Check for updates

Article

A Psychometric Analysis of the Structured Clinical Interview for the *DSM-5* Alternative Model for Personality Disorders Module I (SCID-5-AMPD-I): Level of Personality Functioning Scale

Assessment
1–14
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1073191120967972
journals.sagepub.com/home/asm
SAGE

Benjamin Hummelen[1], Johan Braeken[1,2], Tore Buer Christensen[2,3], Tor Erik Nysaeter[3], Sara Germans Selvik[4,5], Kristoffer Walther[1], Geir Pedersen[1,2], Ingeborg Eikenaes[1,6], and Muirne C. S. Paap[1,7]

Abstract

The current study aims to examine the psychometric properties of the Structured Clinical Interview for the *DSM-5* Alternative Model for Personality Disorders Module I (SCID-5-AMPD-I) assessing the Level of Personality Functioning Scale (LPFS) in a heterogeneous sample of 282 nonpsychotic patients. Latent variable models were used to investigate the dimensionality of the LPFS. The results indicate that the LPFS, as assessed by the SCID-5-AMPD-I, can be considered as a unidimensional construct that can be measured reliably across a wide range of the latent trait. Threshold parameters for the 12 indicators of the LPFS increased gradually over the latent scale, indicating that the five LPFS levels were ordered as predicted by the model. In general, the increase of threshold parameters was relatively small for the shift from Level 2 to Level 3. A better distinction among the different severity levels might be obtained by fine-tuning the interview guidelines or the Level 2 indicators themselves.

Keywords

Level of Personality Functioning, Alternative *DSM-5* Model for Personality Disorders, polytomous item response theory

The fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*; American Psychiatric Association [APA], 2013) has incorporated two systems for diagnosing personality disorders (PD): (a) The traditional diagnostic system, which is identical to the DSM-IV system (APA, 1994); and (b) The Alternative *DSM-5* Model for Personality Disorders (AMPD), which is a hybrid model containing both PD dimensions and diagnostic categories. The dimensional aspects of personality pathology are identified by two main criteria. The A criterion, or Level of Personality Functioning Scale (LPFS), aims to assess the presence and general severity of personality pathology by delineating five levels of impairment of personality functioning, ranging from little or no impairment (Level 0) to extreme impairment of personality functioning (Level 4). The B criterion describes 25 pathological personality traits that are categorized into five broad domains (Negative Affectivity, Detachment, Antagonism, Disinhibition, and Psychoticism). To establish a PD diagnosis, a moderate or greater impairment of personality functioning is required (Level 2), as well as the presence of at least one pathological personality trait. The fact that two diagnostic systems for PDs are included in *DSM-5* reflects the decision of the APA Board of Trustees to maintain the traditional PD system in section II of *DSM-5* and consign the entire new proposal to Section III; emerging measures and models (see Zachar et al., 2016, for an outline of the processes leading to this decision).

[1]Oslo University Hospital, Oslo, Norway
[2]University of Oslo, Oslo, Norway
[3]Sørlandet Hospital, Arendal, Norway
[4]Hospital Namsos, Namsos, Norway
[5]Norwegian University of Science and Technology, Trondheim, Norway
[6]Vestfold Hospital Trust, Tønsberg, Norway
[7]University of Groningen, Groningen, The Netherlands

Corresponding Author:
Muirne C. S. Paap, Department of Research and Innovation, Clinic for Mental Health and Addiction, Oslo University Hospital, P.O.Box 4956 Nydalen, 0424 Oslo, Norway.
Email: m.c.s.paap@rug.nl

By embracing both the A and B criteria, the AMPD unites two dimensional paradigms: the psychodynamic paradigm and the multivariate paradigm (Wiggins, 2003). The multivariate paradigm is represented by the trait model and is the result of the lexical trait tradition in personality science with a solid background in factor-analytical studies (e.g., Watters & Bagby, 2018; Widiger & Simonsen, 2005). The psychodynamic paradigm is embodied by the A criterion (LPFS), since the content of this scale is derived from psychodynamic, attachment, social–cognitive, and interpersonal traditions, reviewed by Bender et al. (2011). A synthesis of the various concepts across the six measures included in this review and additional analyses by Morey et al. (2011) resulted in a final proposal for the LPFS, which was included in the AMPD after some minor revisions.

The LFPS was developed to measure a unidimensional construct relating to self- and interpersonal functioning, covering four areas of impairment, that is, Identity, Self-direction, Empathy, and Intimacy, each of which contains three narrower indicators (Bender et al., 2011; see also Tables 1 and 2). During and after the publication of *DSM-5*, several self-report questionnaires were developed to assess the 12 indicators of the LPFS, reviewed by Zimmermann et al. (2019). Though self-report questionnaires have important merits, like obtaining a quick first impression of the severity of personality pathology or evaluating clinical change in treatment studies, their use in clinical decision making is limited, and since the LPFS is a diagnostic tool, this scale must be assessed by clinicians using structured clinical interviews. For this purpose, Bender et al. (2018) developed the "Structured Clinical Interview for the *DSM-5* Alternative Model for Personality Disorders, module I" (SCID-5-AMPD-I). This interview has a "funnel structure," which implies that the 12 indicators of the LPFS are assessed by a combination of screener questions and questions for level determination (see also Method section). Screener questions are used to make an initial judgment of the most likely level of functioning for that indicator; for example, "Do you sometimes have the experience of not really knowing who you are or how you are unique in the world?" (Sense of Self). The interviewer continues by posing questions for level determination, starting with the level that the interviewee is assumed to match, then continues to the next level, and so on until the interviewee clearly does not qualify for that level of impairment. Since it is not necessary to ask questions at all levels, the funnel structure makes it possible to conduct the interview more efficiently, presumably without undermining the reliability and validity of the instrument. The first test–retest interrater reliability study of the SCID-5-AMPD-I found a good reliability estimate for the global LPFS score (interrater reliability coefficient = .75) and overall sufficient reliability estimates for most indicator scores (Buer Christensen et al., 2018).

It seems reasonable to examine the factor structure and psychometric properties of the SCID-5-AMPD-I along the same lines as the instruments that aim at capturing the B criterion (Zimmermann et al., 2019). Hitherto, virtually all factor analytical studies are based on self-report questionnaires. Among these, the most authoritative is probably the LPFS-Self Report, an 80-items self-report questionnaire developed by Morey (2017). In the first factor analytical study of this instrument, Morey (2017) found support for the contention that the LPFS reflects a single dimension. However, in a subsequent study, Sleep et al. (2019a) found poor fit for both a single factor model and a four-factor model that aligns with the *DSM-5* description. A limitation of these studies is the reliance on self-report questionnaires and the use of community dwellers recruited from Amazon's Mechanical Turk, limiting the opportunity to generalize the findings to clinical samples or make assumptions about the theoretical underpinnings of the LPFS. The next phase of research on the LPFS should take in use structured clinical interviews in samples that are representative for those whom the AMPD is made for, that is, patients with clinically relevant personality pathology.

In a discussion about the factor structure of the LPFS between Morey (2019) and Sleep et al. (2019b), it was brought up by Sleep et al. (2019b) that the LPFS would imply a multidimensional impairment model since the specific PD diagnoses in the AMPD, as well as PD-Trait Specified, require that impairment in personality functioning is manifested by difficulties in two or more of the four areas. However, this requirement appears not to be a strict decision rule; it is the presence of moderate or greater impairment in personality functioning that is decisive for assigning a PD diagnosis (Morey, 2019; Skodol, 2012; Skodol et al., 2015). Here, we assume that virtually all psychological constructs, including the LPFS, are inherently multidimensional, and that the question is rather to what *degree* this multidimensionality is present. Given the theory underlying the LPFS, we would expect psychometric analyses of instruments measuring the LPFS to show support for (essential) unidimensionality.

It is important to note that the five levels of LPFS are not scored using a Likert-type scale; rather, PD prototype aspects are included at some levels. For instance, the Desire and Capacity for Closeness indicator at Level 2 is clearly descriptive for narcissistic PD ("Intimate relationships are predominantly based on meeting self-regulatory and self-esteem needs, with an unrealistic expectation of being perfectly understood"), whereas at Level 3, this indicator is more characteristic for borderline PD ("Relationships are based on a strong belief in the absolute need for the intimate other, and/or expectations of abandonment and abuse"). Compared with the AMPD, the SCID-5-AMPD-I accentuates prototypical differences even more at several locations.

For instance, it is not obvious that the feature "Somewhat goal-inhibited" (Level 1 of "Ability to Pursue Meaningful Goals"; p. 775 of *DSM-5*) reflects an obsessive–compulsive personality trait. However, the SCID-5-AMPD-I recommends assessing this feature by the following question: "Does needing to get things just right make it hard to set or achieve behavior for yourself?" which clearly implies an obsessive–compulsive trait.

Although the prototype features are not salient enough throughout the entire interview to posit that each level represents a specific PD type, it is a good illustration of the presence of qualitative differences at each level. Still it is not clear whether the qualitative differences are indeed associated with increasing levels of severity of personality pathology. Regarding the specific PDs, there is some empirical support for the viewpoint that patients with borderline PD and schizotypal PD are more severely disturbed than patients with obsessive-compulsive PD with respect to impairment in work, social relationships, and leisure (Skodol et al., 2002). Patients with narcissistic PD, however, are not necessarily characterized by functional disability in work and social situations, though their personality functioning might be severely impaired (Ronningstam, 2009). Thus, from an empirical perspective, it is not obvious that the SCID-5-AMPD-I levels should indicate increasing degrees of severity. A useful way to assess whether the scoring levels are ordered would be to apply polytomous item response theory (IRT) modeling. The first LPFS study using polytomous IRT modeling was conducted by Zimmermann et al. (2015). Data were collected through an online survey, in which 515 lay persons and 145 therapists were asked to assess a personal acquaintance (for lay persons) or a patient (for therapists). Personality functioning was rated using a list of 60 items that closely followed the descriptions of the LPFS. In essence, only one indicator (Depth and Duration of Connections) displayed an ordering that was in line with all five severity levels according to the LPFS. A main limitation of this study is that the scorings were obtained by means of a checklist filled out by lay persons or by therapists with limited knowledge of the AMPD, which clearly illustrates the importance and necessity for studies using well-designed structured clinical interviews.

Aims

The aim of the current study is to evaluate the psychometric properties of the LPFS as operationalized by the SCID-5-AMPD-I in a heterogeneous sample of nonpsychotic patients, representing a wide range of PD severity levels. We expect that the 12 SCID-5-AMPD-I indicators will constitute a single dimension that can be measured reliably along the entire latent severity trait. We will compare a unidimensional model with several competing multidimensional models using the IRT framework.

According to *DSM-5*, disturbances in self and interpersonal functioning are supposed to be situated on a continuum, and our results should therefore support the notion that the five levels of the indicators are positioned on an ordinal scale, ranging from no or minimal impairment in personality functioning (Level 0) to extreme impairment in personality functioning (Level 4). We will use nominal IRT modeling to explore whether the levels are indeed ordered as expected. Finally, we will assess known groups validity using the number of PDs assessed employing the traditional categorical system as an external variable.

Method

Participants

The current study is part of a larger clinical multicenter study evaluating the reliability, validity, and clinical utility of the AMPD with main focus on the LPFS (e.g., Buer Christensen et al., 2018). The clinical sample consists of 282 patients, recruited at a variety of clinical sites in different Norwegian health regions. Most patients were female (n = 182; 65%), and mean age was 32 years (SD = 10; range 16 to 72).

Initially, the clinical sample comprised 286 patients. One patient was excluded because of missing diagnostic information, and three patients were excluded because of diagnostic contraindications (i.e., autism spectrum disorder, diagnosed after inclusion in the study). Other exclusion criteria were schizophrenia spectrum disorder (except schizotypal PD), sequelae after brain injury, severe ongoing substance abuse, intellectual disability, and other pervasive developmental disorders besides autism spectrum disorders, and lack of understanding of the Norwegian language.

The sample (N = 282) also includes 30 patients who had participated in a former test–retest interrater reliability study (Buer Christensen et al., 2018). In this study, the SCID-5-AMPD-I was administered by seven raters: three experienced clinicians and four inexperienced clinicians. All patients were assessed separately by two raters, performing the interviews at a maximum interval of 2 weeks. Thirteen patients (5%) were interviewed by inexperienced clinicians only, both the first and second interviews. Results of the first interview were used in the current study. For the remaining 17 patients, only results of the interviews conducted by experienced clinicians were used. Three out of 33 patients participating in the interrater reliability study were not included in the current study due to a large discrepancy between the first and second interview (difference of mean LPFS larger than 1.0). Both the first and second interviews

were conducted by inexperienced clinicians. See Buer Christensen et al. (2018) for details.

Traditional PD diagnoses were assessed before inclusion in the study by therapists at the clinical units where the patients were recruited, using the Structured Clinical Interview for *DSM-IV* Axis II PDs (SCID-II, First, 1994). The quality of the SCID-II assessments was ascertained through training referring therapists in consensus building, organized by the Clinic Mental Health and Addiction at the Oslo University Hospital. A former study conducted at this clinic found good reliability estimates of PD diagnoses assessed by the SCID-II (Arnevik et al., 2009). Diagnostic information of 276 patients was available. Among these, 188 (70%) fulfilled criteria for one or more PDs. The most common PD diagnosis was avoidant PD (n = 81; 29%), followed by borderline PD (n = 70, 25%), and PD not otherwise specified (n = 45, 16%). Antisocial PD was relatively common (n = 30, 11%) due to the inclusion of two addiction clinics, one of which served the local prison. The prevalence of schizotypal, schizoid, histrionic, and narcissistic PD was less than 2%. Among the 276 patients whose PD diagnoses were available, 141 had one PD diagnosis (51%), 28 had two PD diagnoses (10%), and 24 had three or more PD diagnoses (9%).

As with PD diagnoses, symptom disorders were assessed by referring therapists, using the Mini-International Neuropsychiatric Interview for Axis I diagnoses (Sheehan et al., 1994). Information of 254 patients concerning symptom disorders was available. The mean number of symptom diagnoses among these 254 patients was 1.7 (SD = 1.3, range 0-8). However, 93% had one or more symptom diagnoses; the most common symptom diagnosis was major depression (48%), followed by social phobia (22%), substance and alcohol use disorder (17%), posttraumatic stress disorder (14%), and generalized anxiety disorder (12%).

All participants provided written informed consent prior to participation in this study. The study is approved by the Regional Committee for Medical Research Ethics—South East Norway. Due to restrictions imposed by the Medical Research Ethics Committee regarding patient confidentiality, data are only available on request from the corresponding author. Requests can be sent to the Data Protection Officer at the Oslo University Hospital at: personvern@ous-hf.no.

Recruitment Sites and Raters

Patients were recruited at general outpatient departments (n = 70), general inpatient departments (n = 34), addiction outpatient departments (n = 30), addiction inpatient departments (n = 2), and specialized PD treatment units (n = 146). The PD treatment units were all part of the Norwegian Network for Personality Focused Treatment Programs (Karterud et al., 2003).

Most interviews (95%) were administered by experienced clinicians (four psychiatrists and three clinical psychologists), trained in consensus building by Donna Bender during a 2-day workshop, described in more detail by Buer Christensen et al. (2018). The inexperienced clinicians (three psychology students and one medical student) were trained by two of the experienced clinicians during a workshop that was virtually identical to the workshop given by Dr. Bender.

SCID-5-AMPD Module I. The SCID-5-AMPD-I (Bender et al., 2018) is a semistructured interview that covers the 12 indicators of the LPFS. The interviewer starts the assessment of personality functioning by posing eight general questions to obtain a basic sense of the interviewee's view of self and the quality of interpersonal relationships, for example, "How would you describe yourself as a person," or "What are your relationships with other people like?" After these initial questions, the 12 items are assessed separately by posing a combination of screener questions and questions for level determination. Based on the interviewee's responses to these screener questions and the responses to the eight preliminary questions, the interviewer conducts a preliminary evaluation of the level at which the interviewee may be functioning, and proceeds by posing determination questions pertaining to that level. The interviewer continues to pose questions corresponding to increasing levels of impairment, until the interviewee clearly does not qualify for that level of impairment, which would imply a score just beneath that level. If none of these levels are applicable, the interviewer carries on posing questions at the level just beneath the lowest level already assessed and continues in descending order. However, it should be noted, that by the time this study started, the instructions were not fully elaborated, and the interviewers were instructed by Donna Bender to start with the determination question at the level below the level at which they assumed the interviewee might be functioning.

The number of screener questions to choose from ranges from one (Self-Esteem) to five (Desire and Capacity for Closeness), and the number of level determination questions ranges from one (Self-Esteem, Comprehension and Appreciation of Others' Experiences and Motivations, and Tolerance of Differing Perspectives—all at Level 1) to six (Sense of Self and Prosocial Standards of Behavior, both at Level 3). A dual-design interrater reliability study of the SCID-5-AMPD-I conducted by the current research group found excellent intraclass correlation coefficients for indicator scores based on a video-based design (median = .84), and acceptable estimates for indicator scores based on a test–retest design (median = .55, Buer Christensen et al., 2018).

Psychometric Analyses

Dimensionality Analyses. To explore the (uni)dimensionality of the SCID-5-AMPD-I, seven competing models were estimated and compared using an IRT framework:

1. A unidimensional graded response model (GRM)
2. A GRM with two uncorrelated factors
3. A correlated traits GRM with two factors
4. A GRM with four uncorrelated factors;
5. A correlated traits GRM with four factors;
6. A bifactor model with two specific factors;
7. And a bifactor model with four specific factors.

The GRM (Samejima, 1997) can be used to model ordered categorical item scores. The model is a so-called indirect model: the probability that a patient is scored in a particular category k is based on differences between cumulative response probabilities. More detailed information about this model can be found in Paap et al. (2020) and the online supplement accompanying the current study. The main distinguishing feature of the bifactor model (Cai, 2010; Gibbons & Hedeker, 1992; Reise, 2012) is that the items load on both the general factor and the so-called group factors; where, in a correlated-trait model, items load on their own respective factors, and these factors are allowed to correlate. We refer the interested reader to the online supplement accompanying the article by Paap et al. (2015) for a more detailed comparison of bifactor analysis to other commonly used techniques for assessing dimensionality.

For models with two (specific) factors, items[1] relating to self-functioning were assigned to the first factor, and items relating to interpersonal functioning to the second factor. In models with four (specific) factors, the items were assigned based on the four areas of impairment; that is, Identity, Self-direction, Empathy, and Intimacy. Table 1 illustrates the item grouping.

Several outcomes were considered when comparing the models. First, we looked at overall fit, which was evaluated using the following indices and rules-of-thumb: the comparative fit index (CFI), good fit if CFI $\geq$ 0.95 and acceptable fit if CFI was between 0.90 and 0.95; the Tucker–Lewis index (TLI), good fit if TLI $\geq$ 0.90, and the root mean square error of approximation (RSMEA), good fit if RSMEA $\leq$ 0.06, acceptable fit if RMSEA was between 0.06 and 0.08 (Cook et al., 2009; Hu & Bentler, 1999). Note that the reported fit statistics were derived from the M_2^* statistic introduced by Cai and Hansen (2013). For the bifactor models, we also looked at the percentage of explained common variance (ECV; Reise et al., 2010; Ten Berge & Sočan, 2004) that was attributable to the general factor and to group factors, and at the model-based reliability associated with each factor. These indices were used to evaluate whether the specific factors had incremental value over and above the general factor, or whether the bifactor analyses supported essential unidimensionality. Reise et al. (2010) showed that, when the ECV for the general factor in a bifactor model is larger than 60%, the estimated factor loading for a unidimensional model are close to the true loadings on the general factor in the bifactor model. Therefore, an ECV >60% can be interpreted as indicative of essential unidimensionality. The model-based reliability is based on the sample distribution of the ability estimates. In this study, the Maximum A Posteriori estimator was used to calculate ability estimates.

Table 1. The Four Areas of Impairment and 12 Indicators of the *DSM-5* LPFS.

Identity (Self)
1. Experience of oneself as unique, with clear boundaries between self and others
2. Stability of self-esteem and accuracy of self-appraisal
3. Capacity for, and ability to regulate, a range of emotional experience
Self-direction (Self):
4. Pursuit of coherent and meaningful short-term and life goals
5. Utilization of constructive and prosocial internal standards of behavior
6. Ability to self-reflect productively
Empathy (Interpersonal)
7. Comprehension and appreciation of others' experiences and motivations
8. Tolerance of differing perspectives
9. Understanding of one's own behavior on others
Intimacy (Interpersonal)
10. Depth and duration of connection with others
11. Desire and capacity for closeness
12. Mutuality of regard reflected in interpersonal behavior

Note. The indicators are evaluated on a continuum, operationalized on a scale ranging from little or no impairment (i.e., healthy, adaptive functioning; Level 0), to some (Level 1), moderate (Level 2), severe (Level 3), and extreme (Level 4) impairment. Reproduced with permission of the American Psychiatric Association.

Additional Model Fit Analyses. Because inferences that are based on scores obtained from poorly fitting IRT models might be misleading, goodness-of-fit for the model chosen in the previous step was inspected in more detail. This was done using local fit statistics; these indices indicate *which parts* of the model are supported by the data, and which ones are not. We use predictive checks (Berkhof et al., 2000) based on a parametric bootstrap (Efron & Tibshirani, 1993) with 5,000 samples to assess local fit. See the online supplement for more information.

Category Use and Ordinal Score. To investigate the assumed ordering or gradation of response categories (higher category scores should be indicative of higher latent trait scores) and the actual usage of the categories per item, we plotted and inspected the item category trace curves, which depict the probability of choosing a particular response category as a function of the latent trait (here: personality functioning). We also fitted a competing model that does not assume a category order: the nominal response model (NRM; Bock, 1972). Fitting the NRM allows us to formally assess, whether the levels are indeed ordered as expected. See the

Table 2. Model Fit Statistics.

Model	logLik	*df*	TLI	CFI	RMSEA
Unidimensional GRM	−4065	244140564	.740	.843	.132
GRM, two uncorrelated factors	−4179	244140564	.488	.692	.186
Correlated traits GRM, two factors	−3991	244140563	1.00	1.00	.038
GRM, four uncorrelated factors	−4493	244140564	.603	.762	.163
Correlated traits GRM, four factors	−4050	244140558	.922	.969	.072
Bifactor model, two specific factors	−3936	244140552	1.00	1.00	.000
Bifactor model, four specific factors	−3947	244140552	.990	.998	.025

Note. logLik = log likelihood; *df* = degrees of freedom; TLI = Tucker–Lewis index; CFI = comparative fit index; RMSEA = root mean square error of approximation; GRM = graded response model.

online supplement for a more detailed description of the NRM.

Local Reliability: Test Information Function and Targeting. In the IRT framework, measurement error is conceptualized in terms of information: more information signifies higher levels of precision and less error of measurement. In contrast to classical reliability, test information can vary across the latent trait scale and is a direct function of the item characteristics. Plotting this test information across the latent trait scale allows for evaluating whether we can reliably measure patients across the entire relevant latent trait range (e.g., from −3 to +3). To ease interpretation, information can be converted to a local reliability estimate. Given that the squared standard error of measurement $SE\left(\theta_p\right)^2$ is equal to the reciprocal of the test information $I\left(\theta_p\right)$, an estimate of local reliability can be computed as follows:

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

The first equation is related to the classic 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 equation is based on the reciprocal information-error relation and the fact that our GRM-based latent trait scale metric is standardized such that $VAR\left(\theta_p\right) = 1$.

Known Groups Validity. The distribution of person estimates obtained using IRT models was compared for three groups: patients with 0, 1, and 2 or more PDs (as assessed by the SCID-II), respectively. The expectation was that the mean theta estimate would increase as the number of PDs increased. This was tested employing one-way analysis of variance and Tukey post hoc tests.

Software. All statistical analyses were coded and performed in the open source software program R version 3.4.3 (R Development Core Team, 2017). All models were estimated using a full information maximum likelihood approach in the R package mirt version 1.26.3 (Chalmers, 2012). Further analyses and diagnostics were custom coded in R.

Results

Dimensionality of the SCID-5-AMPD-I Items

Table 2 shows model fit statistics for all estimated models. Poor fit was found for the GRMs with two or four uncorrelated factors. The fit for the unidimensional model was somewhat better, but still inadequate. Adding correlations for the multidimensional models greatly improved model fit. For the GRM with two correlated factors, a correlation of .83 was found. For the GRM with four correlated factors, correlations ranged between .36 and .81. The bifactor models showed good fit. The ECV exceeded .80 for both bifactor models (.81 for the model with two specific factors, and .82 for the model with four specific factors). This indicates that 81% to 82% of the common variance in responses can be attributed to a common general factor. This number greatly exceeds the 60% threshold. For the bifactor model with two specific factors the reliability estimates equaled .92, .68 and .59 for the general, first specific and second specific factor, respectively. For the bifactor model with four specific factors the reliability estimates equaled .91, .49, .14, .44, and .61 for the general and specific factors, respectively. Taking these findings together, some multidimensionality was present, but there was strong evidence for a highly dominant global factor. Therefore, the subsequent analyses (including more detailed fit and local dependency analyses) were performed for the unidimensional GRM. More detailed results for the correlated traits and bifactor models can be found in the online supplement.

GRM: Item Parameters

The IRT parameters estimated for the GRM are reported in Table 3. Characteristic of clinical settings (Reise & Waller, 2009), the discrimination parameters were quite large, ranging from 2.0 to 3.0. The threshold values (b_{i1}, b_{i2}, b_{i3}, and b_{i4})

Table 3. Item Parameters Based on the Graded Response Model and Item Fit Statistics.

Indicator/item *i*	a_i	b_{i1}	b_{i2}	b_{i3}	b_{i4}	$r(Y_i, S_i)$	95% CI	$\Delta\chi^2$	*LID(i, j): j*
Identity									
1. Sense of Self	2.57	−1.28	−0.26	0.20	1.63	0.80	[.76, .85]	[−19, 42]	2, 3, 4
2. Self-Esteem	2.37	−2.13	−0.85	−0.22	1.79	0.75	[.70, .81]	[−40, 70]	1, 3, (−9)
3. Emotional Range and Regulation	2.56	−1.68	−0.61	0.14	2.37	0.76	[.73, .83]	[−25, 52]	1, 2
Self-direction									
4. Ability to Pursue Meaningful Goals	2.29	−1.32	−0.52	0.03	1.91	0.75	[.72, .82]	[−33, 55]	5
5. Constructive, Prosocial Internal Standards of Behavior	2.27	−1.58	−0.30	0.68	2.11	0.76	[.72, .82]	[−25, 26]	4
6. Self-Reflective Functioning	3.03	−1.54	−0.32	0.27	1.92	0.82	[.79, .86]	[−18, 44]	
Empathy									
7. Comprehension and Appreciation of Others' Experiences	2.57	−0.84	−0.05	0.72	2.53	0.79	[.76, .84]	[−05, 19]	9
8. Tolerance of Differing Perspectives	2.33	−0.83	0.22	0.92	1.99	0.78	[.72, .82]	[−27, 22]	9
9. Understanding of Effects of Own Behavior on Others	2.03	−0.64	0.23	1.15	2.24	0.74	[.68, .79]	[−21, 26]	(−2), 7, 8
Intimacy									
10. Depth and Duration of Connections	2.37	−1.16	−0.67	0.15	1.83	0.78	[.73, .83]	[−24, 45]	11, 12
11. Desire and Capacity for Closeness	2.72	−1.20	−0.37	0.13	1.63	0.81	[.77, .85]	[−23, 43]	10, 12
12. Mutuality of Regard Reflected in Interpersonal Behavior	2.59	−1.07	0.11	0.42	1.94	0.80	[.75, .84]	[−29, 36]	10, 11
Mean	2.47	−1.27	−0.28	0.38	1.99				

Note. The descriptions of the indicators deviate slightly from the formulations in *DSM-5* but are identical to the formulations in the SCID-5-AMPD. CI = confidence interval; LID = local item dependence.

indicate the position on the latent trait, where the probability of being assigned one of the next categories becomes larger than .5. Ideally, the threshold values within one item are evenly spaced, covering a large range of the latent trait. However, some of the threshold values within the same item were grouped quite closely together, especially for the step from Level 2 to Level 3 (from b_{i2} to b_{i3}). For instance, for the last item (Mutuality of Regard Reflected in Interpersonal Behavior), the difference in threshold values between Step 2 to Step 3 ($b_{i2} - b_{i3}$) was .31, whereas the difference between Step 3 to 4 was 1.52. It should be noted that threshold values of the Empathy items were relatively large, which is best observed by comparing these values with the mean values in the bottom row of Table 3.

The category response curves (Figure 1), which depict the probability of choosing a particular response category (here: level) as a function of the latent trait (here: personality functioning), show that, for 6 out of 12 items, one of the categories hardly ever had a higher probability of being chosen compared with the adjacent categories. For five items (1, 2, 4, 11, and 12), this concerns Category 2. For instance, it is shown that for the twelfth item (Mutuality of Regard Reflected in Interpersonal Behavior), the category response curve of Category 2 is dominated by the category response curves of the two adjacent categories (i.e., Levels 1 and 3). This suggests that the probability of being assigned Level 2 is almost always smaller than the probability of being assigned Level 1 or Level 3. It should also be noted that for most items, category response curves for Level 2 were smaller (i.e., lower and narrower) than the other curves, indicating that Category 2 is chosen less frequently than the other categories.

Taking a Closer Look at Model Fit and Local Dependency

We performed additional model fit analyses by means of predictive checks (see Table 3), which entails comparing observed data characteristics with expected data characteristics under the fitted model. The *observed* correlations between item response and total sum score were well within the 95% confidence intervals (CIs) for their *expected* counterparts. The same holds for the first order chi-square statistics $\Delta\chi^2$ capturing the difference between observed category frequencies and equally distributed category frequencies per item. These findings imply good item fit.

With respect to the second-order chi-square statistics $\Delta\chi^2$ capturing the difference between observed pairwise frequency tables and corresponding pairwise-independent frequency tables, some discrepancies could be observed that may suggest local item dependencies. As the last column of Table 3 shows, local dependency occurred within all areas of impairment; that is, items within one area were stronger associated with other items in the same area than with other items. For instance, Item 1 is associated with Items 2, 3, and 4. Local dependency was most pronounced for Intimacy and least pronounced for Self-direction. An overall predictive check based on the item correlation matrix resulted in a 95% CI for $\Delta\chi^2$ of [−504, 100], which implies that in general the expected item dependency structure reflected the observed dependency structure reasonably well. Hence, overall fit was deemed acceptable. As the bifactor analyses indicated, the percentage of ECV that could be attributed to the specific factors was low. Taking these findings together, the effect of ignoring the

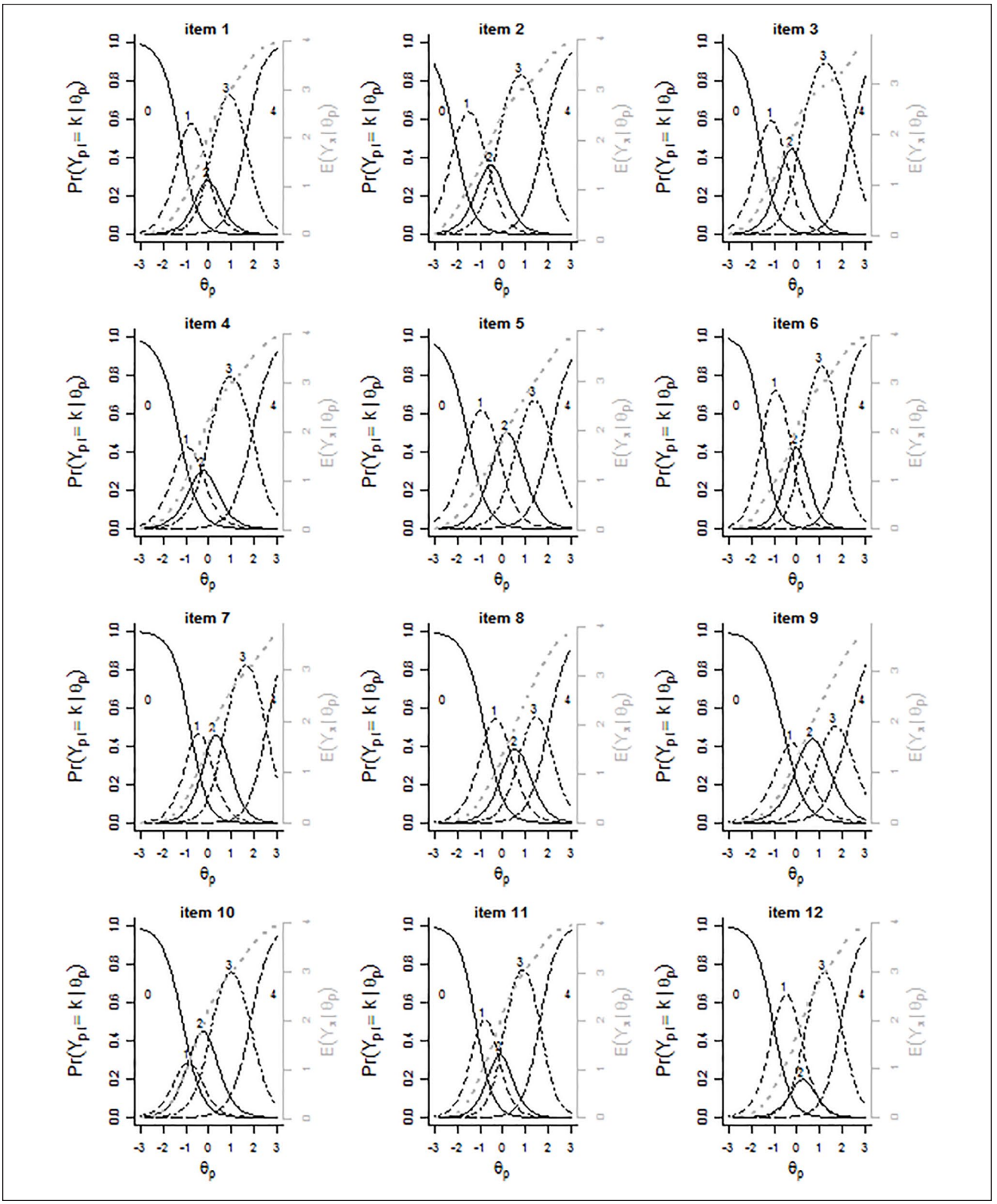

Figure 1. Category response curves for the 12 indicator (items) of the SCID-5-AMPD-I (expected item score conditional on latent trait value in grey).

Note. The item numbers correspond to the descriptions of the 12 indicators given in Table 1.

Table 4. Item Parameters (Slope Parameters and 95% CIs) Based on the Nominal Response Model.

Indicator/item *i*	a_i	a^*_{i1}	a^*_{i2}	a^*_{i3}	a^*_{i4}
Identity					
1. Sense of Self	1.94	0.99 [0.57, 1.43]	0.87 [0.45, 1.30]	**0.53** [0.17, 0.88]	**1.60** [1.08, 2.13]
2. Self-Esteem	1.85	1.33 [0.70, 1.98]	0.74 [0.31, 1.17]	0.99 [0.58, 1.40]	0.94 [0.55, 1.33]
3. Emotional Range and Regulation	2.74	1.50 [0.79, 2.19]	**0.60** [0.27, 0.93]	**0.50** [0.25, 0.76]	1.40 [0.65, 2.15]
Self-direction					
4. Ability to Pursue Meaningful Goals	1.64	**1.63** [1.06, 2.20]	0.70 [0.24, 1.18]	0.71 [0.33, 1.08]	0.96 [0.56, 1.37]
5. Constructive, Prosocial Internal Standards of Behavior	1.70	1.36 [0.86, 1.85]	0.78 [.42, 1.14]	0.84 [0.49, 1.19]	1.02 [0.53, 1.52]
6. Self-Reflective Functioning	2.48	0.98 [0.54, 1.43]	0.76 [0.40, 1.10]	0.90 [0.54, 1.25]	1.36 [0.80, 1.92]
Empathy					
7. Comprehension and Appreciation of Others' Experiences	1.85	1.02 [0.57, 1.46]	0.73 [0.33, 1.14]	1.07 [0.65, 1.50]	1.18 [0.53, 1.84]
8. Tolerance of Differing Perspectives	1.78	0.71 [0.37, 1.06]	0.89 [0.47, 1.31]	1.07 [0.58, 1.55]	1.32 [0.72, 1.93]
9. Understanding of Effects of Own Behavior on Others	1.47	0.71 [0.34, 1.09]	0.67 [0.25, 1.08]	1.41 [0.84, 1.97]	1.21 [0.55, 1.88]
Intimacy					
10. Depth and Duration of Connections	1.62	1.11 [0.56, 1.67]	0.86 [0.32, 1.39]	1.17 [0.77, 1.58]	0.86 [0.47, 1.25]
11. Desire and Capacity for Closeness	2.08	1.42 [0.90, 1.94]	0.99 [0.50, 1.47]	0.95 [0.54, 1.37]	**0.63** [0.33, 0.94]
12. Mutuality of Regard Reflected in Interpersonal Behavior	1.92	0.88 [0.51, 1.25]	1.04 [0.54, 1.54]	0.64 [0.19, 1.09]	1.44 [0.91, 1.97]

Note. Bold numbers specify slope parameters associated with confidence intervals (in square brackets) that did not include "1" in their range, indicating a sharper distinction between the adjacent categories (if the lower bound is larger than "1") or a less clear distinction between the adjacent categories (if the upper bound is smaller than "1").

local dependency that was identified is expected to be minor.

Response Categories

Table 4 summarizes the relevant results from the NRM. All category boundary slope parameters were clearly positive (i.e., all positive CIs) indicating that there was no problem with the assumed category ordering. However, the conditional discrimination parameters $a_i a^*_{ik}$ associated with the four sets of adjacent categories were not equal. Across all items, the $a_i a^*_{ik}$ parameters associated with the consecutive conditional item response curves $\Pr\left(Y_{pi} = k \mid \theta_p; (k \text{ or } k-1)\right)$ equaled 2.22, 1.52, 1.66, and 2.28; for adjacent category sets 0-1, 1-2, 2-3, and 3-4, respectively. This means that determining whether a patient's PD level should be scored as 1 rather than 2 (or vice versa) was less straightforward than choosing between 0 and 1, or 3 and 4. When inspecting the findings for each item separately, it can be seen that the upper bound of the 95% CIs for the a^*_{ik} parameter was smaller than "1" for four of the item-category combinations (Table 4); implying that in these cases it was especially challenging to differentiate between the adjacent categories in question.

Local Reliability

Figure 2 shows the test information across the latent trait dimension. It can be seen that a wide range of latent trait scores is sufficiently covered with this instrument: from about −2 to +3. The information function has two peaks ($\theta_p = -.14$ and $\theta_p = 1.85$). Even though a small dip in the information function occurs around a latent trait score of about 1, the local reliability still exceeds .9 in that region.

Known Groups Validity

Figure 3 shows that the degree of impairment in personality functioning (as measured with the SCID-5-AMPD-I and estimated using the GRM where categories "2" and "3" were merged) increased as the number of PDs increased. The mean theta score equaled −.40, .29, and .97 for the 0, 1, and 2+ PD groups, respectively. These differences were significant at the $\alpha = .05$ level. To facilitate interpretation of the theta scores, Figure S1 in the online supplement shows the correlation between theta scores and global (average) LPFS scores.

Discussion

In the dimensionality analyses, a strong general factor came to the fore, supporting the unidimensionality of the LPFS construct as measured by the SCID-5-AMPD-I in this clinical sample. The unidimensional score was strongly related to the number of PDs assessed with the SCID-II, which supports the validity of the instrument. Local reliability was high across a wide range of the latent trait

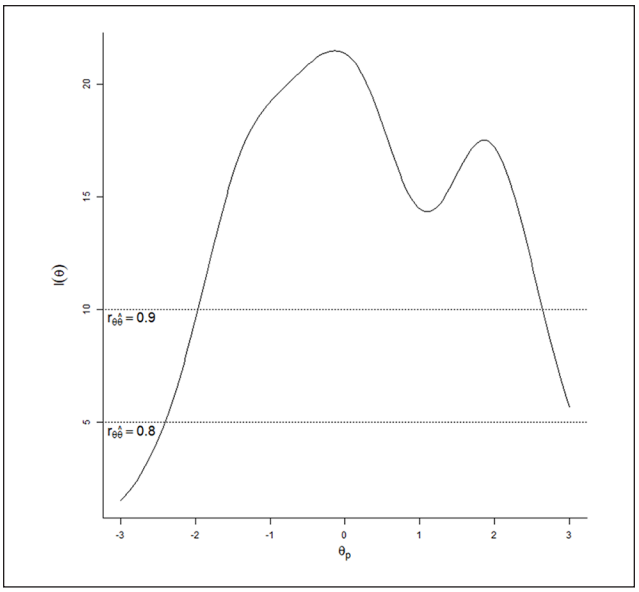

Figure 2. Test information function of the SCID-5-AMPD-I with the latent trait score on the *x*-axis and the Information on the *y*-axis.
Note. An information value of 5 corresponds to a local reliability estimate of .8, and an information value of 10 to a local reliability estimate of .9 (see horizontal dotted lines).

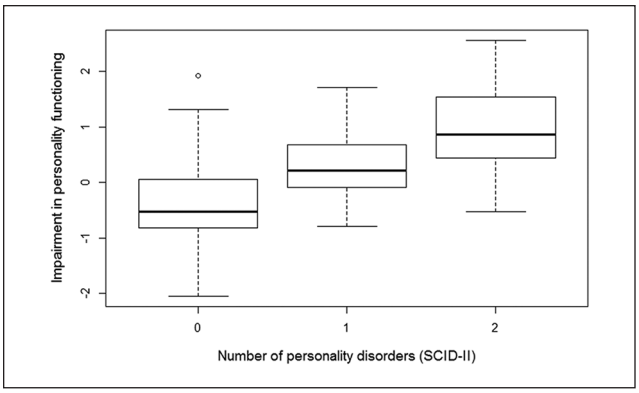

Figure 3. Boxplots with the degree of impairment in personality functioning (theta) on the *y*-axis, and number of PDs as assessed using the traditional categorical approach on the *x*-axis.
Note. Degree of impairment is expressed on a theta-scale, with a mean of 0. The theta estimates were obtained using a graded response model where scoring levels "2" and "3" were merged prior to analysis. Category 2 on the *x*-axis constitutes patients who had two *or more* PD's.

(personality functioning) and discrimination parameters obtained from the GRM analyses were large, indicating that the items were able to differentiate well among patients who were situated at different points on the latent personality functioning scale. The four threshold parameters within each item increased gradually over the latent scale, suggesting that the five LPFS levels were ordered as predicted by the theoretical model. However, for several indicators, the increase of threshold parameters was relatively small for the shift from Level 2 to Level 3, and relatively large for the shift from Level 3 to Level 4. The category response curves showed that for most indicators, Level 2 hardly ever had a higher probability of being chosen compared with the adjacent categories. Moreover, the results of the NRM analyses indicated that the distinction between Level 2 and its adjacent levels (Levels 1 and 3) was less pronounced compared with the distinction between Level 0 to 1 and Level 3 to 4. Overall, these findings suggest that Level 2 is the most problematic category from an IRT perspective, and raise the question whether Level 2 should be merged with Level 1 or with Level 3.

The finding that as much as 82% of the common variance was explained by the general factor supports the notion that the LPFS can be used as a unidimensional scale when assessed by the SCID-5-AMPD-I. Only a small proportion of the variance was left unexplained by the general factor in the bifactor analysis, emerging as local item dependency in the IRT analyses. This effect was more pronounced for items belonging to the same area of impairment, that is, Identity, Self-direction, Empathy, and Intimacy, indicating that items within the same area were still related to each other after the association with the general factor was accounted for. If a brief (or computerized adaptive) version of the instrument were to be developed, local dependence could be reduced by only including one item of an item pair showing a relatively high degree of dependence. More specifically, one of the Intimacy and/or Empathy items could be omitted. This notion is in line with our clinical experience with the SCID-5-AMPD-I, since we found that the content of the indicators of Empathy and Intimacy is less heterogeneous as compared with the content of the indicators of Identity and Self-direction. In fact, the indicators of Intimacy are rather narrow in their scope, focusing on nuanced differences of the quality of close relationships without taking into account other types of relationships.

The category response curves, which depict the probability of choosing a particular response level as a function of the latent trait, revealed that Level 2 was chosen relatively rarely, resulting in increased uncertainty with regard to parameter estimation. The low frequency of Level 2 may be explained in various ways. It could very well be due to sample characteristics, for instance. Our population had a low prevalence of narcissistic PD; and since Level 2 descriptions contain elements reflecting narcissistic personality, patients may only rarely have identified with these descriptions. However, a low frequency of narcissistic PD is a common phenomenon in clinical samples, at least in Northern and Western European outpatient clinics (Hummelen et al., 2006; Soeteman et al., 2008), and thus these findings may be

expected to generalize to other clinical samples. Another possible explanation could be that Level 2 puts too much emphasis on grandiose narcissism and too little emphasis on vulnerable narcissism. There is growing consensus among narcissism researchers that vulnerable narcissism is as important as grandiose narcissism (Baskin-Sommers et al., 2014; Miller et al., 2013). Vulnerable narcissism comprises characteristics like "My feelings are easily hurt by ridicule or by the slighting remarks of others," and "I often interpret the remarks of others in a personal way," and "I easily become wrapped up in my own interests and forget the existence of others." Miller et al. (2013) found that vulnerable narcissism was positively correlated with avoidant PD, the most common PD diagnosis in our sample. Broadening the range of personality pathology by including questions reflecting vulnerable narcissism may improve the utility of Level 2. The potential impact of such a change needs to be considered carefully, since vulnerable narcissism may convey more severe pathology than grandiose narcissism. In the study of Miller et al. (2013), vulnerable narcissism was associated with PD severity, which might imply that aspects of vulnerable narcissism should rather be included at Level 3.

With respect to the Identity indicators, the NRM analysis showed that the distinction between Levels 2 and 3 was relatively unclear for Sense of Self since the upper bound of the 95% CI was smaller than "1." Thus, there seems to be no clear category distinction between "to be excessively dependent on others for identity definition" (Level 2 description) and "having a weak sense of autonomy/agency; experience of a lack of identity, or emptiness" (Level 3 description). For Emotional Range and Regulation, the upper bound of the 95% CI was also smaller than "1" for Level 3. A clinical interpretation is that the Level 3 descriptions of Emotional Range and Regulation convey less severe personality pathology than expected when compared with Level 2. One possible explanation could be that Level 3 questions in the SCID-5-AMPD-I do not focus on the negative social consequences of emotional dysregulation, whereas Level 2 questions lay much emphasis on the social aspects of emotional dysregulation. In a similar vein, Level 4 questions reflect serious social corollaries, which may explain the relatively large increase of the location parameter in the GRM analyses for Emotional Range and Regulation. In future editions of the SCID-5-AMPD-I, more emphasis could be put on the social consequences of emotional dysregulation at Level 3.

We found that the threshold parameters of the Empathy indicators were somewhat larger than the threshold parameters of the other indicators. A clinical interpretation is that the Empathy indicators are well-suited to identify patients with more severe personality pathology. However, these findings could also be interpreted in a less favorable way; they could be indicative of assessment bias. That is, patients might have overestimated their own empathic capacities, whereas clinicians might not have been sufficiently able to estimate these capacities accurately at the first meeting, with no prior information about the patient, as was the case in this study. Similar concerns have been raised by Zimmermann et al. (2014), who suggested that direct questions might not be very helpful in the assessment of empathy (e.g., "Do you appreciate others' experiences and motivations?"). The authors recommended to include questions probing for reflective functioning (e.g., "Why did your parents behave as they did?"). It would be advisable for future studies on the assessment of personality functioning to focus explicitly on how to obtain a valid evaluation of empathic capacities.

Desire and Capacity for Closeness (Intimacy) was the only indicator with a conditional slope parameter that was notably smaller than 1 at Level 4. Thus, Level 4 questions for this indicator might be more indicative of severe impairment in personality functioning than extreme impairment. Many patients with PD, also those with moderate PD, would probably give an affirmative answer to questions like "Is it hard to trust people?" (Level 4 question). Moreover, narcissistic patients may respond affirmatively to questions like "Do you only interact with people when it's necessary to get what you need?" Thus, a more obvious distinction between Level 3 and Level 4 could be obtained by revising the SCID-5-AMPD-I (e.g., by focusing more on the social consequences of not having any desire or capacity for closeness).

Since Level 2 conveyed little information in our analyses, we repeated the analyses collapsing Levels 2 and 3 across all items (data not shown). This modification had no substantial impact on the factor structure and model fit in our analyses. A possible implication of these findings might be that the LPFS could be simplified by reducing the number of levels without loss of information. Level 1 descriptions could include the current obsessive–compulsive PD features and some narcissistic features, whereas the Level 2 descriptions could be revised to include the current Level 3 descriptions features as well as vulnerable narcissistic features. Although a simplified model may be easier to implement in clinical practice, it might imply an infringement on the content validity of the scale. For instance, Caligor et al. (2018) have outlined the clinical relevance of distinguishing between Level 2 and Level 3 in treatment planning and psychotherapeutic interventions. We recommend that future research should focus on the clinical utility of the LPFS, with special emphasis on the clinical application of the five levels.

Our findings may have some consequences for the assessment of PDs according to the International Classification of Diseases, 11th Revision (ICD-11, World Health Organization, 2018). The ICD-11 has a lot in common with the AMPD; it also makes a distinction between personality functioning and personality traits, for instance. Like the

AMPD, the ICD-11 assesses aspects of personality functioning that contribute to severity determination in PD using four out of five levels: Personality Difficulty, Mild Personality Disorder; Moderate Personality Disorder; and Severe Personality Disorder. However, there are three important differences between the AMPD and the ICD-11 approach to PDs. First, Level 0 used in the AMPD, that is, healthy personality functioning, is not defined in ICD-11. Second, it appears that the severity levels in ICD-11 are not fully compatible with the severity levels in *DSM-5*. That is, mild PD in ICD-11 seems less severe than moderate impairment in the AMPD; moderate PD seems less severe than severe impairment; and severe PD is obviously less severe than extreme impairment. Third, ICD-11 defines severity in terms of the number of impairments. For instance, mild PD is defined as "Disturbances affect some areas *but not others"* (e.g., problems with self-direction in the absence of problems with stability and coherence of identity or self-worth). In spite of these differences, we expect that the SCID-5AMPD-I may be useful in assessing severity according to the ICD-11; at least as long as no instruments are available that were specifically designed to operationalize the severity construct according to ICD-11. The reasons are that our study showed support for the SCID-5-AMPD-I measuring a unidimensional construct, as well as a clear relation to the number of PDs assessed with the SCID-II. Moreover, a previous study of the Norwegian Multicenter Study of the AMPD found that the five levels of impairment were meaningfully associated with other indicators of severity (Buer Christensen et al., 2020). To address the differences in the definition of the severity levels between the two systems, some adaptations could be made to the definition and the use of the levels in the SCID-5-AMPD-I to make them more compatible with the ICD-11.

The SCID-5-AMPD-I has a funnel structure to allow for more efficient assessment compared with a linear administration (i.e., assessing all levels for each question). Although this approach has clear advantages, such as a reduced test administration time, it should be acknowledged that the implicit assumption is made that this approach does not affect the quality of the psychometric properties of the instrument. Although we have no reason to believe that the assessed scores would have been different if a linear approach had been used, this is not something we can rule out at this point. To reduce the chance of this potential bias occurring, the clinicians participating in this study did assess adjacent levels. However, it is uncertain whether an interviewee would have met the criteria for the levels that were not assessed. We are currently conducting a study in which all the levels of the LPFS are assessed for each indicator, to establish whether the funnel structure of the SCID-5-AMPD-I is indeed a valid approach.

It should be noted that the high local reliability could be partly due to the "halo-effect." The SCID-5-AMPD-I might be particularly sensitive to this effect, because clinicians are instructed to obtain an overall impression of the interviewee's level of personality functioning at the beginning of the interview by asking eight screening questions, and use this level as a reference in the subsequent assessment of the indicators. Thus, clinicians might be prone to give scorings that are in line with their preconceived idea of the interviewee's level of functioning. Potential consequences of the presumed halo-effect were not investigated in this study. Another limitation of this study is that the sample was relatively small from an IRT perspective. As a result, the standard errors associated with the parameter estimates are expected to be larger than they would have been, if a larger sample had been used. Finally, in a recent study conducted by members of our research group, some indicators were shown to have rather poor test–retest interrater reliability (Buer Christensen et al., 2018). Notwithstanding, most indicators had acceptable interrater reliability, and in contrast to the interrater reliability study, most patients in the current study were assessed by experienced clinicians. Moreover, interrater reliability estimates based on video recordings were excellent.

In sum, we found convincing support for the unidimensional structure of the SCID-5-AMPD-I, indicating that the general severity score obtained by this instrument can be used to determine the severity and presence of personality pathology as conceptualized in the AMPD. However, moderate level of impairment (Level 2) appeared to be problematic from an IRT perspective. Poor category distinction seemed to be most pronounced for the shift from Level 2 to Level 3 for several indicators. This less clear category distinction could be an empirical phenomenon, but could also be interpreted as a need for better guidelines and training to enable a more informed choice between the middle categories. It is advisable that these findings be taken into consideration for further refinement of the SCID-5-AMPD-I.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a FRIPRO Young Research Talent grant for the last author (Grant no. NFR 286893), awarded by the Research Council of Norway.

Supplemental Material

Supplemental material for this article is available online.

Note

1. In the present study, the term "items" is used to refer to the 12 SCID-5-AMPD indicators

References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.).

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.).

Arnevik, E., Wilberg, T., Urnes, Ø., Johansen, M., Monsen, J. T., & Karterud, S. (2009). Psychotherapy for personality disorders: Short-term day hospital psychotherapy versus outpatient individual therapy: A randomized controlled study. *European Psychiatry*, *24*(2), 71-78. https://doi.org/10.1016/j.eurpsy.2008.09.004

Baskin-Sommers, A., Krusemark, E., & Ronningstam, E. (2014). Empathy in narcissistic personality disorder: From clinical and empirical perspectives. *Personality Disorders*, *5*(3), 323-333. https://doi.org/10.1037/per0000061

Bender, D. S., Morey, L. C., & Skodol, A. E. (2011). Toward a model for assessing level of personality functioning in *DSM-5*, Part I: A review of theory and methods. *Journal of Personality Assessment*, *93*(4), 332-346. https://doi.org/10.1080/00223891.2011.583808

Bender, D. S., Skodol, A. E., First, M. B., & Oldham, J. M. (2018). Module I: Structured Clinical Interview for the Level of Personality Functioning Scale. In M. B. First, A. E. Skodol, D. S. Bender, & J. M. Oldham (Eds.), *Structured Clinical Interview for the DSM-5 Alternative Model for Personality Disorders (SCID-5-AMPD)*. American Psychiatric Association.

Berkhof, J., van Mechelen, I., & Hoijtink, H. (2000). Posterior predictive checks: Principles and discussion. *Computational Statistics*, *15*(3), 337-354. https://doi.org/10.1007/s001800000038

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51. https://doi.org/10.1007/bf02291411

Buer Christensen, T., Eikenaes, I., Hummelen, B., Pedersen, G., Nysæter, T. E., Bender, D. S., Skodol, A. E., & Selvik, S. G. (2020). Level of personality functioning as a predictor of psychosocial functioning-concurrent validity of Criterion A. *Personality Disorders*, *11*(2), 79-90. https://doi.org/10.1037/per0000352

Buer Christensen, T., Paap, M. C. S., Arnesen, M., Koritzinsky, K., Nysaeter, T. E., Germans Selvik, S., Walther, K., Torgersen, S., Bender, D. S., Skodol, A. E., Kvarstein, E., Pedersen, G., & Hummelen, B. (2018). Interrater reliability of the Structured Interview for the *DSM-5* Alternative Model for Personality Disorders Module I: Level of Personality Functioning Scale. *Journal of Personality Assessment*, *100*(6), 630-641. https://doi.org/10.1080/00223891.2018.1483377

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581-612. https://doi.org/10.1007/s11336-010-9178-0

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 245-276. https://doi.org/10.1111/j.2044-8317.2012.02050.x

Caligor, E., Kernberg, O. F., Clarkin, J. F., & Yeomans, F. E. (2018). *Psychodynamic therapy for personality pathology: Treating self and interpersonal functioning*: American Psychiatric Association.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R Environment. *Journal of Statistical Software*, *48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research*, *18*(4), 447-460. https://doi.org/10.1007/s11136-009-9464-4

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

First, M. B. (1994). *Structured Clinical Interview for DSM-IV Axis II Personality Disorders (SCID II)*. New York State Psychiatric Institute.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423-436. https://doi.org/10.1007/bf02295430

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Hummelen, B., Wilberg, T., Karterud, S., & Pedersen, G. (2006). An investigation of the validity of the *DSM-IV* avoidant personality disorder construct as a prototype category and the psychometric properties of the diagnostic criteria. *Comprehensive Psychiatry*, *47*(5), 376-383. https://doi.org/10.1016/j.comppsych.2006.01.006

Karterud, S., Pedersen, G., Bjordal, E., Brabrand, J., Friis, S., Haaseth, O., Haavaldsen, G., Irion, T., Leirvåg, H., Tørum, E., & Urnes, Ø. (2003). Day treatment of patients with personality disorders: Experiences from a Norwegian treatment research network. *Journal of Personality Disorders*, *17*(3), 243-262. https://doi.org/10.1521/pedi.17.3.243.22151

Miller, J. D., Gentile, B., Wilson, L., & Campbell, W. K. (2013). Grandiose and vulnerable narcissism and the *DSM-5* pathological personality trait model. *Journal of Personality Assessment*, *95*(3), 284-290. https://doi.org/10.1080/00223891.2012.685907

Morey, L. C. (2017). Development and initial evaluation of a self-report form of the *DSM-5* Level of Personality Functioning Scale. *Psychological Assessment*, *27*(10), 1302-1308. https://doi.org/10.1037/pas0000450

Morey, L. C. (2019). Thoughts on the assessment of the *DSM-5* alternative model for personality disorders: Comment on Sleep et al. (2019). *Psychological Assessment*, *31*(10), 1192-1199. https://doi.org/10.1037/pas0000710

Morey, L. C., Berghuis, H., Bender, D. S., Verheul, R., Krueger, R. F., & Skodol, A. E. (2011). Toward a model for assessing level of personality functioning in *DSM-5*, Part II: Empirical articulation of a core dimension of personality pathology. *Journal of Personality Assessment*, *93*(4), 347-353. https://doi.org/10.1080/00223891.2011.577853

Paap, M. C. S., Braeken, J., Urnes, Ø., Karterud, S., Wilberg, T., Pedersen, G., & Hummelen, B. (2020). A psychometric evaluation of the *DSM-IV* criteria for antisocial personality disorder: Dimensionality, local reliability, and differential

item functioning across gender. *Assessment*, *27*(1), 89-101. https://doi.org/10.1177/1073191117745126

Paap, M. C. S., Brouwer, D., Glas, C. A. W., Monninkhof, E. M., Forstreuter, B., Pieterse, M. E., & van der Palen, J. (2015). The St George's Respiratory Questionnaire revisited: A psychometric evaluation. *Quality of Life Research*, *24*(1), 67-79. https://doi.org/10.1007/s11136-013-0570-y

R Development Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667-696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*(6), 544-559. https://doi.org/10.1080/00223891.2010.496477

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

Ronningstam, E. (2009). Narcissistic personality disorder: Facing *DSM-V*. *Psychiatric Annals*, *39*(3), 111-121. https://doi.org/10.3928/00485713-20090301-09

Samejima, F. (1997). The graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). Springer.

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

Skodol, A. E. (2012). Personality disorders in *DSM-5*. *Annual Review of Clinical Psychology*, *8*, 317-344. https://doi.org/10.1146/annurev-clinpsy-032511-143131

Skodol, A. E., Gunderson, J. G., McGlashan, T. H., Dyck, I. R., Stout, R. L., Bender, D. S., Grilo, C. M., Shea, M. T., Zanarini, M. C., Morey, L. C., Sanislow, C. A., & Oldham, J. M. (2002). Functional impairment in patients with schizotypal, borderline, avoidant, or obsessive-compulsive personality disorder. *American Journal of Psychiatry*, *159*(2), 276-283. https://doi.org/10.1176/appi.ajp.159.2.276

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*(7), 606-613. https://doi.org/10.1176/appi.ajp.2015.14101220

Sleep, C., Lynam, D., Widiger, T. A., Crowe, M. L., & Miller, J. (2019a). An evaluation of *DSM-5* Section III Personality Disorder Criterion A (Impairment) in accounting for psychopathology *Psychological Assessment*, *31*(10), 1181. https://doi.org/10.31234/osf.io/z48tv

Sleep, C. E., Lynam, D. R., Widiger, T. A., Crowe, M. L., & Miller, J. D. (2019b). Difficulties with the conceptualization and assessment of Criterion A in the *DSM-5* alternative model of personality disorder: A reply to Morey. *Psychological Assessment*, *31*(10), 1200-1205. https://doi.org/10.1037/pas0000758

Soeteman, D. I., Hakkaart-van Roijen, L., Verheul, R., & Busschbach, J. J. (2008). The economic burden of personality disorders in mental health care. *Journal of Clinical Psychiatry*, *69*(2), 259-265. https://doi.org/10.4088/JCP.v69n0212

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613-625. https://doi.org/10.1007/BF02289858

Watters, C. A., & Bagby, R. M. (2018). A meta-analysis of the five-factor internal structure of the Personality Inventory for *DSM-5*. *Psychological Assessment*, *30*(9), 1255-1260. https://doi.org/10.1037/pas0000605

Widiger, T. A., & Simonsen, E. (2005). Alternative dimensional models of personality disorder: Finding a common ground. *Journal of Personality Disorders*, *19*(2), 110-130. https://doi.org/10.1521/pedi.19.2.110.62628

Wiggins, J. S. (2003). *Paradigms of personality assessment*: Guilford Press.

World Health Organization. (2018). *International statistical classification of diseases and related health problems* (11th Rev.). https://icd.who.int/browse11/l-m/en

Zachar, P., Krueger, R., & Kendler, K. J. P. M. (2016). Personality disorder in *DSM-5*: An oral history, *46*(1), 1-10. https://doi.org/10.1017/S0033291715001543

Zimmermann, J., Benecke, C., Bender, D. S., Skodol, A. E., Schauenburg, H., Cierpka, M., & Leising, D. (2014). Assessing *DSM-5* level of personality functioning from videotaped clinical interviews: A pilot study with untrained and clinically inexperienced students. *Journal Personal Assessment*, *96*(4), 397-409. https://doi.org/10.1080/00223891.2013.852563

Zimmermann, J., Böhnke, J. R., Eschstruth, R., Mathews, A., Wenzel, K., & Leising, D. (2015). The latent structure of personality functioning: Investigating Criterion A from the alternative model for personality disorders in *DSM-5*. *Journal of Abnormal Psychology*, *124*(3), 532-548. https://doi.org/10.1037/abn0000059

Zimmermann, J., Kerber, A., Rek, K., Hopwood, C. J., & Krueger, R. (2019). A brief but comprehensive review of research on the Alternative DSM-5 Model for Personality Disorders. *Current Psychiatry Reports*, *21*(9), Article 92. https://doi.org/10.1007/s11920-019-1079