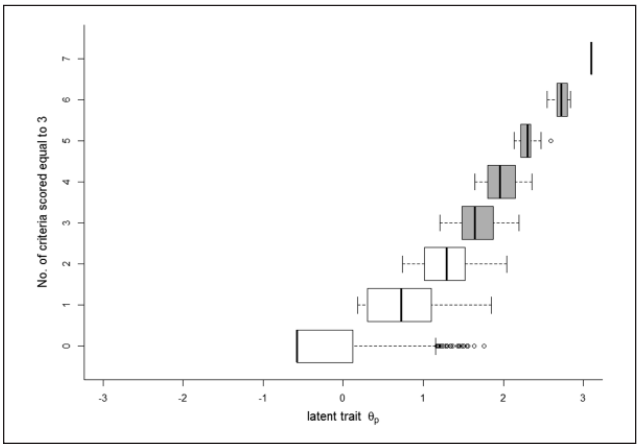

Figure 5. Boxplots showing the distribution of latent trait values (θ_p) for seven subgroups in the sample; patients were assigned to these subgroups on the basis of the number of criteria they scored a "3" on. Note that at least three 3s are needed in order to qualify for a diagnosis of antisocial personality disorder (cutoff). To facilitate interpretation, the boxplots for patients scoring above the cut-off are printed in gray.

Using an IRT model that ignored DIF (constraining the item parameters to be equal across groups), resulted in a lower mean θ_p for females compared with males ($\Delta = -.56$, $p < .001$). After having corrected for DIF, the group difference was somewhat smaller but still significantly different from zero ($\Delta = -.52$, $p < .001$). No difference was found in variance.

Diagnostic Relevance of Conduct Disorder

Finally, we compared latent trait distributions for four diagnostic groups: (1) patients with three or more ASPD criteria and CD (*ASPD-CD*), (2) patients with three or more ASPD criteria without CD (*ASPD-late onset*), (3) patients with fewer than three ASPD criteria with CD (*CD-only*), and (4) patients with fewer than three ASPD criteria and absence of CD (*noASPD-noCD*). The results

Table 3. Overview of the Differential Item Functioning Model Comparison Results.

	Model	LL	Reference model used for comparison	*df*	χ	*p*
0	*Equivalent*	H-				
1	*Nonequivalent*	−9,818	Equivalent	19	97.39	<.001
2	DIF: Item #1	−9,866	Equivalent	3	1.80	.615
3	DIF: Item #2	−9,866	Equivalent	3	.98	.807
4	DIF: Item #3	−9,843	Equivalent	3	47.82	<.001
5	DIF: Item #4	−9,864	Equivalent	3	4.88	.181
6	DIF: Item #5	−9,837	Equivalent	3	59.20	<.001
7	DIF: Item #6	−9,866	Equivalent	3	.99	.803
8	DIF: Item #7	−9,866	Equivalent	3	.95	.812
9	DIF: Items #3 and #5	−9,822	Equivalent	6	89.38	<.001
			Nonequivalent	13	8.02	.843

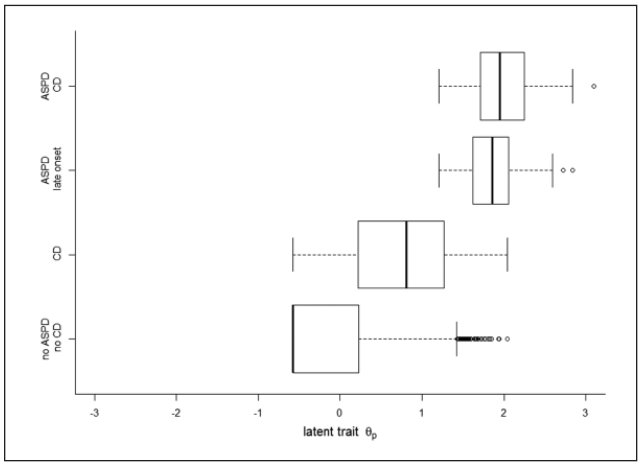

Figure 6. Boxplots showing the distribution of latent trait values (θ_p) for four diagnostic subgroups in the sample.

are displayed in Figure 6. Using the *noASPD-noCD* as a reference group, the following 95% confidence intervals were found: [1.2, 1.6] for *CD-only*; [2.5, 3.8] for *ASPD-late onset*; and [2.8, 3.9] for *ASPD-CD*. If the confidence interval contains the value 0, this means that they do *not* significantly differ from the reference group with respect to average latent trait score. If the confidence intervals overlap, this indicates that the groups in question are *not* significantly different from each other. In this case, the only confidence intervals overlapping were those of ASPD-late onset and ASPD-CD. This can also be seen in Figure 6: the score distributions of the two ASPD groups are both clearly situated at the high end of the latent trait continuum, followed by the *CD-only* group and finally the *noASDP-noCD* group which was placed around the midpoint of the scale. Notice that the location of the peak of the TIF matches well with the location on the trait scale of the two ASPD groups.

Discussion

This study of a large clinical sample sought to examine the psychometric properties of the ASPD criteria as defined by *DSM-IV* and assessed by the SCID-II (First, 1994). The results of the analyses indicate that ASPD is a unidimensional construct that can be measured reliably at the upper range of the latent trait scale. There was some DIF across gender, but this had little impact on the latent ASPD trait level and was restricted to two items, that is, Items 3 (impulsivity) and 5 (reckless disregard). Patients with three or more ASPD criteria without CD (*ASPD-late onset*) had similar levels of the underlying antisocial dimension as ASPD according to *DSM-IV* (*ASPD-CD*).

Our IRT analyses showed that the SCID-II ASPD items had good item discrimination and covered the upper range of the latent trait scale, from 1 *SD* above the mean to 2.5 *SD*s above the mean. This fits the purpose of the *DSM* criteria well: differentiating among people with and without the disorder. If the aim would be to differentiate along the entire scale (low from average severity, average from high, etc.), new items would have to be added with lower item threshold values.

In accordance with previous studies (Bolt et al., 2004; Cooke & Michie, 1997; Jane et al., 2007), items that revealed DIF were more behavior focused. More specifically, female patients were more likely to be considered as impulsive compared with male patients with similar ASPD scores. Male patients, on the other hand, were more likely to be considered as reckless compared with female patients with similar ASPD scores. Importantly, the effect of the DIF we found did not have a substantial effect on latent trait scores at the group level. This is in line with the statement made by Reise and Waller (2009, p. 38): " . . . the presence of item-level DIF does not necessarily lead to bias at the level of scale scores." Nevertheless, we concur with Reise and Waller (2009) and Orlando and Marshall (2002) that it

is important to test for and detect DIF rather than to ignore potential DIF-related problems, even if DIF does not always lead to bias at the scale/group level. In our sample, there was a marked gender imbalance (72% women). Ideally, from a statistical viewpoint, one would prefer to have equal group sizes when studying DIF. However, if both groups are sufficiently large, the group imbalance has much less impact than it would have in smaller samples. In our study, the smallest group was still quite large (N = 924), so we are confident that we had sufficient power to detect DIF.

As mentioned in the introduction, Jane et al. (2007) found DIF for three of the seven adult ASPD criteria. Of these three items, only one showed DIF in our sample: recklessness. In contrast to Jane and colleagues, we also found DIF for impulsivity (where they found none). Another difference is that the DIF found by Jane and colleagues all went in the same direction: men were more likely to endorse the DIF items than women (for similar trait levels). However, in our study the two DIF items had opposite directionality. DIF of the recklessness item could be explained by the operationalization of the ASPD items in the SCID-II, that is, the recklessness questions in the SCID-II are focused on driving behavior and unsafe sex, which might be considered as examples of male-like behavior. DIF of the impulsivity item might be explained by the fact that our study concerns a clinical sample with a high prevalence of borderline PD. High comorbidity rates between ASPD and borderline PD is a common phenomenon in clinical samples of patients with severe personality pathology (Bateman et al., 2016). Overall, our results do not support the assertion of Jane et al. (2007) that the current ASPD criteria do not adequately reflect how the construct is expressed in women. In clinical populations, measurement bias across gender may be less prominent, at least when assessed by experienced clinicians using a structured clinical interview.

For patients with high latent antisocial trait values, being diagnosed with CD did not lead to a further increase in trait levels. This finding corroborates earlier reports from clinical samples that did not find clinically significant differences between antisocial patients with and without CD (Black & Braun, 1998; Perdikouri et al., 2007). However, this finding is at odds with the study of Walters and Knight (2010), who found that the presence of CD was associated with more severe antisociality. This discrepancy might be explained by methodological differences, since Walters and Knight included a more comprehensive assessment of antisocial features, for example, criminal thinking style and egocentricity. It might also be due to sample differences, that is, a forensic sample with only male individuals versus a clinical sample with predominantly female patients of whom most had a PD.

In our sample, the C*D-only* patients (patients with childhood CD but without ASPD) had higher levels of the latent ASPD trait than the *noASPD-noCD* patients. These results suggest that even though the majority of children with CD may not develop a "full-blown" ASPD (Robins, 1978), they are still at risk for developing antisocial traits. Moreover, the ASPD-related traits and behavior of what we labeled the *ASPD-late onset* group may not be adequately addressed/recognized since these patients did not receive a formal diagnosis (in spite of their high latent scores). It should be kept in mind, however, that the CD criteria were assessed retrospectively. It is uncertain to what degree a retrospective CD diagnosis accurately reflects the presence of CD during childhood. Furthermore, the CD criteria were assessed in concordance with the SCID-II interview, which requires the presence of at least three CD criteria. In *DSM-IV* and *DSM-5*, however, the required number of CD criteria is not explicitly specified.

It is an implicit assumption, supported by empirical research, that personality traits lie along a continuum (Widiger & Simonsen, 2005). Accordingly, in the SCID-II (First, 1994), personality traits are rated within one of three response categories: 1 = not present/do not fulfill; 2 = partly true/subthreshold; and 3 = personality trait present. IRT provides a means of analyzing how subthreshold scores may be helpful in assessing ASPD. By using the graded response model, we found that the scoring of subthreshold criteria did not result in additional/richer information compared with using Categories 1 and 3 only. Since the current version of the SCID-II is not accompanied by guidelines as to how subthreshold diagnostic values should be scored, the use of subthreshold values might have been confounded by how the diagnostic rules were used by the clinicians participating in this study. For example, it may be that clinicians assessed the middle category less carefully, since clear guidelines are lacking. Another possibility is that subthreshold scores may have been used intentionally at times, to avoid setting an ASPD diagnosis. We suggest that in future versions of the SCID-II, items should either be rated dichotomously, or there should be clear rules regarding the use of a middle category (i.e., they should be taken into account in the diagnostic process). A recent study (Huprich, Paggeot, & Samuel, 2015) used IRT analyses to compare the SCID-II borderline personality disorder scale to the corresponding scale in the Personality Disorder Interview for *DSM-IV* (PDI-IV; Widiger, Mangine, Corbitt, Ellis, & Thomas, 1995). For both interviews, items are scored on a 3-point Likert-type scale; but in contrast to the SCID-II, the PDI-IV is accompanied by explicit scoring guidelines also for the middle category. The middle category takes on a different meaning in the PDI-IV as compared with how it is treated in the SCID-II, however. The middle response options from the PDI-IV are verbally most similar to the highest response option for the SCID-II: They indicate the presence of a criterion, as does the highest response category of the SCID-II. This was also reflected by the IRT parameters that were

found for the two measures: SCID-II items allowed for higher precision in the subthreshold range, whereas the PDI-IV items covered a broader range of latent trait values (and thus provided more information about individuals scoring above the diagnostic threshold as compared with the SCID-II). Although these findings are highly interesting, and show that the choice of diagnostic interview can influence how disorders are diagnosed, it is important to keep in mind that the middle category may not be used in a consistent manner for the SCID-II; which may have influenced the results.

In the past decades, increased recognition of the limitations of the categorical approach paired with an increasing body of empirical evidence supporting the continuum approach has resulted in a call for abandoning the categorical model in favor of a continuous one (Hopwood et al., 2011; Tyrer et al., 2011; Widiger & Simonsen, 2005). In the alternative *DSM-5* model for PDs (APA, 2013), continuous scores are reported in addition to categorical ones (Skodol, Morey, Bender, & Oldham, 2015). This approach can be supported by IRT analysis. Taking our results as an example, one would have information about whether or not the formal criteria for ASPD were fulfilled as well as to what degree a patient scored relatively high or low on an antisocial trait/behavior continuum.

In sum, the results of our study suggest that ASPD can be measured with minimal measurement bias across gender in clinical samples—at least when assessed by experienced clinicians using the SCID-II. Overall, the SCID-II ASPD items appear to fit the purpose of the *DSM* well, that is, differentiating among persons in the upper ranges of the latent trait continuum (ASPD). If the aim would be to differentiate among individuals with less severe antisocial personality features, items would have to be added with lower item threshold values. Finally, we did not find differences in score distributions between *ASPD-CD* and *ASPD-late onset* groups. In other words, these two groups show similarly high levels of antisocial behavior and antisocial traits, irrespective of childhood diagnosis of CD.

Acknowledgments

We wish to thank the patients and staff from the Norwegian Network of Personality-Focused Treatment Programs for their contribution to this study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. We preferred a fully modeled based approach to DIF assessment over for instance Mantel–Haenszel tests or logistic regression, as it avoids having to use a proxy variable for the latent trait scores. Furthermore, the IRT modelling approach to DIF is generally better understood and studied, and potential DIF findings can be directly related to the underlying latent dimension(s).

References

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: Author.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd rev. ed.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Bateman, A., O'Connell, J., Lorenzini, N., Gardner, T., & Fonagy, P. (2016). A randomised controlled trial of mentalization-based treatment versus structured clinical management for patients with comorbid borderline personality disorder and antisocial personality disorder. *BMC Psychiatry*, *16*(1), 304.

Black, D. W., & Braun, D. (1998). Antisocial patients: A comparison of those with and those without childhood conduct disorder. *Annals of Clinical Psychiatry*, *10*(2), 53-57.

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A multigroup item response theory analysis of the Psychopathy Checklist–Revised. *Psychological Assessment*, *16*, 155-168.

Braeken, J., & van Assen, M. A. (2017). An empirical Kaiser criterion. *Psychological Methods*, *22*, 450-466. doi:10.1037/met0000074

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29.

Chou, Y. H., Lee, C. P., Liu, C. Y., & Hung, C. I. (2017). Construct validity of the Depression and Somatic Symptoms Scale: Evaluation by Mokken scale analysis. *Neuropsychiatric Disease and Treatment*, *13*, 205-211. doi:10.2147/ndt.s118825

Cooke, D. J., & Michie, C. (1997). An item response theory analysis of the Hare Psychopathy Checklist–Revised. *Psychological Assessment*, *9*(1), 3-14.

Dolan, M., & Vollm, B. (2009). Antisocial personality disorder and psychopathy in women: A literature review on the reliability and validity of assessment instruments. *International Journal of Law & Psychiatry*, *32*(1), 2-9.

Edens, J. F., Kelley, S. E., Lilienfeld, S. O., Skeem, J. L., & Douglas, K. S. (2015). DSM-5 antisocial personality disorder: Predictive validity in a prison sample. *Law and Human Behavior*, *39*, 123-129.

Edens, J. F., Marcus, D. K., Lilienfeld, S. O., & Poythress, N. G., Jr. (2006). Psychopathic, not psychopath: Taxometric evidence for the dimensional structure of psychopathy. *Journal of Abnormal Psychology*, *115*(1), 131-144.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

First, M. B. (1994). *Structured clinical interview for DSM-IV Axis II personality disorders (SCID II)*. New York: New York State Psychiatric Institute.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, *18*, 454-474. doi:10.1037/a0030005

Gelhorn, H. L., Sakai, J. T., Price, R. K., & Crowley, T. J. (2007). DSM-IV conduct disorder criteria as predictors of antisocial personality disorder. *Comprehensive Psychiatry*, *48*, 529-538.

Grant, B. F., Hasin, D. S., Stinson, F. S., Dawson, D. A., Chou, S. P., Ruan, W. J., & Huang, B. (2005). Co-occurrence of 12-month mood and anxiety disorders and personality disorders in the US: Results from the national epidemiologic survey on alcohol and related conditions. *Journal of Psychiatric Research*, *39*(1), 1-9.

Guay, J. P., Ruscio, J., Knight, R. A., & Hare, R. D. (2007). A taxometric analysis of the latent structure of psychopathy: Evidence for dimensionality. *Journal of Abnormal Psychology*, *116*, 701-716.

Hare, R. D., & Neumann, C. S. (2008). Psychopathy as a clinical and empirical construct. *Annual Review of Clinical Psychology*, *4*, 217-246. doi:10.1146/annurev.clinpsy.3.022806.091452

Harford, T. C., Chen, C. M., Saha, T. D., Smith, S. M., Hasin, D. S., & Grant, B. F. (2013). An item response theory analysis of DSM-IV diagnostic criteria for personality disorders: Findings from the national epidemiologic survey on alcohol and related conditions. *Personality Disorders: Theory, Research, & Treatment*, *4*(1), 43-54.

Haslam, N. (2003). The dimensional view of personality disorders: A review of the taxometric evidence. *Clinical Psychology Review*, *23*(1), 75-93.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hopwood, C. J., Malone, J. C., Ansell, E. B., Sanislow, C. A., Grilo, C. M., McGlashan, T. H., . . . Morey, L. C. (2011). Personality assessment in DSM-5: Empirical support for rating severity, style, and traits. *Journal of Personality Disorders*, *25*, 305-320.

Huprich, S. K., Paggeot, A. V., & Samuel, D. B. (2015). Comparing the Personality Disorder Interview for DSM-IV (PDI-IV) and SCID-II borderline personality disorder scales: An item-response theory analysis. *Journal of Personality Assessment*, *97*(1), 13-21. doi:10.1080/00223891.2014.946606

Jane, J. S., Oltmanns, T. F., South, S. C., & Turkheimer, E. (2007). Gender bias in diagnostic criteria for personality disorders: An item response theory analysis. *Journal of Abnormal Psychology*, *116*, 166-175.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141-151. doi:10.1177/001316446002000116

Karterud, S., Pedersen, G., Bjordal, E., Brabrand, J., Friis, S., Haaseth, Ø., Haavaldsen, G., Irion, T., Leirvåg, H., Tørum, E., & Urnes, Ø. (2003). Day hospital treatment of patients with personality disorders. Experiences from a Norwegian treatment research network. *Journal of Personality Disorders*, *17*, 173–193. doi:10.1521/pedi.17.3.243.22151

Kendler, K. S., Aggen, S. H., & Patrick, C. J. (2012). A multivariate twin study of the DSM-IV criteria for antisocial personality disorder. *Biological Psychiatry*, *71*, 247-253. doi:10.1016/j.biopsych.2011.05.019

Kim-Cohen, J., Caspi, A., Moffitt, T. E., Harrington, H., Milne, B. J., & Poulton, R. (2003). Prior juvenile diagnoses in adults with mental disorder: Developmental follow-back of a prospective-longitudinal cohort. *Archives of General Psychiatry*, *60*, 709-717.

Lenferink, A., Effing, T., Harvey, P., Battersby, M., Frith, P., van Beurden, W., . . . Paap, M. C. S. (2016). Construct validity of the Dutch version of the 12-item Partners in Health Scale: Measuring patient self-management behaviour and knowledge in patients with chronic obstructive pulmonary disease. *PLoS One*, *11*(8), e0161595. doi:10.1371/journal.pone.0161595

Maffei, C., Fossati, A., Agostoni, I., Barraco, A., Bagnato, M., Deborah, D., . . . Petrachi, M. (1997). Interrater reliability and internal consistency of the Structured Clinical Interview for DSM-IV Axis II Personality Disorders (SCID-II), version 2.0. *Journal of Personality Disorders*, *11*, 279-284.

Marcus, D. K., Lilienfeld, S. O., Edens, J. F., & Poythress, N. G. (2006). Is antisocial personality disorder continuous or categorical? A taxometric analysis. *Psychological Medicine*, *36*, 1571-1581.

Mariani, J. J., Horey, J., Bisaga, A., Aharonovich, E., Raby, W., Cheng, W. Y., . . . Levin, F. R. (2008). Antisocial behavioral syndromes in cocaine and cannabis dependence. *American journal of Drug and Alcohol Abuse*, *34*, 405-414.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Moffitt, T. E., Arseneault, L., Jaffee, S. R., Kim-Cohen, J., Koenen, K. C., Odgers, C. L., . . . Viding, E. (2008). Research review: DSM-V conduct disorder: Research needs for an evidence base. *Journal of Child Psychology and Psychiatry*, *49*(1), 3-33.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, Netherlands: Mouton.

Murray, A. L., McKenzie, K., Murray, K. R., & Richelieu, M. (2014). Mokken scales for testing both pre- and postintervention: An analysis of the Clinical Outcomes in Routine Evaluation-Outcome Measure (CORE-OM) before and after counseling. *Psychological Assessment*, *26*, 1196-1204. doi:10.1037/pas0000015

Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment*, *14*(1), 50-59.

Pedersen, G., Karterud, S., Hummelen, B., & Wilberg, T. (2013). The impact of extended longitudinal observation on the assessment of personality disorders. *Personality and Mental Health*, *7*, 277-287. doi:10.1002/pmh.1234

Perdikouri, M., Rathbone, G., Huband, N., & Duggan, C. (2007). A comparison of adults with antisocial personality traits with and without childhood conduct disorder. *Annals of Clinical Psychiatry*, *19*(1), 17-23.

Pfohl, B., Blum, N., & Zimmerman, M. (1997). *Structured Interview for DSM-IV Personality: SIDP-IV*. Washington, DC: American Psychiatric Press.

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reise, S. P., & Revicki, D. A. (2014). *Handbook of item response theory modeling: Applications to typical performance assessment*. New York, NY: Routledge.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*(1), 27-48. doi:10.1146/annurev.clinpsy.032408.153553

Robins, L. N. (1978). Sturdy childhood predictors of adult antisocial behaviour: Replications from longitudinal studies. *Psychological Medicine*, *8*, 611-622.

Rosenström, T., Ystrom, E., Torvik, F. A., Czajkowski, N. O., Gillespie, N. A., Aggen, S. H., . . . Reichborn-Kjennerud, T. (2017). Genetic and environmental structure of DSM-IV criteria for antisocial personality disorder: A twin study. *Behavior Genetics*, *47*, 265-277.

Samejima, F. (1996). The graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York, NY: Springer.

Sheehan, D. V., Lecrubier, Y., Janavs, J., Knapp, E., Weiller, E., & Bonora, L. I. (1994). *Mini International Neuropsychiatric Interview (MINI)*. Tampa: University of South Florida Institute for Research in Psychiatry and INSERM-Hôpital de la Salpétrière.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Skodol, A. E., Morey, L. C., Bender, D. S., & Oldham, J. M. (2015). The alternative DSM-5 model for personality disorders: A clinical application. *American Journal of Psychiatry*, *172*, 606-613.

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry*, *24*, 399-411.

Stewart, M. E., Allison, C., Baron-Cohen, S., & Watson, R. (2015). Investigating the structure of the autism-spectrum quotient using Mokken scaling. *Psychological Assessment*, *27*, 596-604. doi:10.1037/pas0000058

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*, 209-220. doi: 10.1037/a0023353

Torgersen, S., Kringlen, E., & Cramer, V. (2001). The prevalence of personality disorders in a community sample. *Archives of General Psychiatry*, *58*, 590-596.

Trull, T. J., Jahng, S., Tomko, R. L., Wood, P. K., & Sher, K. J. (2010). Revised NESARC personality disorder diagnoses: Gender, prevalence, and comorbidity with substance dependence disorders. *Journal of Personality Disorders*, *24*, 412-426.

Tyrer, P., Crawford, M., Mulder, R. T., Blashfield, R. K., Farnam, A., Fossati, A., . . . Reed, G. M. (2011). The rationale of the reclassificaiton of personality disorder in the 11th revision of the International Classification of Disease (ICD-11). *Personality and Mental Health*, *4*, 246-259. doi:10.1002/pmh.190

van den Berg, S. M., Paap, M. C. S., & Derks, E. M. (2013). Using multidimensional modeling to combine self-report symptoms with clinical judgment of schizotypy. *Psychiatry Research*, *206*(1), 75-80.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41-71). Boston, MA: Springer.

Walters, G. D., & Knight, R. A. (2010). Antisocial personality disorder with and without antecedent childhood conduct disorder: Does it make a difference? *Journal of Personality Disorders*, *24*, 258-271.

Watson, R., van der Ark, L. A., Lin, L. C., Fieo, R., Deary, I. J., & Meijer, R. R. (2012). Item response theory: How Mokken scaling can be used in clinical practice. *Journal of Clinical Nursing*, *21*, 2736-2746. doi:10.1111/j.1365-2702.2011.03893.x

Weertman, A., Arntz, A., Dreessen, L., van Velzen, C., & Vertommen, S. (2003). Short-interval test-retest interrater reliability of the Dutch version of the Structured Clinical Interview for DSM-IV personality disorders (SCID-II). *Journal of Personality Disorders*, *17*, 562-567.

Widiger, T. A. (1998). Invited essay: Sex biases in the diagnosis of personality disorders. *Journal of Personality Disorders*, *12*, 95-118.

Widiger, T. A., Cadoret, R., Hare, R., Robins, L., Rutherford, M., Zanarini, M., . . . Frances, A. (1996). DSM-IV antisocial personality disorder field trial. *Journal of Abnormal Psychology*, *105*(1), 3-16.

Widiger, T. A., Mangine, S., Corbitt, E. M., Ellis, C. R., & Thomas, G. V. (1995). *Personality Disorder Interview-IV: A semi-structured interview for the assessment of personality disorder*. Odessa, FL: Psychological assessment resources.

Widiger, T. A., & Simonsen, E. (2005). Alternative dimensional models of personality disorder: Finding a common ground. *Journal of Personality Disorders*, *19*, 110-130.

Zimmerman, M., Rothschild, L., & Chelminski, I. (2005). The prevalence of DSM-IV personality disorders in psychiatric outpatients. *American Journal of Psychiatry*, *162*, 1911-1918.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432-442.