

Personality Disorders: Theory, Research, and Treatment

Assessing DSM-5-Oriented Level of Personality Functioning: Development and Psychometric Evaluation of the Semi-Structured Interview for Personality Functioning DSM-5 (STiP-5.1)

Joost Hutsebaut, Jan H. Kamphuis, Dine J. Feenstra, Laura C. Weekers, and Hilde De Saeger
Online First Publication, June 16, 2016. <http://dx.doi.org/10.1037/per0000197>

CITATION

Hutsebaut, J., Kamphuis, J. H., Feenstra, D. J., Weekers, L. C., & De Saeger, H. (2016, June 16). Assessing DSM-5-Oriented Level of Personality Functioning: Development and Psychometric Evaluation of the Semi-Structured Interview for Personality Functioning DSM-5 (STiP-5.1). *Personality Disorders: Theory, Research, and Treatment*. Advance online publication. <http://dx.doi.org/10.1037/per0000197>

Assessing *DSM–5*-Oriented Level of Personality Functioning: Development and Psychometric Evaluation of the Semi-Structured Interview for Personality Functioning *DSM–5* (STiP-5.1)

Joost Hutsebaut

Viersprong Institute for Studies on Personality Disorders,
Halsteren, the Netherlands

Jan H. Kamphuis

University of Amsterdam and Viersprong Institute for Studies
on Personality Disorders, Halsteren, the Netherlands

Dine J. Feenstra

Viersprong Institute for Studies on Personality Disorders,
Halsteren, the Netherlands, and Erasmus Medical Centre,
Rotterdam, the Netherlands

Laura C. Weekers and Hilde De Saeger

Viersprong Institute for Studies on Personality Disorders,
Halsteren, the Netherlands

The alternative model for personality disorders (AMPD) in the *Diagnostic and Statistical Manual of Mental Disorders (DSM–5)* features a Level of Personality Functioning Scale, measuring intrinsic personality processes that include identity, self-direction, empathy, and intimacy. This study describes the development and psychometric evaluation of a semistructured interview schedule for the multi-item assessment of the level of personality functioning, the Semi-Structured Interview for Personality Functioning *DSM–5* (STiP-5.1). Eighty patients and 18 community subjects completed the STiP-5.1. Patients additionally completed the Brief Symptom Inventory, the Severity Indices of Personality Problems, and the Structured Clinical Interview for *DSM–IV* Axis I and Axis II Personality Disorders. Good interrater reliability was observed in subsamples of patients ($n = 40$) and nonpatients ($n = 18$). Associations between the interview scores and conceptually relevant external measures consistently supported the construct validity of the instrument. The STiP-5.1 thus offers a brief, relatively user-friendly instrument with generally favorable psychometric properties for the assessment of level of personality functioning of the *DSM–5* AMPD.

Keywords: personality disorders, assessment, *DSM–5*, Level of Personality Functioning Scale, AMPD

The fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders (DSM–5)*; American Psychiatric Association, 2013) features in its Section III an alternative model for personality disorders (AMPD). At the core of this newly proposed classification of personality disorders is an assessment of level of impairment in personality functioning (Criterion A). Impairments constituting *personality* pathology are presumed to manifest in self and/or interpersonal functioning (Bender, Morey, & Skodol, 2011). Self-functioning, according to this model, refers to a range of adaptive psychological abilities related to the elements of iden-

tity and self-direction. Identity encompasses the facets of experience of oneself as unique, stability of self-esteem, and capacity for emotion regulation, whereas self-direction refers to the facets of pursuit of meaningful goals, utilization of prosocial internal standards of behavior, and the ability to self-reflect productively. The domain of interpersonal functioning refers to capacities related to the elements of empathy and intimacy. Empathy includes the ability to understand the mental world of others, to tolerate differing perspectives, and to understand the impact of one's own behavior. Finally, intimacy encompasses the ability to establish durable and meaningful relations with others, to experience and tolerate closeness, and to cooperate on the basis of mutuality of regard.

The *DSM–5* AMPD's Level of Personality Functioning Scale (LPFS) uses each of these 12 facets to differentiate five levels of severity of personality pathology, ranging from little or no impairment (Level 0) to extreme impairment (4). Criterion A requires a moderate or more-severe level of impairment (Level 2 or greater) for the classification of a categorical personality disorder (PD), on the basis of a study by the Personality and Personality Disorders Work Group members showing that a "moderate or greater" impairment demonstrated 84.6% sensitivity and 72.7% specificity for identifying patients who met criteria for at least one PD diagnosis according to the fourth edition of the *Diagnostic and Statistical*

Joost Hutsebaut, Viersprong Institute for Studies on Personality Disorders, Halsteren, the Netherlands; Jan H. Kamphuis, Department of Clinical Psychology, University of Amsterdam and Viersprong Institute for Studies on Personality Disorders; Dine J. Feenstra, Viersprong Institute for Studies on Personality Disorders and Department of Medical Psychology and Psychotherapy, Erasmus Medical Centre, Rotterdam, the Netherlands; Laura C. Weekers and Hilde De Saeger, Viersprong Institute for Studies on Personality Disorders.

Correspondence concerning this article should be addressed to Joost Hutsebaut, Viersprong Institute for Studies on Personality Disorders, PB 7, 4660 AA Halsteren, Halsteren, the Netherlands. E-mail: joost.hutsebaut@deviersprong.nl

Manual of Mental Disorders (DSM-IV; American Psychiatric Association, 1994; Morey, Bender, & Skodol, 2013). Level of impairment may also serve as a specifier to a categorical diagnosis, thus offering relevant supplemental information to the clinician for treatment planning (American Psychiatric Association, 2013). Also, in the absence of a specific type of PD, a diagnosis of *personality disorder—trait specified* could be assigned on the basis of the level of impairment and prevailing maladaptive traits.

The introduction of impairment of personality functioning (i.e., Criterion A) in the new model was designed to remedy some of the major well-documented shortcomings of the existing Section II model, including limited diagnostic reliability, extensive co-occurrence among PDs, and large heterogeneity within categories with limited possibilities to represent variance within PD types (Morey, Benson, Busch, & Skodol, 2015). Assessment of impairment in self and interpersonal functioning may serve the twin goals of capturing some of the assumed “essential commonalities” among PDs of different types (Morey et al., 2011, p. 347) on the one hand and to demarcate PDs more specifically from other types of mental disorders on the other.

Although both Criteria A and B have demonstrated incremental value in treatment planning, the AMPD seems to prioritize a taxation of severity (Criterion A) above a stylistic characterization of PDs (Criterion B), much in line with psychodynamic tradition (McWilliams, 1994). Indeed, severity of PD was found to be a better predictor of therapy outcome than was PD classification (Bernstein, 1998), the best predictor of prospectively assessed functional impairment in patients with PD after 10 years of follow-up (Hopwood et al., 2011), a strong predictor of treatment outcome for depression (Oleski, Cox, Robinson, & Grant, 2012), and a significant predictor for differential outcomes for specialist versus generalist treatment of borderline PD (Bateman & Fonagy, 2013).

Although there now seems to be general acceptance of the notion that assessment of severity should be included in any (new) diagnostic system for PDs (Kim & Tyrer, 2010; Tyrer, Reed, & Crawford, 2015), the specific conceptualization differs across the International Classification of Diseases—11th Revision proposal (Tyrer et al., 2015) and Sections II and III of the *DSM-5* diagnostic systems (American Psychiatric Association, 2013). *DSM-5*'s AMPD sought to relate severity directly to *personality processes*, independent of social and vocational outcomes, or experienced burden of disease. The alternative *DSM-5* proposal emanates—much in line with recent psychodynamic models—from a model of healthy adaptive functioning, which extends beyond the mere absence of pathological personality symptoms, as is implied by the Section II *DSM-5* PD criteria (PDM Task Force, 2006). Instead, healthy personality functioning emerges from the gradual development of adaptive capacities related to the way one perceives oneself and relates to others. Impairments can vary in terms of severity, resulting in personality problems that often underpin other mental disorders, such as mood or anxiety disorders, or in severe PDs, often—but not necessarily—characterized by complex comorbidity and restricted social and professional functioning. This approach to personality pathology makes it useful to assess areas of impairments in self and interpersonal functioning for each patient, independent of the obvious presence of a categorical PD diagnosis.

Initial statements from the *DSM-5* work group suggested that assessment of severity of impairments as operationalized in the LPFS might be a relatively easy task:

If not evident from chief complaints or the history of the presenting problems, a few basic questions about how patients feel about themselves and about the nature of their relationships with others should enable clinicians to say with some confidence whether a personality problem exists. (Skodol et al., 2011, p. 24)

However, several researchers criticized the concept for being abstract, vague, and theory-laden (Pincus, 2011; Shedler et al., 2010). Zimmermann and colleagues (2014) pointed out that reliable and valid ratings would likely require extensive training, several years of clinical experience, and long periods of observation or contact with the target patient. Supporting these concerns were the findings by Few et al. (2013), who asked trained graduate student interviewers to rate the level of personality functioning (LPF) after administration of the Structured Clinical Interview for *DSM-IV* Axis II Personality Disorders (SCID-II; First, Spitzer, Gibbon, Williams, & Benjamin, 1996). These videotaped interviews were double-rated by a second interviewer and resulted in interrater reliability estimates ranging from .47 to .49 for the elements of identity, self-direction, empathy, and intimacy, all well below the threshold of .75 put forward by Bender and colleagues (2011).

A series of studies on the German Operationalized Psychodynamic Diagnosis (OPD) by the Zimmermann group (2014, 2012) further highlights the seemingly complex nature of the assessment of severity in terms of underlying personality processes. The OPD interview operationalizes personality-related constructs, some of which are similar to those of the LPFS. This interview requires 60 hr of training, 10–15 hr of which focus on the LPFS-related structure axis, the OPD Levels of Structural Integration Axis (OPD-LSIA; Zimmermann et al., 2012). Zimmermann and colleagues (2014) demonstrated that it was feasible for inexperienced and minimally trained students to apply the LPFS on the OPD interview material, reaching acceptable estimated reliability ratings (intraclass correlation [ICC] = .51). The authors of the study concluded that assessing the LPFS might require less-extensive training or clinical experience than was assumed by many critics of the model. However, the same results could also be read in a different way. First, the estimated reliability ratings remained well short of the .75 threshold (Zimmermann et al., 2014). Moreover, it deserves mention that the rated interviews were conducted by experienced and thoroughly trained clinicians, “who are supposed to draw on the interviewee’s emphases, omissions, nonverbal behaviors, and reenactments, as well as on his or her own countertransference” (Zimmermann et al., 2012, p. 8). Furthermore, Zimmermann and colleagues (2012) pointed out that “these additional data sources are of great importance because impairments in basic capacities are especially apparent in how the interviewee copes with the demands of the interview situation” (p. 8). It is therefore highly questionable whether students or untrained clinicians would be able to gather the rich interview material the ratings were based upon. Moreover, the average interview duration was 74 min, making it less suitable for many standard assessment procedures (Zimmermann et al., 2014). However, as Zimmermann et al. (2014) pointed out, the OPD-LSIA was not designed in accordance with the LPFS criteria, and reliability of the ratings may well be improved when the interview focuses more explicitly on the con-

cepts of the LPFS. In sum, there is a clear need for a time-efficient, reliable instrument that is specifically oriented to the LPFS.

The Current Investigation

This study describes the development and preliminary psychometric evaluation of an interview schedule that was developed specifically for assessing the level of personality functioning as operationalized by the LPFS in Section III of the *DSM-5*: the Semi-Structured Interview for Personality Functioning *DSM-5* (STiP-5.1). To our knowledge, this is the first interview schedule that was developed with this specific aim, although the *DSM-5* work group members are currently developing the SCID-AMPD (D. S. Bender, personal communication, October, 29, 2014). We first describe the interview in detail and then present data on its interrater reliability and associations with *DSM-IV* diagnoses of PDs as well as with self-report measures of (mal)adaptive functioning and symptom severity (much in line with the research agenda for Criterion A put forward by the work group; Morey et al., 2011).

Development and Description of the Semi-Structured Interview for Personality Functioning *DSM-5* (STiP-5.1)

The Semi-Structured Interview for Personality Functioning *DSM-5* (STiP-5.1) was developed by the Centre of Expertise on PDs in the Netherlands (Kenniscentrum Persoonlijheidsstoornissen). The aim was to develop a relatively brief (i.e., between 30 and 60 min) interview schedule that would yield a reliable multi-item assessment of the facets constituting the Level of Personality Functioning Scale. Moreover, its format should be sufficiently user-friendly such that, after only a brief training, clinicians without specialized experience would be able to competently administer it. The interview schedule was developed, tested, discussed, refined, and tested again through a series of case interviews before a first version of the interview (Berghuis, Hutsebaut, Kaasenbrood, De Saeger, & Ingenhoven, 2013) was tested in a preliminary pilot study (Kamphuis, De Saeger, & Hutsebaut, 2014). On the basis of the pilot results, the interview schedule was significantly modified, and interview techniques and instructions were described in a manual in order to guide standardized administration of the interview. The result was the STiP-5.1 (Hutsebaut, Berghuis, De Saeger, Kaasenbrood, & Ingenhoven, 2014), which was used and evaluated in this study.

The interview schedule¹ features three columns and is organized around the facets of the LPFS. The criteria for the different levels of each facet were translated in Dutch and are displayed in the left column. On the basis of these criteria, the developers determined what information should be collected in order to rate the different LPF levels. These required aspects are described in the outer right column, providing the interviewer with a “shorthand” version for each section of the interview. Finally, the middle column contains the specific questions that should be posed to the patient. To illustrate, we describe the structure for the first facet: experience of oneself as unique, with clear boundaries between self and others. Information about two aspects of this facet should be collected in order to differentiate between the different levels of this facet according to the LPFS: (a) whether respondents have a clear idea

of who they are and (b) how well respondents can maintain this “sense of self” in stressful situations and/or in contact with other people. Each new aspect of information is represented by an open question (“How would you describe yourself?” respectively, “To what extent are you able to be really yourself, even in stressful situations or in contact with others?”). The open question is followed by a couple of auxiliary questions, which often ask more directly about the different criteria (e.g., “Does it happen to you that you adapt yourself so much to the expectations of others that you feel like you are no longer your usual self?”). The interviewer is required to score each facet of the LPFS before proceeding to the next section of the interview.

The LPFS contains 60 descriptors of severity (12 facets \times 5 levels each). To check each of these 60 descriptors separately might lead to a very long interview indeed. The authors therefore opted for a “funnel” strategy, aimed at narrowing down possible levels of impairment through a sequence of questions (Hutsebaut et al., 2014). Each section of the interview starts with a broad open question (e.g., Facet 1 [experience of oneself as unique, with clear boundaries between self and others]: How would you describe yourself? What kind of person are you?). Typically, on the basis of the information provided by interviewees, it is quite feasible to exclude two or three levels of severity (e.g., because interviewees do not have a clear sense of who they are) and use further questioning to narrow down between the two or three remaining levels. In other words: Contingent upon the answer to the open question, interviewers may use auxiliary questions to subsequently focus on the remaining levels, using the answers to differentiate between the remaining levels of impairment for that facet (e.g., Facet 1: “Do you sometimes feel ‘empty’ or that you no longer know who you are?” referring to Levels 3 and 4). Finally, the assumed level is checked by reframing respondents’ information in the terms that match the explicit description of the level in the LPFS (i.e., through the use of specific check and/or test questions; e.g., “Am I correct that you are saying you often do not know who you are, feeling empty, and recognizing within yourself a tendency to adapt overly to the people you are with?” as a check for Level 3). By not having to follow up on all levels of functioning, this strategy allows for a more time-efficient interview schedule, and matching the collected information to the stated descriptions in the LPFS enhances reliability. In our experience, this procedure allows for a transparent and collaborative interview style and reduces overly subjective appraisals. In the STiP-5.1 manual, specific examples of testing or checking questions are provided for each facet and each level.

Ratings of each facet should be performed during the interview. Ultimately, the STiP-5.1 is a clinician-rated interview, and clinician are encouraged to use their clinical judgment in making the final ratings. However, to avoid overly subjective appraisals, the ratings should first and foremost be based upon the patient’s report. Furthermore, interviewers are encouraged to give one score only (as opposed to leaving more than one option open). Scores for the domains as well as the total score are based upon a clinical evaluation of the separate ratings for the respective comprising facets or domains. Accordingly, when, for example, self-

¹ A free (Dutch or English) copy of the interview schedule and manual can be downloaded from www.kenniscentrumtrumps.nl

functioning is more impaired than interpersonal functioning, the interviewers are asked to use their judgment to make an overall rating, thus indicating their best estimate of the patient's level of personality functioning.

Method

Participants

Participants were recruited from a clinical ($n = 80$) and a community ($n = 18$) sample. Participants in the clinical sample were treatment-seeking adults who were referred to De Viersprong, the Netherlands Institute for Personality Disorders, a mental health care center specialized in the assessment and treatment of adolescents and adults with personality pathology. All interviews were conducted between July 2013 and August 2014. A total sample of 80 patients were included, 53 of whom (66.3%) were female. Their age ranged from 16 to 61 years, with a mean age of 33.6 ($SD = 12$). Participants from the community sample were recruited through a call for participation among relatives, friends, and neighbors of personnel working at De Viersprong. A first series of interviews were conducted in July 2014, and a second in February 2015. A total sample of 18 nonclinical participants were included, 16 of whom (84.2%) were female. Their age ranged from 18 to 60 years, with a mean age of 39 ($SD = 14.5$). None of these participants had been in treatment for mental disorders in the last 5 years.

Procedure

In addition to the standard admission procedure, which included semistructured interviews for the assessment of Axis I and Axis II disorders, as well as several self-report questionnaires, the STiP-5.1 was completed by all referred patients. Patients were informed about the goals, procedure, and status of the interview. After agreeing to participate, patients signed the informed-consent forms. Extensively trained psychologists, all participating in regular booster sessions to avoid interview drift, administered the structured diagnostic interviews: the Structured Clinical Interview for *DSM-IV* Axis I Personality Disorders (SCID-I; First, Spitzer, Gibbon, & Williams, 1997) and the SCID-II (First et al., 1996). The STiP-5.1 interview was administered by another psychologist, who was given only the name, gender, and age of the participant. The interviewer explained the procedure and goals of the interview, asked permission to videotape the interview, and asked informed consent from the patient to use the recording for scientific purposes, including rescoring of the interview by an independent rater. The interviewer then conducted the interview, rated the LPFS, and registered the patient's scores in the his or her file. Recorded interviews were uploaded on a secure server by the interviewer. A subsample of these interviews ($n = 40$ in the clinical sample) was used to perform reliability ratings. Therefore, a second rater, who was equally uninformed concerning the patient's personal and clinical background, watched the interview and independently applied the LPFS on the interview. Finally, a research collaborator built a data file with both ratings and additional sociodemographic, diagnostic, and other clinical variables. The procedure for participants from the community sample was largely the same, except that the interview was not part of the

standard admission procedure and only additional sociodemographic information was collected. No additional diagnostic interviews (SCID-I and SCID-II) or Self-Report Questionnaires were administered to the community participants.

Measures

Semi-Structured Interview for Personality Functioning *DSM-5* (STiP-5.1). The STiP-5.1 (Hutsebaut et al., 2014) was described extensively in the introduction. The interview consists of 28 open questions, with optional clarifying questions. The average interview duration in this study in the clinical sample was 50 min ($SD = 9.3$, range = 28–70).

Structured Clinical Interview for *DSM-IV* Axis I Disorders (SCID-I). The SCID-I (First et al., 1997; translated by van Groenestijn, Akkerhuis, Kupka, Schneider, & Nolen, 1999) is a semistructured interview designed to assess for *DSM-IV* Axis I disorders. The SCID-I has demonstrated good interrater reliability in a diversity of samples, especially when interviewers had received a formal training (overall $\kappa = .85$; Ventura, Liberman, Green, Shaner, & Mintz, 1998).

Structured Clinical Interview for *DSM-IV* Axis II Personality Disorders (SCID-II). The SCID-II (First et al., 1996, translated by Weertman, ArntZ, & Kerkhofs, 1996) was used to diagnose Axis II PDs. Criteria were scored when the clinician deemed sufficient evidence present that the targeted behaviors were present and pathological, pervasive, and persistent. Participants were classified as having a personality disorder not otherwise specified (PDNOS) when five criteria from personality disorders were present (Verheul, Bartak, & Widiger, 2007). The SCID-II has good interrater and test–retest reliability in PD samples (see, e.g., Maffei et al., 1997; Weertman, ArntZ, Dreessen, Van Velzen, & Vertommen, 2003) with sum ICCs reported as high as .90 for avoidant and .95 for borderline PD in a Dutch sample (Lobbestael, Leurgans, & ArntZ, 2011).

Brief Symptom Inventory (BSI). The BSI (Derogatis, 1975; translated by De Beurs, 2006) was used to assess symptom severity. It consists of 53 items covering nine symptom dimensions. The present study utilized only the BSI total score, which provides an index of the intensity of distress by psychological symptoms during the past week. Respondents rate each item on a 5-point scale ranging from 0 (*not at all*) to 4 (*extremely*). Cronbach's alpha in the present sample was high ($\alpha = .95$).

Severity Indices of Personality Problems (SIPP-118). The SIPP-118 (Verheul et al., 2008) is a dimensional self-report measure designed to assess core components of (mal)adaptive personality functioning. The SIPP-118 asks respondents to think back over the past 3 months and to answer the extent to which they agree with the presented statements. The response categories range from 1 (*fully disagree*) to 4 (*fully agree*). The measure comprises 16 facets, which are clustered into five higher order domains: self-control, identity integration, relational capacities, social concordance, and responsibility. High scores indicate better adaptive functioning. The subscales constituting the SIPP-118 have generally yielded adequate to strong internal consistency in PD samples, with alpha scores ranging from .62 to .89 (Verheul et al., 2008; Feenstra, Hutsebaut, Verheul, & Busschbach, 2011). Internal consistencies in the current sample were consistent with these estimates ($\alpha_s = .72$ –.86).

Interviewers

The STiP-5.1 in the clinical sample was administered by 12 psychologists with varying levels of experience and training. None of the interviewers were involved in the development of the STiP-5.1 interview. All interviewers were given a basic training of 3 hr (by Joost Hutsebaut), providing them with background information on the *DSM-5*, the AMPD, and the STiP-5.1. Training included one self-recorded video demonstration of the STiP-5.1. Prior to entering the study, all interviewers first conducted two pilot interviews to develop facility with its format. All interviews of sufficient audio quality were used in this study until we reached the target number of 40 clinical interviews. The interviews were independently rated by one of the authors (Joost Hutsebaut, Hilde De Saeger, or Dine J. Feenstra) of this article. Subsequent interview scores from the initial interviewers were collected to perform validity ratings. The STiP-5.1 in the community sample was administered by Joost Hutsebaut and Hilde De Saeger. Independent ratings were performed by Joost Hutsebaut, Hilde De Saeger, and Dine J. Feenstra.

Results

Descriptive Analyses: Clinical Characteristics

Clinical characteristics of the sample are presented in Table 1. As can be seen, the predominant PD diagnoses were PD Not

Table 1
Prevalence of DSM Diagnoses and Descriptive Statistics of Self-Reported Personality Functioning and Symptom Severity in the Clinical Sample (N = 80)

| Variable | n | % | M | SD |
|-------------------------------|----|------|-----------------|------|
| Axis I diagnoses ^a | | | | |
| Anxiety disorders | 27 | 33.8 | | |
| Mood disorders | 30 | 37.5 | | |
| Somatization disorders | 11 | 13.8 | | |
| Eating disorders | 7 | 8.8 | | |
| Substance use disorders | 2 | 2.5 | | |
| Psychotic disorders | 1 | 1.3 | | |
| Any Axis I diagnosis | 54 | 67.5 | | |
| Axis II diagnoses | | | | |
| Avoidant PD | 11 | 13.8 | | |
| Obsessive-compulsive PD | 1 | 1.3 | | |
| Schizoid PD | 1 | 1.3 | | |
| Narcissistic PD | 5 | 6.3 | | |
| Borderline PD | 22 | 27.5 | | |
| PD not otherwise specified | 27 | 33.8 | | |
| Any PD | 64 | 80 | | |
| SIPP-118 | | | | |
| Self-control | | | 4.37 (low) | 1.10 |
| Identity integration | | | 3.39 (low) | .72 |
| Responsibility | | | 4.58 (average) | .91 |
| Relational capacities | | | 3.83 (low) | .75 |
| Social concordance | | | 5.58 (average) | .89 |
| BSI total score | | | 1.56 (veryhigh) | .68 |

Note. Axis I and Axis II disorders that are not mentioned in the table were not diagnosed; SIPP-118 scores were compared to the norms of a normal population; DSM = *Diagnostic and Statistical Manual of Mental Disorders*; PD = personality disorder; SIPP-118 = Severity Indices of Personality Problems; BSI = Brief Symptom Inventory.

^a Only current Axis I diagnoses are displayed in this table.

Otherwise Specified (PDNOS; 33.8%), borderline PD (27.5%), and avoidant PD (13.8%). The most prevalent Axis I comorbidity concerned mood (37.5%) and anxiety (33.8%) disorders.

Reliability

Interrater reliability was computed using a one-way random, absolute agreement, single-measures ICC (McGraw & Wong, 1996) to assess the degree to which interviewers provided the same ratings of personality functioning across participants. Internal consistency of the STiP-5.1 was high, with a Cronbach's alpha of .97 for the total scale and .94 for both the self-functioning and interpersonal functioning domains. Interrater reliability was good, with ICCs ranging from .81 to .92 in the total sample and .58 to .80 in the clinical sample (see Table 2).

Construct Validity

We conducted preliminary tests of construct validity, testing for (a) mean level differences between the clinical group and the control group, and within the clinical group for mean level differences between patients with or without a PD diagnosis; (b) associations with pertinent SCID-I- and SCID-II-based scores; and (c) cross-method associations with self-report measures of personality functioning (the SIPP-118) and symptom severity (BSI).

Group differences. A large effect size was observed for the mean level difference between STiP-5.1 total scores of patients in the clinical sample ($M = 2.63$, $SD = 0.66$) and the nonpatients from the community sample ($M = 0.56$, $SD = 0.51$), $t(96) = -12.42$, $p < .001$; $d = 3.27$. Likewise, within the clinical group, a large group effect was observed between patients with a PD diagnosis scored ($M = 2.80$, $SD = 0.54$) and those without a PD diagnosis ($M = 1.94$, $SD = 0.68$), $t(78) = -5.40$, $p < .001$; $d = 1.53$.

Associations with SCID-I and SCID-II diagnoses and features. Table 3 shows the correlations between STiP-5.1 scores and Axis I and Axis II diagnoses. No significant correlation was found between STiP-5.1 scores and the number of Axis I disorders. However, the STiP-5.1 total score showed a strong association with the number of *DSM-IV* PDs ($r = .56$). Moreover, the STiP-5.1 interpersonal domain was more strongly related to the number of PDs than was the self domain. ($r = .52$ vs. $.39$; $p = .01$). Furthermore, patients with more personality disorder features had significantly higher STiP-5.1 scores (on both the total score and the comprising domains), with moderate to large effect sizes (r s ranging from .45 to .54). Table 3 also provides additional information on associations between STiP-5.1 scores and specific PD features for those specific PDs for which the sample included more than one participant, showing borderline PD to be the most impaired type of PD (r s ranging from .29 to .44).

Associations with self-report measures of personality problems and symptom severity. The total STiP-5.1 score and its interpersonal functioning and self-functioning domains were associated with all SIPP-118 domains of personality functioning (see Table 4). It is interesting that, whereas the STiP-5.1 total score and self-functioning domain scores were significantly associated with symptom severity as measured by the BSI (r s = .41 and .43, respectively), no significant associations emerged with the interpersonal domain scores. Table 4 shows the correlation ma-

Table 2
Interrater Reliability: ICC per Facet, Aspect, and Domain of the STiP 5.1

| Scale | Clinical (<i>n</i> = 40) | Total (<i>n</i> = 58) |
|--|------------------------------|---------------------------|
| STiP-5.1 total score | .71 | .89 |
| Self-functioning ^a | .78 | .91 |
| Identity ^b | .76 | .90 |
| Experience of oneself as unique, with clear boundaries between self and others | .79 | .91 |
| Self-esteem | .77 | .89 |
| Emotions | .66 | .87 |
| Self-direction ^b | .64 | .90 |
| Goals | .65 | .86 |
| Norms | .65 | .85 |
| Self-reflection | .58 | .85 |
| Interpersonal functioning ^a | .79 | .91 |
| Empathy ^b | .79 | .87 |
| Understanding others | .67 | .86 |
| Perspectives | .69 | .81 |
| Impact | .58 | .81 |
| Intimacy ^b | .80 | .92 |
| Connection | .82 | .88 |
| Closeness | .60 | .88 |
| Mutuality | .80 | .89 |

Note. All row labels that are not footnoted are facets of personality functioning. ICC = intraclass correlation; STiP-5.1 = Semi-Structured Interview for Personality Functioning *DSM-5*.

^a Domain of personality functioning. ^b Element of personality functioning; normal font = facet of personality functioning

trix, along with Fischer *r* to *z* tests of between-groups differences in correlations. Generally, the observed pattern of correlations showed (slightly) stronger associations between matching domains; that is, STiP-5.1 self-functioning was more highly correlated with the SIPP-118 scale Identity Integration, and the STiP-5.1 interpersonal domain showed a higher correlation with the SIPP-118 scale Social Concordance.

Discussion

This study explored the psychometric properties of the STiP-5.1, a newly developed interview instrument for the assessment of level of personality functioning closely aligned to the *DSM-5* Section III's Level of Personality Functioning Scale (LPFS; Bender et al., 2011). Our study demonstrated good to excellent interrater reliability of the interview-based ratings. Internal consistency and associations between domains and elements were remarkably high. Furthermore, our findings generally provide support for the construct validity for the instrument. The interview-based LPF scores clearly differentiated community from clinical participants, as well as patients with and without personality disorder. Moreover, STiP-5.1 ratings were consistently associated with both (diagnostic) interview-based and self-report measures of severity of personality problems. Finally, STiP-5.1 ratings were more strongly associated with personality pathology measures than with measures of general symptom stress, and borderline PD was confirmed as a particularly severe type of PD within this clinical sample. Taken together, our findings provide substantial albeit preliminary support for the suitability of the STiP-5.1 as an instrument for reliable and valid assessment of the level of severity of personality pathology, as explicated in the AMPD in the *DSM-5*.

The pattern of reliability findings point to some relative strengths and weaknesses of the instrument. First, in view of the satisfactory to good interrater reliability, we were successful in our aim of developing an interview that yielded consistent ratings across clinicians. In terms of consistency across raters, the STiP-5.1 clearly outperformed ratings of the level of personality functioning as based upon SCID-II interviews (Few et al., 2013) or upon a psychodynamic interview (Zimmermann et al., 2014). Moreover, the observed reliability rates met or approximated the threshold put forward by the *DSM-5* work group (Bender et al., 2011). These results are particularly remarkable because prior to the study none of the interviewers were familiar with the alternative model for personality disorders or with the LPFS. Thus, after a basic 3-hr training and minimal practice (i.e., two trial interviews), these (nonexpert) clinicians were able to competently administer the STiP-5.1 and produce ratings of satisfactory to good reliability. Of note, the interrater reliability of a previous version of the STiP interview schedule (Berghuis et al., 2013) was much lower. Upon close inspection of the pattern of findings, we significantly restructured our interview format. The main changes were (a) the revised interview schedule focused on each facet of the LPFS separately, (b) the questions in the revised interview format were more closely aligned to the severity criteria as described in the LPFS, (c) we introduced control questions that required the interviewer to reformulate the information collected during the interview and match it with the descriptive criteria of the LPFS, and (d) we wrote an extensive manual that included examples of control questions for all levels and all facets. It may be that such a structured, systematic approach is essential to ratings of sufficient reliability.

The high internal consistency of the STiP-5.1 and, relatedly, the strong intercorrelations of its facets and domains may point to an area for further development of the instrument. On the one hand, (aspects of) self and interpersonal functioning have been theorized as dynamically related constructs (Blatt, 2004). Clearly, from a developmental perspective, the ability to develop a sense of self (self) is inextricably interwoven with the capacity to engage with others (interpersonal). As such, significant associations between ratings of self-functioning and interpersonal functioning should be expected, and our findings might also be considered as supporting the validity of using a single rating

Table 3
Correlations of STiP-5.1 Scores With SCID-I and SCID-II Indices

| Variable | STiP-5.1 Total | Interpersonal | Self |
|---------------------------------|----------------|---------------|-------|
| Interpersonal functioning | .89** | | |
| Self-functioning | .90** | .87** | |
| No. of PDs | .56** | .52** | .39** |
| No. of PD features | .54** | .45** | .48** |
| No. of Axis I diagnoses | .10 | .02 | .01 |
| Avoidant PD (<i>n</i> = 11) | .01 | -.09 | -.03 |
| Narcissistic PD (<i>n</i> = 5) | .15 | .17 | .08 |
| Borderline PD (<i>n</i> = 22) | .44** | .29** | .39** |
| PDNOS (<i>n</i> = 27) | .01 | .17 | -.02 |

Note. *N* = 77–80 (variation is due to missing values). STiP-5.1 = Semi-Structured Interview for Personality Functioning *DSM-5*; SCID-I = Structured Clinical Interview for *DSM-IV* Axis I Disorders; SCID-II = Structured Clinical Interview for *DSM-IV* Axis II Personality Disorders; PD = personality disorder; PDNOS = personality disorder not otherwise specified.

** *p* = .01.

Table 4
Pearson Correlations With Self-Report Measures of Personality Problems and Symptom Severity
($N = 79$)

| Measure | STiP-5.1 total score | Interpersonal functioning | Self-functioning | Diff (p) |
|-----------------------|----------------------|---------------------------|------------------|--------------|
| SIPP-118 | | | | |
| Self-control | -.40** | -.40** | -.47** | .17 |
| Identity integration | -.41** | -.32** | -.43** | .04* |
| Responsibility | -.36** | -.44** | -.35** | .09 |
| Relational capacities | -.51** | -.47** | -.42** | .33 |
| Social concordance | -.40** | -.39** | -.25* | .01** |
| BSI total score | .41** | .24 | .43** | <.001** |

Note. STiP-5.1 = Semi-Structured Interview for Personality Functioning *DSM-5*; Diff = p -score of the difference between the Interpersonal functioning and Self-functioning correlation; SIPP-118 = Severity Indices of Personality Problems; BSI = Brief Symptom Inventory.

* $p = .05$. ** $p = .01$.

for level of personality functioning, as proposed by the AMPD. However, the strong overlap might also in part be due to a rating artifact. The STiP-5.1 interview draws on the introspective capacities of the patients, and at times patients struggled with the questions they were asked. When, even after using the specific probes and control questions, no full clarity was achieved, raters were still required to make their best estimate. It seems reasonable that clinician raters rather quickly develop a global impression of the level of personality functioning of the patient and then draw on this impression to “fill in the blanks.” Or alternatively, raters may have had a tendency to assign scores in line with the previous ratings for this patient. Such cognitive processes might have a homogenizing effect on the ratings. These reservations notwithstanding, there was also some evidence supporting the discriminant validity of the STiP-5.1 self and interpersonal functioning domains, as reflected in the pattern of theoretically meaningful differential correlations with dimensions of self-reported personality functioning (SIPP-118), and symptom severity (BSI). Nevertheless, strengthening the separate evaluation of its facets and domains may be a good target for future development, or alternatively, shorter versions of the present interview may be explored.

Our findings also further elucidate the nomological net of the construct of level of personality functioning. First, there is clear evidence that the STiP-5.1 is successful in capturing personality pathology rather than symptomatic stress in general. Second, from their moderate intercorrelation, it is also evident that the number of personality disorder features and the LPF as measured by the STiP-5.1 are not identical. Further research may clarify to what extent the LPF adds relevant information to the assessment of personality pathology that could inform clinical decisions such as informing treatment assignment or predicting ruptures in treatment relationships.

Some strengths and limitations of the present study are notable. First, we believe that our design is strong on ecological validity. The administration of the STiP-5.1 was integrated in the regular assessment procedure of the De Viersprong setting, and the clinical interviews were conducted by its regular staff clinicians. Another strength of our study is the rather large clinical sample, along with the use of high-quality external measures, that is, the extensive assessment of Axis I and Axis II pathology using structured interviews. On the other hand, our community sample was smaller and demographically rather homogeneous. Moreover, interviewers were not blind to clinical status, which may have caused some bias among raters, who may have expected healthier scores from community respondents. However, the

concerns about potential bias are somewhat mitigated by the fact that the STiP-5.1 scores also differentiated within the clinical sample between PD and non-PD patients, even in the absence of information about the PD diagnosis in both interviewer and patient. Finally, our clinical sample consisted mostly of Cluster B or C patients, which presents a limitation of the present study. It remains to be established in future research whether the interview is also suited for use with patients with Cluster A PDs or patients with antisocial PDs. More specifically, it is an interesting issue whether an interview schedule that strongly draws upon self-report such as the STiP-5.1 will require taking in observational data too for assessing level of impairment. If so, this might impact upon the level of training necessary in order to combine all relevant sources of data—self-report, observation, and interpretation—for assessing severity of personality pathology. At De Viersprong, we also have started administering the STiP-5.1 to adolescents (starting from the age of 12), but psychometric data are still forthcoming. Finally, future research may focus on the predictive validity of the STiP-5.1 for various treatment outcomes, including response to treatment, dropout, and crisis.

References

- 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.). Arlington, VA: American Psychiatric Publishing.
- Bateman, A., & Fonagy, P. (2013). Impact of clinical severity on outcomes of mentalisation-based treatment for borderline personality disorder. *British Journal of Psychiatry*, *203*, 221–227. <http://dx.doi.org/10.1192/bjp.bp.112.121129>
- 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*, 332–346. <http://dx.doi.org/10.1080/00223891.2011.583808>
- Berghuis, H., Hutsebaut, J., Kaasenbrood, A., De Saeger, H., & Ingenhoven, T. (2013). *Semi-Structured Interview for Personality Functioning DSM-5 (STiP-5)*. Utrecht, the Netherlands: Trimbos Institute.
- Blatt, S. (2004). *Experiences of depression: Theoretical, clinical, and research perspectives*. Washington, DC: American Psychological Association. <http://dx.doi.org/10.1037/10749-000>
- Bernstein, R. F. (1998). Reconceptualizing personality disorder diagnosis in the *DSM-5*: The discriminant validity challenge. *Clinical Psychology*:

- Science and Practice*, 5, 333–343. <http://dx.doi.org/10.1111/j.1468-2850.1998.tb00153.x>
- De Beurs, E. (2011). *Brief Symptom Inventory 18 (BSI): Handleiding [Manual]*. Leiden, the Netherlands: PITS.
- Derogatis, L. R. (1975). *Brief Symptom Inventory*. Baltimore, MD: Clinical Psychometric Research.
- Feenstra, D. J., Hutsebaut, J., Verheul, R., & Busschbach, J. J. V. (2011). Severity Indices of Personality Problems (SIPP-118) in adolescents: Reliability and validity. *Psychological Assessment*, 23, 646–655. <http://dx.doi.org/10.1037/a0022995>
- Few, L. R., Miller, J. D., Rothbaum, A. O., Meller, S., Maples, J., Terry, D. P., . . . MacKillop, J. (2013). Examination of the Section III *DSM-5* diagnostic system for personality disorders in an outpatient clinical sample. *Journal of Abnormal Psychology*, 122, 1057–1069. <http://dx.doi.org/10.1037/a0034878>
- First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1997). *Structured Clinical Interview for DSM-IV Axis I Disorders*. Washington, DC: American Psychiatric Press.
- First, M. B., Spitzer, R. L., Gibbon, M., Williams, J. B. W., & Benjamin, L. (1996). *Structured Clinical Interview for DSM-IV Axis II Personality Disorders (SCID-II)*. Washington, DC: American Psychiatric Press.
- 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. <http://dx.doi.org/10.1521/pepi.2011.25.3.305>
- Hutsebaut, J., Berghuis, H., De Saeger, H., Kaasenbrood, A., & Ingenhoven, T. (2014). *Semi-Structured Interview for Personality functioning DSM-5 (STiP 5.1)*. Utrecht, the Netherlands: Trimbos Institute.
- Kamphuis, J. H., De Saeger, H., & Hutsebaut, J. (2014). *Semi-Structured Interview for Personality functioning DSM-5 (STiP5): A feasibility study*. Paper presented at the annual meeting for the Society for Personality Assessment (SPA), Chicago, IL.
- Kim, Y. R., & Tyrer, P. (2010). Controversies surrounding classification of personality disorder. *Psychiatry Investigation*, 7, 1–8. <http://dx.doi.org/10.4306/pi.2010.7.1.1>
- Lobbestael, J., Leurgans, M., & ArntZ, A. (2011). Inter-rater reliability of the Structured Clinical Interview for *DSM-IV* Axis I Disorders (SCID I) and Axis II Disorders (SCID II). *Clinical Psychology & Psychotherapy*, 18, 75–79. <http://dx.doi.org/10.1002/cpp.693>
- 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. <http://dx.doi.org/10.1521/pepi.1997.11.3.279>
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30–46. <http://dx.doi.org/10.1037/1082-989X.1.1.30>
- McWilliams, N. (1994). *Psychoanalytic diagnosis*. New York, NY: Guilford Press.
- Morey, L. C., Bender, D. S., & Skodol, A. E. (2013). Validating the proposed *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed., severity indicator for personality disorder. *Journal of Nervous & Mental Disease*, 201, 729–735. <http://dx.doi.org/10.1097/NMD.0b013e3182a20ea8>
- Morey, L. C., Benson, K. T., Busch, A. J., & Skodol, A. E. (2015). Personality disorders in *DSM-5*: Emerging research on the alternative model. *Current Psychiatry Reports*, 17, 1–9. <http://dx.doi.org/10.1007/s11920-015-0558-0>
- 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, 347–353. <http://dx.doi.org/10.1080/00223891.2011.577853>
- Oleski, J., Cox, B. J., Robinson, J., & Grant, B. (2012). The predictive validity of Cluster C personality disorders on the persistence of major depression in the National Epidemiologic Survey on Alcohol and Related Conditions. *Journal of Personality Disorders*, 26, 322–333. <http://dx.doi.org/10.1521/pepi.2012.26.3.322>
- PDM Task Force. (2006). *Psychodynamic diagnostic manual*. Silver Spring, MD: Alliance of Psychoanalytic Organizations.
- Pincus, A. L. (2011). Some comments on nomenclature, diagnostic process, and narcissistic personality disorder in the *DSM-5* proposal for personality and personality disorders. *Personality Disorders: Theory, Research, and Treatment*, 2, 41–53. <http://dx.doi.org/10.1037/a0021191>
- Shedler, J., Beck, A., Fonagy, P., Gabbard, G. O., Gunderson, J., Kernberg, O., . . . Westen, D. (2010). Personality disorders in *DSM-5*. *American Journal of Psychiatry*, 167, 1026–1028. <http://dx.doi.org/10.1176/appi.ajp.2010.10050746>
- Skodol, A. E., Bender, D. S., Oldham, J. M., Clark, L. A., Morey, L. C., Verheul, R., . . . Siever, L. J. (2011). Proposed changes in personality and personality disorder assessment and diagnosis for *DSM-5* Part II: Clinical application. *Personality Disorders: Theory, Research, and Treatment*, 2, 23–40. <http://dx.doi.org/10.1037/a0021892>
- Tyrer, P., Reed, G. M., & Crawford, M. J. (2015, February 21). Classification, assessment, prevalence, and effect of personality disorder. *The Lancet*, 385, 717–726. [http://dx.doi.org/10.1016/S0140-6736\(14\)61995-4](http://dx.doi.org/10.1016/S0140-6736(14)61995-4)
- van Groenestijn, M. A. C., Akerhuis, G. W., Kupka, R. W., Schneider, N., & Nolen, W. A. (1999). *Gestructureerd klinisch interview voor de vaststelling van DSM-IV As-I stoornissen* [Structured clinical interview for the assessment of *DSM-IV* Axis-I disorders]. Lisse, the Netherlands: Swets Test.
- Ventura, J., Liberman, R. P., Green, M. F., Shaner, A., & Mintz, J. (1998). Training and quality assurance with the Structured Clinical Interview for *DSM-IV* (SCID-IP). *Psychiatry Research*, 79, 163–173. [http://dx.doi.org/10.1016/S0165-1781\(98\)00038-9](http://dx.doi.org/10.1016/S0165-1781(98)00038-9)
- Verheul, R., Andrea, H., Berghout, C. C., Dolan, C., Busschbach, J. J. V., van der Kroft, P. J. A., . . . Fonagy, P. (2008). Severity Indices of Personality Problems (SIPP-118): Development, factor structure, reliability, and validity. *Psychological Assessment*, 20, 23–34. <http://dx.doi.org/10.1037/1040-3590.20.1.23>
- Verheul, R., Bartak, A., & Widiger, T. (2007). Prevalence and construct validity of personality disorder not otherwise specified (PDNOS). *Journal of Personality Disorders*, 21, 359–370. <http://dx.doi.org/10.1521/pepi.2007.21.4.359>
- 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. <http://dx.doi.org/10.1521/pepi.17.6.562.25359>
- Weertman, A., ArntZ, A., & Kerkhofs, M. (1996). *Gestructureerd klinisch interview voor DSM-IV as II persoonlijkheidsstoornissen* [Structured Clinical Interview for *DSM-IV* Axis II personality disorders]. Amsterdam, the Netherlands: Harcourt Test.
- Zimmermann, J., Benecke, C., Bender, D. S., 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 of Personality Assessment*, 96, 397–409. <http://dx.doi.org/10.1080/00223891.2013.852563>
- Zimmermann, J., Ehrental, J. C., Cierpka, M., Schauenburg, H., Doering, S., & Benecke, C. (2012). Assessing the level of structural integration using Operationalized Psychodynamic Diagnosis (OPD): Implications for *DSM-5*. *Journal of Personality Assessment*, 94, 522–532. <http://dx.doi.org/10.1080/00223891.2012.700664>