Assessment
1–14
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10731911211015309
journals.sagepub.com/home/asm
SAGE

Development of Measures for the Hierarchical Taxonomy of Psychopathology (HiTOP): A Collaborative Scale Development Project

Leonard J. Simms[1], Aidan G. C. Wright[2], David Cicero[3], Roman Kotov[4], Stephanie N. Mullins-Sweatt[5], Martin Sellbom[6], David Watson[7], Thomas A. Widiger[8], and Johannes Zimmermann[9]

Abstract

In this article, we describe the collaborative process that is underway to develop measures for the Hierarchical Taxonomy of Psychopathology (HiTOP). The HiTOP model has generated much interest in the psychiatric literature in recent years, but research applications and clinical translation of the model require measures that are specifically keyed to the model. To that end, the Measures Development Workgroup of HiTOP has been engaged in a collaborative effort to develop both questionnaire and interview methods that (a) are specifically tied to the elements of the HiTOP structure, and (b) provide one means of testing that structure. The work has been divided among five subgroups that are focused on specific HiTOP spectra. Our scale development methods are rooted in the principles of construct valid scale development. This report describes Phase 1 of this project, summarizes the methods and results thus far, and discusses the interplay between measurement and HiTOP model revisions. Finally, we discuss future phases of the scale development and the steps we are taking to improve clinical utility of the final measures.

Keywords

HiTOP, psychiatric classification, scale development, construct validity

Introduction to the HiTOP Model and Consortium

The Hierarchical Taxonomy of Psychopathology (HiTOP; Kotov et al., 2017) is the product of a large, grass-roots consortium of mental health researchers who have come together to build a psychiatric classification system that is rooted in the quantitative classification tradition rather than consensus judgments of experts (e.g., Williams & Simms, 2020). Existing classification systems, such as the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*; American Psychiatric Association, 2013) and the *International Classification of Diseases* (ICD; World Health Organization, 2020) have long been noted for exhibiting important shortcomings that detract from their value in applied and research settings. Among the well-documented concerns with these systems are (a) excessive co-occurrence of disorders (i.e., comorbidity) that raises questions about the distinctiveness of many disorders (e.g., Clark et al., 2017); (b) significant within-diagnosis heterogeneity in which individuals with the same diagnostic label can present in markedly different ways (e.g., Galatzer-Levy & Bryant, 2013; Wright et al., 2013), (c) poor diagnostic reliability for numerous diagnostic categories (e.g., Chmielewski et al., 2015; Regier et al., 2013); (d) arbitrary diagnostic thresholds that erroneously imply that psychopathology varies in kind rather than degree (e.g., Aslinger et al., 2018; Carragher et al., 2014; Clark et al., 2017; Haslam et al., 2020; Markon et al., 2011); and (e) concerns regarding clinical utility (e.g., Ruggero et al., 2019).

[1]University at Buffalo, Buffalo, NY, USA
[2]University of Pittsburgh, Pittsburgh, PA, USA
[3]University of North Texas, Denton, TX, USA
[4]Stony Brook University, Stony Brook, NY, USA
[5]Oklahoma State University, Stillwater, OK, USA
[6]University of Otago, Dunedin, New Zealand
[7]University of Notre Dame, Notre Dame, IN, USA
[8]University of Kentucky, Lexington, KY, USA
[9]University of Kassel, Kassel, Germany

Corresponding Author:
Leonard J. Simms, Department of Psychology, University at Buffalo, Park Hall 218, Buffalo, NY 14221, USA.
Email: ljsimms@buffalo.edu

In response to such concerns, the HiTOP consortium was organized by Roman Kotov, Robert Krueger, and David Watson, in 2015, to develop an alternative psychiatric classification system that is rooted in quantitative evidence. Kotov, Krueger, and Watson assembled a group of psychologists and psychiatrists, from a variety of backgrounds and perspectives, who nonetheless were guided by the unifying goal of tethering psychiatric classification to empirical data rather than to expert consensus. In particular, the HiTOP consortium has been heavily influenced by the quantitative tradition in psychopathology, which is focused on using structural statistical methods (e.g., factor analyses, latent variable modeling, etc.) to identify important and distinct psychiatric phenomena based on patterns of symptom/feature covariation (Williams & Simms, 2020).

The consortium currently includes over 140 members who span psychology, psychiatry, and neuroscience, and the group is actively working to build bridges to other allied professions. Work in the consortium is divided among multiple workgroups focused on a range of topics relevant to psychiatric classification (e.g., quantitative methods, genetics, measurement, personality, clinical translation, and neurobiology). The work summarized in this article and special issue more generally is the product of the Measures Development Workgroup, which is the largest workgroup in the consortium. The mission of the workgroup is to develop HiTOP-specific measures, use the measure development process to provide guidance to the broader consortium regarding the lower order symptom and trait dimensions that serve as the foundation on which the full HiTOP model is built, and provide a measurement basis for informing the full hierarchical structure of HiTOP. Our goals with this article are to (a) briefly describe the HiTOP model and consortium and the rationale for a comprehensive HiTOP-specific measure, (b) describe the steps we have taken toward that goal, and (c) discuss the remaining steps in the process and our future directions.

The HiTOP model has been fully described and depicted in numerous other publications (e.g., Kotov et al., 2017; Kotov et al., 2021; Ruggero et al., 2019). Here we summarize the model briefly and specifically describe how the measurement development efforts interface with the original model. The HiTOP model is hierarchical, meaning that it includes dimensions of psychopathology at multiple levels of generality versus specificity. See Figure 1 for a summary of the HiTOP classification hierarchy. At the top of the hierarchy, Kotov et al. proposed a broad level reserved for structural *superspectra* that have been identified in the quantitative classification literature (e.g., p Factor, or general psychopathology, Caspi et al., 2014; Lahey et al., 2012). Such superspectra have been shown to account for much of the statistical covariation in psychiatric symptomatology at a very broad level (i.e., little specificity), but the meaning of such superspectra remains a matter of substantial debate (e.g., Levin-Aspenson et al., 2020; Smith et al., 2020).

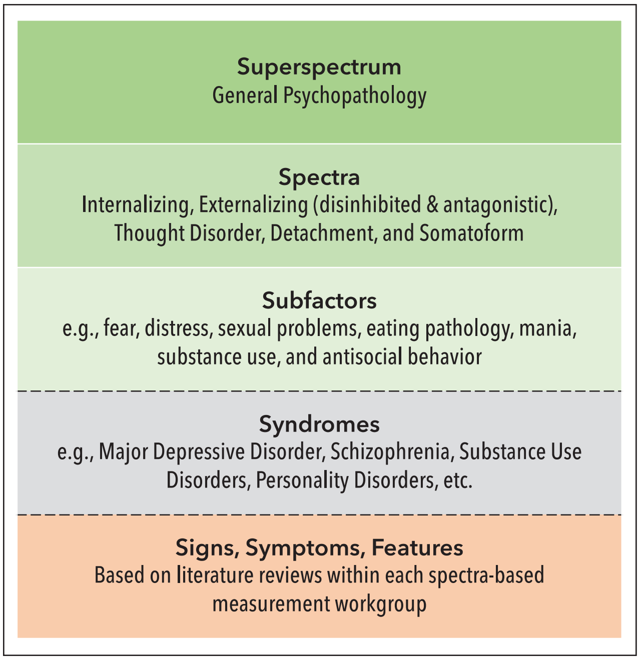
Superspectrum
General Psychopathology
Spectra
Internalizing, Externalizing (disinhibited & antagonistic), Thought Disorder, Detachment, and Somatoform
Subfactors
e.g., fear, distress, sexual problems, eating pathology, mania, substance use, and antisocial behavior
Syndromes
e.g., Major Depressive Disorder, Schizophrenia, Substance Use Disorders, Personality Disorders, etc.
Signs, Symptoms, Features
Based on literature reviews within each spectra-based measurement workgroup

Figure 1. HiTOP system levels of hierarchy.
Note. HiTOP measurement has been organized by spectra and focused on building scales from the bottom-up, focused on signs, symptoms, and features. HiTOP = Hierarchical Taxonomy of Psychopathology.

One level down, the model lists six psychopathology *spectra*—internalizing, disinhibited externalizing, antagonistic externalizing, thought disorder, detachment, and somatoform—the first five of which have been robustly identified across multiple literatures spanning normal-range personality (e.g., the five-factor model; Goldberg, 1993; McCrae & Costa, 1997), pathological personality (e.g., Alternative Model of Personality Disorders, American Psychiatric Association, 2013; personality psychopathology–5 model, Harkness et al., 1995; Harkness et al., 2012), and psychopathology measures more generally (e.g., Minnesota Multiphasic Personality Inventory–2 Restructured Form; see Sellbom, 2019, for a review). Notably, the somatoform spectrum has not been sufficiently studied with respect to its placement relative to the other five spectra, and so it currently exists as a separate provisional spectrum, awaiting additional research to guide its placement in the model (but see Watson et al., in press).

The next level down in the hierarchy summarizes what is known regarding *subfactors* of the spectra. For example, the model offers four subfactors for internalizing: fear, distress, eating pathology, and sexual problems. Mania is a subfactor that currently sits provisionally between the internalizing and thought disorder spectra. Substance abuse is a subfactor of the disinhibited externalizing spectrum, whereas antisocial behavior is nested between both disinhibited and antagonistic externalizing in the model. Notably, additional subfactors are likely as data on the model accumulate in the literature (e.g., Forbes et al., 2021).

In particular, the work of the Measures Development Workgroup is well positioned to provide evidence to support (or refute) current subfactors and to suggest new subfactors to be included in the model.

The fourth level depicted in the original hierarchy is where syndromes are represented (e.g., Major Depressive Disorder, Generalized Anxiety Disorder, and Schizophrenia). In the original model, these disorders are neatly nested under spectra and subfactors, based on studies showing how traditional disorders relate to structural models like HiTOP (Kotov et al., 2011; Krueger, 1999; Wright & Simms, 2015), but these placements likely are much more complicated given some of the known limitations of these disorders as summarized above. In particular, symptom heterogeneity within disorders means that certain aspects of traditional disorders likely "load" on different spectra or subfactors. For example, borderline personality disorder currently is listed under two spectra in the model, but it actually includes criteria that are related to four different aspects of the HiTOP model: the distress subfactor of internalizing, antagonistic externalizing, disinhibited externalizing, and thought disorder. Given examples like this, the syndrome level of HiTOP is best considered a rough cross-walk between *DSM*/ICD and other levels of the HiTOP model, rather than a distinct, cohesive level in its own right. To that end, in more recent depictions of the HiTOP model, the syndrome level is clarified to represent empirical syndromes derived from other levels of the HiTOP model rather than traditional *DSM*/ICD diagnoses per se.

Finally, at the lowest level of the hierarchy are listed the most narrow-band elements of the model: Specific signs, symptoms, and features of psychopathology. Kotov et al. (2017) provided a list of such lower order psychopathology features, but this level of the hierarchy remains the most tentative aspect of the model. Even a casual examination of the psychopathology measurement literature will reveal a wide range of models and measures of psychopathology features that could be used to identify the lower order signs, symptoms, and features of psychopathology, but none of these are comprehensive, and all have limitations relative to the others. As such, in the Measures Development Workgroup, we opted to focus our efforts on measuring this level of the hierarchy as a way to build a comprehensive set of measures for each broad HiTOP spectrum. Rather than be limited by the facets listed in the original HiTOP paper, we instead opted to use these as a starting point to building a more comprehensive list of candidate facets within each spectrum. Details of this process are provided next.

The Need for HiTOP-Specific Measures

The HiTOP consortium recognized early in the process that the identification and development of HiTOP measures would determine the ultimate impact of their work. Without adequate measures, the HiTOP model would risk being seen as an intellectually interesting yet practically useless exercise, which would represent a tremendous missed opportunity given the evidence showing the problems with traditional psychiatric classification methods. HiTOP requires measurement tools to fully realize its potential to transform the diagnosis of psychopathology in research and practice. In research, HiTOP-specific measures will be needed to fully study all elements in the model, including the placement of certain provisional elements (e.g., somatoform symptomatology, mania). In the clinic, HiTOP-specific measures are needed to offer practicing mental health clinicians with a viable alternative to traditional classification methods (Ruggero et al., 2019).

To that end, the consortium has worked along two independent routes to promote and develop HiTOP measures. First, the Clinical Translation Workgroup has identified a set of HiTOP-consistent measures that can be used immediately. Notably, as part of a field trial into the feasibility of HiTOP in clinical practice, the Clinical Translation Workgroup has developed the HiTOP Self-Report, a suggested battery of existing measures for all six spectra. HiTOP Self-Report is available as an online instrument free for use in clinical settings on request at https://hitop.unt.edu/hitop-clinical-field-trials. In addition, while waiting for HiTOP-specific measures, those interested in measuring aspects of the HiTOP model are encouraged to do so using one or more existing measures that, although not developed specifically for the HiTOP model, nonetheless represent ways of conceptualizing and measuring psychopathology in a manner consistent with this model. No single instrument measures the full model, but users may combine several instruments to measure a substantial portion of it. A nonexhaustive list of HiTOP-consistent measures is presented in Table 1. Some are free for use, whereas others require purchase. Moreover, many have extensive norms, have been adapted for use across different populations, and include other features that are valuable in clinical applications (e.g., validity scales).

Second, the Measures Development Workgroup has undertaken the task of developing both questionnaire and interview methods that (a) are specifically tied to the elements of the HiTOP structure (Kotov et al., 2017), and (b) provide one means of testing that structure. As noted above, measures certainly exist for all domains within HiTOP, so the primary need that the HiTOP measure is trying to fill is for a unified set of measures that span the full breadth of HiTOP constructs, something that does not yet exist in the assessment world. Having a unified set of measures is helpful in several ways. First, it will permit researchers and clinicians to measure the full model with a single, open-source and freely available method, as opposed to needing to string together measures of each domain separately, some of

Table 1. Examples of HiTOP-Friendly Measures.

Instrument	Coverage	How to access measure
Achenbach System of Empirically Based Assessment (ASEBA)	Internalizing and disinhibited externalizing spectra	https://store.aseba.org/
Child and Adolescent Psychopathology Scale (CAPS)	Internalizing and disinhibited externalizing spectra	https://www.parinc.com/Products/Pkey/9
Externalizing Spectrum Inventory (ESI)	Disinhibited and antagonistic externalizing spectra	Contact author: cpatrick@psy.fsu.edu
Inventory for Depression and Anxiety Symptoms (IDAS-II)	Internalizing spectrum	Contact author: db.watson@nd.edu
Interview for Mood and Anxiety Symptoms (IMAS)	Internalizing spectrum	https://renaissance.stonybrookmedicine.edu/system/files/IMASInterview.pdf
Scale for the Assessment of Negative Symptoms (SANS)	Thought disorder spectrum	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd000807.2
Scale for the Assessment of Positive Symptoms (SAPS)	Thought disorder spectrum	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd000837.1
Schedule for Nonadaptive and Adaptive Personality–2nd ed. (SNAP-2)	Personality disorder traits	Contact author: lclark6@nd.edu
Personality Inventory for *DSM-5* (PID-5)	Personality disorder traits	https://www.psychiatry.org/File%20Library/Psychiatrists/Practice/DSM/APA. . .
Five Factor Form (FFF)	Personality disorder traits	Contact author: widiger@uky.edu
Five-Factor Model Personality Disorder Scales	Personality disorder traits	Contact author: widiger@uky.edu
Comprehensive Assessment of Traits Relevant to Personality Disorder (CAT-PD)	Personality disorder traits	http://3plab.org/cat-pd/
Dimensional Assessment of Personality Pathology—Basic Questionnaire (BQ)	Personality disorder traits	https://www.sigmaassessmentsystems.com/assessments/dimensional-assessment-of-personality-pathology-basic-questionnaire/
Personality Assessment Inventory (PAI)	Mix of personality traits and psychopathology spectra/syndromes	https://www.parinc.com/products/pkey/287
Minnesota Multiphasic Personality Inventory–2 Restructured Form (MMPI-2-RF)	Mix of personality traits and psychopathology spectra/syndromes; covers all six HiTOP spectra	https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Personality-%26-Biopsychosocial/Minnesota-Multiphasic-Personality-Inventory-2-Restructured-Form/p/100000631.html

which may come with usage costs. This should improve efficiency and clinical translation due to such simplicity. Second, given the cross-domain analytic methods we plan for Phase 2 (described below), discriminant validity and redundancy across domains should be minimized in the HiTOP measure relative to other measures that exist for each domain separately. Third, the development of a full-model measure will facilitate studies of the HiTOP model and will provide an important source of data regarding future revisions of the model. Finally, given the diversity of timeframes, instruction sets, and response formats across exiting measures, development of a single measure that spans the model without the confound of differing methods is an important contribution of this endeavor (Watson et al., this issue).

The Measures Development Workgroup includes more than 40 members who belong to one of five psychopathology spectrum-based subgroups: internalizing psychopathology (chaired by David Watson), disinhibited and antagonistic externalizing psychopathology (chaired by Stephanie Mullins-Sweatt), thought disorder (chaired by David Cicero), detachment (co-chaired by Tom Widiger and Johannes Zimmermann), and somatoform and eating pathology (chaired by Martin Sellbom). The first two authors of this article serve as the workgroup's chair (LJS) and statistical advisor (AGCW), respectively. As with the HiTOP consortium as a whole, the Measures Development Workgroup members have volunteered their time and resources to support the development of strong and useful measures of the HiTOP model.

The subgroup themes largely parallel the HiTOP spectrum structure as hypothesized by Kotov et al. (2017). However, several pragmatic considerations influenced this process as well. First, although there are six spectra in the published HiTOP model, we have five spectrum-based subgroups: Although we initially had separate subgroups for antagonistic and disinhibited externalizing, as articulated in the HiTOP model, we combined them under a single Chair given their joint influence on externalizing behavior (e.g., substance abuse, antisocial behavior). Second, despite being listed as a subfactor within internalizing, eating pathology was assigned to the somatoform group to help balance the workload across groups. Third, mania, which is placed provisionally between internalizing and thought disorder in the HiTOP model, was assigned to internalizing for development, but the thought disorder group administered the same items in their data collections too. Finally, certain detachment constructs, which are theoretically relevant to both the thought disorder and detachment subgroups, were independently conceptualized and defined across those groups, and then the groups combined their work and administered the complete set of detachment items in both groups' data collections.

Phases of HiTOP Scale Development. Development of this new measure is proceeding through three phases of data collection and analyses, each with its own goals. The goals of Phase 1 were to develop conceptual definitions for constructs within each domain, build the initial item pool, collect multiple rounds of response to these items within each subgroup, and develop a set of preliminary scales. Phase 1 is now complete for four of five spectrum-based subgroups (the externalizing subgroup will complete Phase 1 by summer 2021. This article primarily is focused on describing the rationale and methods for Phase 1. The accompanying articles summarize the Phase 1 activities to date for each spectrum-based subgroup. Watson et al. (this issue) describe the Phase 1 results for the internalizing subgroup. Sellbom et al. (this issue) describe the Phase 1 results for the somatoform and eating pathology scale development efforts. Cicero et al. (this issue) describe the Phase 1 results for the thought disorder subgroup. Zimmermann et al. (this issue) describe the Phase 1 results for the detachment scale development work. Finally, Mullins-Sweatt et al. (this issue) described the Phase conceptual work they have completed thus far.

All of the constructs and items emerging from Phase 1 then will be combined together in Phase 2, the primary goal of which is to finalize the HiTOP scales. Finally, Phase 3 will focus on external validation of the HiTOP questionnaire and improving clinical utility. More on Phases 2 and 3 is presented below in the "Next Steps" section.

Building Construct Valid HiTOP Measures

The desire to build HiTOP-specific measures can be traced back to well before Kotov et al.'s (2017) introductory article was published. However, the work began in earnest in Baltimore in 2017, at the annual HiTOP meeting, where the group discussed many aspects of the scale development process, debating best practices both at the broad, superordinate level (e.g., how to build scales for a model that is designed to be dynamic and responsive to data) and at the level of specific measurement decisions that needed to be made (e.g., item format, response format, measure timeframe). We decided at the Baltimore meeting to ground our measurement efforts in the principles of construct valid scale development (Clark & Watson, 2019; Loevinger, 1957; Simms & Watson, 2007), which treat validity less as a static outcome and more of a process of dynamically infusing construct validity into the scales at all stages of the process.

Loevinger (1957) was the first to systematically describe a theory-driven method of scale development that is grounded in construct validity (Cronbach & Meehl, 1955). Cronbach and Meehl described the general process of construct validation as an exercise in theory development and testing. Loevinger (1957) extended their work to the specific process of scale development. She distinguished among three aspects of construct validity—*substantive validity*, *structural validity*, and *external validity*—that she argued are "mutually exclusive, exhaustive of the possible lines of evidence for construct validity, and mandatory" (pp. 653-654) and are

> closely related to three stages in the test construction process: constitution of the pool of items, analysis of the internal structure of the pool of items and consequent selection of items to form a scoring key, and correlation of test scores with criteria and other variables. (p. 654)

Modern treatments of Loevinger's scale development principles have been described in detail elsewhere (e.g., Clark & Watson, 1995, 2019; Simms & Watson, 2007). Here, we briefly describe construct valid scale development principles as the foundation on which the HiTOP measure development processes have been built.

Substantive Validity

The first element of Loevinger's (1957) scale development model—substantive validity—is centered on the tasks of construct conceptualization and development of the initial item pool. In this stage of the process, scale developers must carefully review the literature(s) related to the constructs they wish to measure, develop a definition of those constructs (i.e., a theory) that fully operationalizes them in terms of their required components, and build a broad item pool that attempts to assess all aspects of the construct as defined.

Construct List Development. In the HiTOP measure development process, the spectrum-based measurement subgroups first reviewed the literatures and existing measures relevant to their particular spectrum and identified an overinclusive list of candidate constructs to be measured. Overinclusiveness is important at the substantive validity phase, because elements not present at the start of the process are unlikely to emerge in the final measurement model. Next, the subgroups developed and edited definitions for each candidate construct based on the results of their literature reviews. This process resulted in a wide range of candidate constructs across subgroups. Given their breadth, the internalizing and externalizing groups generated the most candidate constructs—57 and 78, respectively—whereas the remaining subgroups each identified 20 or fewer constructs initially. The members within subgroups were selected for their diverse expertise relevant to the spectra they were assigned to measure. Thus, each subgroup internally debated the construct lists and iteratively edited the construct definitions until consensus was achieved. In addition, some subgroups (e.g., somatoform and eating pathology) consulted with outside experts as they deemed necessary to ensure that all important constructs were represented in the lists and properly defined.

Notably, the subgroups worked independently to build their construct lists and definitions, which resulted in some interesting areas of overlap that we permitted at this stage to err on the side of overinclusiveness. For example, anhedonia was hypothesized as being relevant across three subgroups: internalizing, detachment, and thought disorder. Anhedonia within the detachment subgroup was defined as "Inability to experience pleasure from activities usually found to be enjoyable." Within the internalizing subgroup, anhedonia was defined as " . . . diminished interest in normally enjoyable activities; these can include such activities as social interaction, hobbies, and sex. High scorers indicate that they are unable to enjoy things that they previously found pleasurable, and that they have little to look forward to in their lives. They report low levels of positive mood states."

Finally, the thought disorder group conceptualized anhedonia as "General deficits in positive emotions and energy levels. High scorers report difficulties experiencing joy and excitement, show little interest in things, and exhibit lethargy, lassitude, and psychomotor slowness." Clearly, these definitions have much in common, such as the central focus on deficits in experiencing positive emotions; however, there are some distinct elements as well, such as the inclusion of lethargy, lassitude, and psychomotor slowness in the thought disorder definition and lack of optimism in the internalizing definition. We permitted each group to define anhedonia (and other overlapping constructs) in a way that made sense in terms of their overarching spectrum. We proceeded this way to provide the opportunity to evaluate different shades in meaning across subgroups. Our analytic plans ultimately will resolve the overlap through structural analyses across all subgroup constructs in Phase 2 of the project.

Another issue that we debated early in the process was whether to operationalize HiTOP psychopathology constructs as unipolar (i.e., with a single maladaptive pole) or bipolar (with poles reflecting pathologically low and high aspects of a given construct). A full treatment of this issue is beyond the scope of this article, but there are clear divisions among psychopathology researchers regarding this issue, especially in the maladaptive personality trait literature, with some arguing for bipolarity at the level of broad personality domains and narrower facets (e.g., Samuel, 2011), and others arguing that maladaptive traits are largely unipolar and thus should be measured that way (e.g., Williams & Simms, 2018). Notably, this is not a controversial issue in most areas of traditional psychopathology assessment (e.g., depression, anxiety, and psychoticism), with constructs and measures typically keyed in a single direction of pathology. Although there were differences across some subgroups, we decided as a workgroup to focus on conceptualizing constructs in a unipolar way and to allow the structural analyses to identify potential points of bipolarity. If the analyses warrant it, two unipolar scales easily can be combined into a single bipolar scale in Phase 2.

Item Pool Development. Following the development of construct lists and definitions, we began the process of developing initial item pools within each subgroup. Similar to above, overinclusiveness was an important principle to follow at this stage (Clark & Watson, 1995). The item pools should be overinclusive in two ways. First, the pool should be broader and more comprehensive than one's theoretical model of the target construct. And second, the pool should include some items that may ultimately be shown to be tangential or unrelated to the target construct. Overinclusiveness is particularly important later in the scale development process when trying to establish the conceptual and empirical boundaries of the target construct(s). As Clark and Watson (1995) noted, "Subsequent psychometric analyses can identify weak, unrelated items that should be dropped from the emerging scale but are powerless to detect content that should have been included but was not" (p. 311).

Content validity—the extent to which a measure's items are relevant to and representative of the construct as it has been theorized or defined—is a second principle that is important in the item development phase of any measurement project (Haynes et al., 1995). To enhance construct validity, the subgroups were instructed to (a) draft 10 to 15 items, at minimum, for each candidate construct, and (b) make sure their item pools have sufficient content to tap all important components of the construct definitions that they wrote. Items were written by all members of each subgroup and compiled and edited by workgroup chairs. Each subgroup handled this task in slightly different ways; details can be found in the accompanying subgroup-driven articles in this special issue.

Before writing a single item, the workgroup as a whole wrestled with numerous decisions—such as response format, measure timeframe, and whether to focus the measure on symptoms, traits, or a combination—that would have implications for how items would need to be written. Although an essential aspect of any objective test, response formats often are afterthoughts in the scale development world. For the HiTOP measurement project, we elected instead to carefully consider the nature and number of response formats before the items were written. Lore rather than data often guide such decisions, but recent work has suggested that response formats with more than six or seven options fail to provide incremental precision benefits (Norman et al., 2003; Simms et al., 2019). Thus, we debated a range of response formats ranging from two to seven options. Arguments for higher numbers focused on the desire to maximize measurement precision at the item level (and ultimately the scale level), but concerns were raised regarding the cognitive load needed to complete more differentiated response formats. In contrast, arguments for fewer response options focused on the desire for a simple format that could be completed easily and quickly by respondents, but 2- and 3-point scales tend to be impoverished with respect to measurement precision (e.g., Simms et al., 2019).

In addition, although data are quite limited on the topic, we opted for an even number of response options to prevent participants from providing a middle or intermediate response for reasons other than a moderate standing on the assessed characteristic. This is especially possible for balanced, Likert-type scales (e.g., inconsistent evidence of such is provided in Simms et al. [2019]), but we were concerned that this also could be the case for any scale with an odd number of options. Thus, we ultimately settled on four response options as a way to accommodate these different issues.

We also debated the labels for the response format, which can take a number of different forms, including agreement-based (e.g., a traditional Likert-type format ranging from strongly disagree to strongly agree), frequency-based (e.g., never to always), and degree-based (e.g., not at all to a lot). Which format to adopt largely is a function of what information is most useful in light of the constructs being measured, as there are no data in the literature claiming any psychometric advantage for one set of labels over others. Moreover, the choice of format has implications for item writing, since items must be written to fit the selected response format. We ultimately opted for a degree-based response format with four choices, including *not at all*, *a little*, *moderately*, and *a lot*, since this format permitted us to write the widest range of items relevant to psychopathology. Other formats were deemed to be too narrow in that they restricted the kinds of items that could be written. For example, traditional Likert-type agreement-based items are best suited for opinion or personality work in which the underlying dimensions are likely to be normally distributed. Similarly, frequency-based formats are best when behavioral counts are desired.

These particular response labels were selected based on an iterative process of discussion among workgroup members who come from a variety of psychometric perspectives. Ultimately, the rationale for these specific anchor labels included that (a) they include relatively simple and straightforward language, which should improve the readability of the measure; and (b) the labels appear to reflect increasing degrees of severity that are consistent with (but not identical to) other common measures in this literature (e.g., Inventory of Depression and Anxiety Symptoms; Watson et al., 2012). The psychometric equidistance between response option points (i.e., true interval scaling) has not been empirically shown for most response formats (e.g., Spratto, 2018); that is true here as well and should be a topic for future research.

We also gave significant consideration to the timeframe and general instructions for the measure, and items were written with these instructions in mind. Psychopathology measures can vary considerably with respect to timeframe, ranging from in-the-moment to lifetime and everything in between (e.g., past week, past year). The choice of timeframe often depends on the application: (a) in clinical work, it often is helpful to have a shorter timeframe (e.g., past week) so that the measure is sensitive to treatment change, whereas (b) in research settings, longer timeframes (e.g., lifetime) often are desired to capture the more stable and trait-like aspects of psychopathology. For the HiTOP item pool, a 1-year time frame was chosen as an intermediate level to permit us to bridge both perspectives and facilitate data collection. That said, we drafted items and instructions such that different timeframes could be substituted and later normed for different context needs (e.g., momentary, past week, past month, past year, lifetime). Future work is planned to study the impact of different timeframes on the measurement of HiTOP dimensions.

Finally, given the breadth of the task before us—to develop a comprehensive measure of psychopathology features within the HiTOP model—we debated whether to focus our work on traditional signs and symptoms, trait manifestations, or some combination of the two. Kotov et al.'s (2017) model depicts both kinds of content at the lowest levels of the hierarchy (see also Figure 1). Some spectra in the model are most clearly related to traditional signs and symptoms (e.g., the distress and anhedonia associated with depression, in the internalizing spectrum), whereas other spectra are more heavily associated with trait manifestations of pathological personality (e.g., the grandiosity and manipulativeness associated with antagonistic

Table 2. Numbers of Constructs and Items Developed in Phase 1 of the HiTOP Measure Development Project.

Subgroup	Initial # of constructs	Initial # of items[a]	# of constructs to Phase 2	# of items to Phase 2
Internalizing	57	430	39	213
Thought disorder[b]	25	365	19	215
Detachment	15	247	10	80
Somatoform and eating pathology	20	240	13	131
Disinhibited and antagonistic externalizing[c]	64	902	—	—

Note. The internalizing group initially developed 1,110 items but reduced them using expert ratings prior top Phase 1 data collection.
HiTOP = Hierarchical Taxonomy of Psychopathology.
[a]Some subgroups (e.g., somatoform, internalizing) started with more items and hone them rationally prior to starting Phase 1 data collection. [b]There is some minimal overlap in constructs and items between the thought disorder and detachment groups. [c]The externalizing group has not yet entered Phase 2.

externalizing). Thus, we decided to write items to reflect a broad range of psychopathology content, including signs, symptoms, features, and traits. Interestingly, there was discussion of creating two different measures, one for symptom dimensions and another for traits, but we ultimately opted to incorporate both in the same measure to minimize complexity and maximize ease of use in research and applied settings. Taken together, the above considerations led us to adopt the following instruction set for items across all domains and subgroups:

> In this survey, you will be asked to respond to a number of statements about your thoughts, feelings, and behavior. Some of these things are pretty common, whereas others are less common. As you complete the survey, please consider whether there have been significant times during the last 12 months during which the following statements applied to you. Then please select the option that best describes how well each statement described you during that period: 0 = not at all; 1 = a little; 2 = moderately; 3 = a lot.

Subgroups were instructed to develop items to work within this format but were given some latitude—given differences in the broad domains we are measuring—regarding how items were written. Some domains were more amenable to specific cognitive, behavioral, or affective instantiations of a given domain. For example, in the internalizing subgroup, items tended toward being more specific symptoms reflective of the domain (e.g., "My mood was unstable and changed very rapidly." In contrast, the detachment subgroup included a mix of items reflective of personality traits relevant to the domain (e.g., "I am generally very distrustful of others") and specific symptoms or behaviors (e.g., "I find it difficult to be generous or warm-hearted towards others" or "I have no interest in romantic relationships.").[1]

Table 2 includes a summary of the numbers of constructs and items that were developed during this phase of scale development, prior to any data collection. In total, over 2,000 items were written initially to measure nearly 200 constructs across the five spectrum-based subgroups.

Structural Validity

The structural validity component of the scale development process is focused on collection of responses to the initial item pool and statistical procedures designed to hone those items into homogeneous and differentiable scales (Clark & Watson 2019; Loevinger, 1957; Simms & Watson, 2007). In this section, we describe the methods that the Measures Development Workgroup has developed to enhance the structural validity of the resultant measure. Loevinger (1957) defined the structural component of construct validity as "the extent to which structural relations between test items parallel the structural relations of other manifestations of the trait being measured" (p. 661).

In the context of psychopathology scale development, this definition suggests that the structural relations among test and nontest manifestations of the target construct should be parallel to the extent possible—what Loevinger (1957) called "structural fidelity"—and ideally, this structure should match that of the theoretical model underlying the construct. Full treatment of structural fidelity is beyond the scope of this article. However, one important implication of this principle is that theory development and scale development are complementary processes that are mutually informative. That is, theory (i.e., a priori construct lists and definitions) informs item development, and structural analyses of responses to those items have the potential to inform the theory of the constructs under investigation. In the end, scale development is an iterative process in which items and constructs go through multiple rounds of item development/revision and data collection/analyses before arriving at a final model of the constructs and a final measure of those constructs.

Phase 1 Statistical Philosophy and Methods. Where the rubber hits the road of new measurement development is in the specific methods and procedures adopted. A unique aspect of this project is that many experts in measurement development were enlisted, without compensation, to collectively build what seeks to be a (reasonably) comprehensive measure of psychopathology. This is no small task, and this is no small cast. From the outset it was clear that as a practical matter, early iterations of the measure development process would suffer by highly prescriptive and centrally defined procedures. That is, we felt that given the expertise in each of the designated teams, a heavy-handed approach to workflow might stymie the important work of content generation and preliminary validation, and thus we opted instead to provide loose guidelines that each team could easily follow while also feeling unencumbered to pursue their own established pipelines of work. This broad and inclusive approach to preliminary item and scale development was safeguarded by the fact that Phase 2 of the project will centralize the data analytic procedures in a fashion that is consistent with contemporary best-practices in scale development methodology.

For Phase 1, the most salient process-based challenges were developing a set of procedures that would ensure commonalities across contributors, while allowing each to flexibly marshal their considerable knowledge of scale design and development. What follows is a set of procedures intended to do just that: Ensure implementation commonalities with enough flexibility to accommodate the unique challenges of each domain of measurement. The principal analytic goal of the first phase was to develop a refined set of constructs and preliminary scales to take forward into a second phase of data collection. Narrower aims under this overarching goal included removing redundant items and identifying scales in need of additional items.

Each team was asked to follow a set of general principles—which were developed collaboratively with all subgroup chairs—in pursuing the analytic steps we describe below. These general principles included addressing each analytic step, unless designated as optional, documenting all decisions for archival purposes, erring on the side of inclusion when deciding on whether to retain or delete items, and retaining deleted items for possible future use if needed. In addition, we proposed that each team keep an "item purgatory" for items that fail these procedures but nonetheless might be useful for re-inclusion at a later stage of scale development (e.g., if a viable scale requires additional items for sufficient reliability).

Phase 1 analyses were designed to follow four steps. First, because item-level factor analyses become exceedingly difficult as the numbers of items and constructs increase, we initially sought to identify smaller groups of constructs on which to conduct item-level analyses.[2] To do this, teams were asked to identify a preliminary domain structure by scoring and factor analyzing their a priori scales (i.e., the items they rationally developed to measure each candidate construct; cf. homogeneous item composites [HICs], Hogan, 1983). We then focused scale development efforts within the factors that emerged from this process. The details of this step (and all analytic steps) appear in each of the subgroup papers that accompany this introductory article. For example, the internalizing subgroup (Watson et al., this issue) started with 52 a priori HICs measured by 395 items, which is too unwieldy to analyze in a single factor analysis. Their factor analyses of these a priori HICs ultimately revealed five factors—labeled Core Distress, Panic/PTSD, Fear, Social/Somatic Anxiety, and Mania—which were more manageably sized to facilitate preliminary scale development efforts (see the second step, next).[3]

Teams were permitted substantial flexibility in modeling decisions, consistent with the principles above, so they were encouraged to use an oblique rotation (e.g., geomin, promax), but which rotation was not specified (Fabrigar et al., 1999). Parallel analysis was specified to guide the maximum number of factors to extract, and the solution with the most interpretable factors that has an eigenvalue larger than randomly generated data was to be retained (Horn, 1965; Velicer et al., 2000). Any reasonable statistical software was permissible (e.g., SAS, SPSS, STATA, M*plus*, and R), but teams were asked to identify and use appropriate procedures for factor analyzing skewed data (e.g., asymptotical distribution free, robust maximum likelihood methods). A priori scales with loadings of at least .50 were assigned to a factor. Orphan scales that failed to load on a factor were retained for subsequent analyses.

For the second analytic step, teams were asked to conduct item-level factor analyses within each resultant factor to arrive at preliminary sets of scales. Teams were asked to use appropriate techniques for item-level ordinal data (e.g., analyses based on polychoric correlation matrices, robust weighted least squares estimation; Holgado-Tello et al., 2010; Jöreskog & Moustaki, 2001). We asked teams to use oblique rotation or alternatives documented with rationale. Similar to above, teams were asked to use parallel analysis and to retain the maximum number of factors that were interpretable and had eigenvalues larger than parallel random data. We set the loading threshold for retention of an item on a factor to .40 or higher. If there were sizeable secondary loadings, we asked for retained items to evidence a .20 difference between primary and secondary loadings to ensure discriminability. Items that failed these thresholds were retained in item purgatory. For example, the internalizing subgroup analyses that were described earlier yielded a provisional set of 35 scales (Watson et al., this issue).

Some teams, but not all, indicated they felt that an additional step was needed to ensure adequate discriminant validity at this stage. Therefore, an optional third step was to compute item-level correlations with other scales to ensure discriminability, and remove items that correlate too highly with theoretically unrelated constructs. Items that failed this were removed to item purgatory. Groups handled this step in a variety of ways. For example, the internalizing group (Watson et al., this issue) examined item-level discriminant validity against their own set of preliminary internalizing scales (i.e., removing items from one scale that correlated too highly with other related scales). In contrast, the detachment subgroup (Zimmermann et al., this issue) evaluated discriminant validity against the domain scales of the Personality Inventory for *DSM-5*–short form (Maples et al., 2015) and Big Five Inventory–2 (Soto & John, 2017).

Fourth, teams were asked to refine preliminary scales and identify scales in need of new items, with a goal of 8 to 10 items for each preliminary scale to carry into Phase 2 data collection.[4] If a scale started this step with more than 10 items, teams were asked to consult McDonald's Omega and/or IRT information curves to identify items to drop, iterating until either Omega dropped below .85 or a set of 10 items was identified (Revelle & Condon, 2019). In contrast, if either there were fewer than 8 to 10 items, Omega was less than .85, or the IRT analyses identified severity gaps, we asked teams to write new items or resurrect previously tested items from purgatory. As with the above, any reasonable software could be used in the estimation of IRT parameters or the calculation of Omega. Based on the item content represented in each scale, we asked teams to label all provisional scales.

Table 2 includes a summary of the numbers of constructs and items that have emerged thus far from Phase 1 data collection following the statistical steps described above. Excluding externalizing, 648 items tapping 81 constructs were identified in Phase 1 of the project.

Next Steps

As noted above, four of five subgroups have completed Phase 1 and have entered Phase 2, which is focused on finalization of the HiTOP scales. More specifically, Phase 2 will be focused on collection of cross-validation data for all preliminary scales together in a large single study, across all subgroups, which will facilitate adjudication of redundancy across domains, studies of the scales' joint structure, collecting representative norms, and examination of moderators of structure, such as gender/sex and ethnicity/race. The data analytic plans for Phase 2 currently are being drafted collaboratively with all subgroup leaders, but several aspects of this plan are clear at this point. First, in contrast to the Phase 1 preliminary scale development analyses described in the articles of this special issue—which were conducted according to a general plan but independently in each subgroup—Phase 2 scale analyses will be centralized and completed by a smaller group of data analysts from within the Measure Development Workgroup. This will ensure that identical procedures are used to finalize all scales. Second, at multiple steps in the scale finalization process, the central data analysts will report interim results to the subgroups so that they may provide conceptual input to the process.

Third, the primary methods will remain within the factor analysis and item response theory toolkits, and all analytic code will be saved and made available for external review. Finally, we will use structural methods (e.g., exploratory structural equation modeling) to study the joint structure of all HiTOP scales and to build broad scales reflecting each HiTOP spectrum. When Phase 2 is completed—slated to be by late 2021 or early 2022—we plan to release the final scales for research use.

As noted earlier, Phase 3 will focus on external validation of the HiTOP questionnaire against relevant test and nontest psychopathology criteria (e.g., prominent measures relevant to each domain, behavioral tasks, ecological momentary assessment, etc.) and on building the HiTOP measure into an open-source product with features designed to improve its clinical utility. A common problem in the structural psychopathology literature is that measures are published in journals with researchers as the primary audience. Such measures may achieve some modicum of success in the research community, but rarely do they see much use in applied settings. Thus, we intend to bridge the disconnect between research and clinical applications of psychopathology measures in a number of ways. First, based on data collected in Phases 2 and 3, we aim to calculate and publish norms representative of all populations within which the measure is designed to be used (e.g., community and psychiatric norms). Second, given the high-stakes nature of some clinical settings (e.g., where there is some motivation to dissemble or malinger for external gain), we plan to integrate validity scales designed to detect a range of problematic responding, including inconsistent responding, underreporting, and overreporting. Third, we will team with the Clinical Translation Workgroup to develop a formal scoring and interpretative manual—as well as training workshops—to aid clinicians in the use and interpretation of the HiTOP measures.

Fourth, we plan to develop short-forms of the measure geared toward different settings, as well as to offer modularization that will permit users to administer only the scales that are desired in a given setting. Fifth, once the constructs and scales are finalized, we plan to develop a companion interview that will permit users to assess the same HiTOP constructs in a manner that permits follow-up questions and clinical judgments regarding the severity of a given psychopathology profile. The interview has not yet been designed, but we anticipate structuring the interview so that all elements of the final self-report measure are assessed with respect to both their presence and severity. Finally, the HiTOP measures will be open-source, free to use, and available in both computerized and paper-and-pencil formats, which provides the flexibility to make the measure practically useful across a variety of clinical settings.

Measurement Informing Revisions of the HiTOP Model

The scales that are developed through this process will be put to good use for many different aims. One that is relatively unique to the HiTOP consortium is using the resulting scales to reinform the model they are designed to measure. Indeed, HiTOP is a living model not yet finalized, and may never be finalized. It draws its strength from adopting a principled approach to identifying the structure of psychopathology. In this sense, it is important to recognize that this scale development process is wholly empirical, and is not designed to fit an extant model, although it is informed by the research that produced the model. By the same token, the model will be informed by this empirical scale development process, although not entirely so, because other sources of data will be important moving forward (e.g., clinical correlates or course). Thus, the model and the scales are by definition related but also separate.

However, it is likely that these scales will play an important role in informing the model and providing much needed information for many sections of the model that currently are underinformed by the extant data. Much of this is a function of the fact that the majority of structural literature informing HiTOP is based on secondary analyses of data that suffer from several limitations leaving them only adequately, but not ideally, suited for the task. Issues such as baked-in structure associated with interviews pre-organized around diagnoses, skip-out rules, spotty or only partial coverage of the psychopathology universe of content, and others contribute to a model that has well-established outlines, but needs much coloring in between the lines, particularly at the low-order levels. This fully bottom-up process of developing scales provides the opportunity to flesh out the model by developing thorough item sets that can be administered in studies designed to provide full coverage.

Consistent with the perspective that HiTOP represents a living model, the Revisions Workgroup recently was formed to develop and implement procedures for making changes or additions to the model. Led by Drs. Miriam Forbes and Aidan Wright, the goal of the Revisions Workgroup is to identify consensual processes that can be used to evaluate proposals for changes and make recommendations for changes or additions based on the strength of the evidence. The Revisions Workgroup is not designed to itself identify and work on making changes to the model. The criteria used to evaluate proposals are loosely based on the Grading of Recommendations Assessment, Development, and Evaluation (Guyatt et al., 2008) system, which is widely used in medicine for rating the quality of evidence for clinical practice recommendations. The focus is on making criteria explicit and clear to ensure transparency and reliability of ratings within and between proposals.

Full detail of the criteria for evaluation extend beyond the scope of this review, but it is important to highlight several key aspects of the process. First, those proposing a change will be asked to prepare a detailed review of the available evidence and its strength. Second, this will include considerations of breadth of measurement, but also sampling, and external validation criteria. As such, the current scales, by virtue of their breadth and detailed item-level assessment of psychopathology, will have an important role in informing these proposals to the Revisions Workgroup. At the same time, the new scales are unlikely to provide sufficient information to address all lingering thorny questions (e.g., location of mania and obsessions), at least not initially. In part, this is because they will need to be considered alongside other relevant existing data, and will need to be evaluated on the same merits and demerits. For instance, considerations of sampling and external validation criteria will be important, but likely to accrue slowly. Thus, we anticipate these scales will provide invaluable data to inform the model, but will necessarily not be the only source considered.

Limitations

One limitation of our measurement work thus far is that we have adopted a relatively simple measurement model in which each HiTOP spectrum is measured by lower order facets that all exist at the same level of specificity or generality. This approach is consistent with most established hierarchical personality and psychopathology structures but very likely is wrong (e.g., Condon et al., 2020). It is likely that psychopathology facets exist at more than the two levels we have focused on in this work (i.e., spectra and their facets). Indeed, HiTOP itself includes multiple hierarchical levels that vary in their specificity or generality. Moreover, even the spectra themselves likely vary in size and breadth and, perhaps, importance. Despite these likelihoods, we asked the subgroups to exhaustively search their literatures for low-order manifestations of their spectra without regard to the size of those manifestations. It will be a task for future research in the workgroup to explicate the size and breadth of the spectra themselves and their low-order manifestations.

Another important limitation is that much of the Phase 1 work is based on nonclinical samples, which raises the risk that some potentially useful but severe items might have been jettisoned too early simply because too few participants endorsed them. This issue is an important one. However, our analytic procedures made clear that promising items that reflect higher severity should not be removed even if they demonstrated poorer than expected psychometric characteristics at this stage of development. Said differently, subgroups were instructed to err on the side of inclusion at this stage. That said, in Phase 2 of the project we intend to recruit a broader range of samples (e.g., clinical, forensic) that include a sufficient severity and range of psychopathology.

Conclusions

HiTOP represents a promising attempt to replace an antiquated and problematic psychiatric classification system with one rooted in modern scientific methods. In particular, given the mass of literature that has accumulated over the past 25 years showing the merits of a quantitatively driven dimensional classification, HiTOP is well-positioned to change the way researchers study mental illness as well as the way it is assessed and treated in clinical settings. However, this new system faces significant challenges given the strong inertia that typically drives practices in clinical psychology and psychiatry. Building strong measures of HiTOP dimensions is one step the consortium is taking to facilitate the uptake of HiTOP into research and practice.

To that end, in this article we introduce the rationale for and methods of the development of HiTOP-specific measures of psychopathology. A large number of measurement experts, nested within five spectrum-based groups, have collaborated to build and implement a principled set of modern scale development methods. This article summarizes the principles of scale development, whereas the other article in this special issue document the specific progress that has been made with respect to the five spectrum-based subgroups. The measures are not yet complete, but we are at a place where we wish to document the significant progress that has been made. Moreover, we view this collaborative effort as a model for how to marshal the resources of many experts in the service of building something that is bigger than any one of us. The 40+ members of the Measures Development Workgroup have worked for years, meeting virtually and in person, to get to this point. In one sense, this exercise has been a little like the process of herding cats (very smart and opinionated cats!) and getting them all to do the same thing. However, the diversity of opinions and perspectives has had the net effect of making a better product that represents the collective wisdom of many scholars who have devoted their careers to scale development and psychiatric classification research.

We have made significant progress thus far, and functional measures should be available soon for research purposes, with clinically useful versions to follow shortly thereafter. Part of the rationale for publishing this and the accompanying articles now, in the middle of the process, is a desire to proceed in the spirit of the open science movement. To that end, the data and methods that we accumulate through this process will be made fully available to others via an open-science platform when the measure is released for research purposes. The present set of articles represents our attempt to document the work that has been done to this point.

Acknowledgments

We wish to thank the following individuals who have contributed to the Measures Development Workgroup with their time and resources: Anna Docherty, Ashley Watts, Avshalom Caspi, Camilo Ruggero, Chris Conway, Craig Rodriguez-Seijas, Donald Lynam, Doug Samuel, Elizabeth Martin, Giorgia Michelini, Holly Levin-Aspenson, J. D. Haltigan, Jennifer Tackett, Jonathan Berzon, Josh Miller, Kasey Stanton, Katherine Jonas, Katherine Keyes, Kelsie Forbush, Kristian Markon, Kristin Naragon-Gainey, Lee Anna Clark, Les Morey, Marina Bornovalova, Matt Sunderland, Michael Chmielewski, Michael Witthöft, Mike Bagby, Miri Forbes, Molly Nikolas, Natacha Carragher, Noah Venables, Praveetha Patalay, Rachael Graziolplene, Robert Krueger, Sara Gould, Shereen Khoo, Ulrich Reininghaus, and Yuliya Kotelnikova.

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: Preparation of this article was supported by a research grant to L. J. Simms and Roman Kotov from the National Institute of Mental Health (No. R56MH122537).

ORCID iDs

Leonard J. Simms https://orcid.org/0000-0001-8081-380X
David Cicero https://orcid.org/0000-0002-5666-9139
Martin Sellbom https://orcid.org/0000-0002-2883-6163
David Watson https://orcid.org/0000-0002-9605-0576
Johannes Zimmermann https://orcid.org/0000-0001-6975-2356

Notes

1. Another way the groups differed is the verb tense of items. Most items were written in the past tense, whereas the detachment items were written in the present tense. At this stage, this was not a concern since we were conducting analyses only within domains. Prior to Phase 2 data collection, which came after the analyses described in this collection of articles, all items were harmonized to be in the past tense.
2. An alternative first step was considered in which we would first factor analyze the items within each a priori HIC to look for items that do not cohere as expected. We opted against this approach because we wanted to leave open the possibility that items that do not work well for their intended scales may nonetheless be useful for other preliminary scales.
3. These analyses did not include the 35 sexual dysfunction items, which were examined separately. The sexual dysfunction analyses produced four scales with a total of 15 items.
4. For preliminary scales requiring additional items, we asked teams to aim for scales of at least 12 items to leave room for cross-validation shrinkage in Phase 2.

References

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

Aslinger, E. N., Manuck, S. B., Pilkonis, P. A., Simms, L. J., & Wright, A. G. C. (2018). Narcissist or narcissistic? Evaluation of the latent structure of narcissistic personality disorder. *Journal of Abnormal Psychology*, *127*(5), 496-502. https://doi.org/10.1037/abn0000363

Carragher, N., Krueger, R. F., Eaton, N. R., Markon, K. E., Keyes, K. M., Blanco, C., Saha, T. D., & Hasin, D. S. (2014). ADHD and the externalizing spectrum: direct comparison of categorical, continuous, and hybrid models of liability in a nationally representative sample. *Social Psychiatry and Psychiatric Epidemiology*, *49*(8), 1307-1317. https://doi.org/10.1007/s00127-013-0770-3

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p Factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*(2), 119-137. https://doi.org/10.1177/2167702613497473

Chmielewski, M., Clark, L. A., Bagby, R. M., & Watson, D. (2015). Method matters: Understanding diagnostic reliability in *DSM-IV* and *DSM-5*. *Journal of Abnormal Psychology*, *124*(3), 764-769. https://doi.org/10.1037/abn0000069

Clark, L. A., Cuthbert, B., Lewis-Fernández, R., Narrow, W. E., & Reed, G. M. (2017). Three approaches to understanding and classifying mental disorder: ICD-11, *DSM-5*, and the National Institute of Mental Health's Research Domain Criteria (RDoC). *Psychological Science in the Public Interest*, *18*(2), 72-145. https://doi.org/10.1177/1529100617727266

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309-319. https://doi.org/10.1037/1040-3590.7.3.309

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412-1427. https://doi.org/10.1037/pas0000626

Condon, D. M., Wood, D., Mõttus, R., Booth, T., Costantini, G., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A. G. C., Ziegler, M., & Zimmermann, J. (2020). Bottom up construction of a personality taxonomy. *European Journal of Psychological Assessment*, *36*(6), 923-934. https://doi.org/10.1027/1015-5759/a000626

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302. https://doi.org/10.1037/h0040957

Fabrigar, L., Wegener, D., MacCallum, R., & Strahan, E. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272-299. https://doi.org/10.1037/1082-989X.4.3.272

Forbes, M. K., Sunderland, M., Rapee, R. M., Batterham, P. J., Calear, A. L., Carragher, N., Ruggero, C., Zimmerman, M., Baillie, A. J., Lynch, S. J., Mewton, L., Slade, T., & Krueger, R. F. (2021). A detailed hierarchical model of psychopathology: From individual symptoms up to the general factor of psychopathology. *Clinical Psychological Science*, *9*(2), 139-168. https://doi.org/10.1177/2167702620954799

Galatzer-Levy, I. R., & Bryant, R. A. (2013). 636,120 ways to have posttraumatic stress disorder. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *8*(6), 651–662. https://doi.org/10.1177/1745691613504115

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*(1), 26-34. https://doi.org/10.1037/0003-066X.48.1.26

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650), 924-926. https://doi.org/10.1136/bmj.39489.470347.AD

Harkness, A. R., Finn, J. A., McNulty, J. L., & Shields, S. M. (2012). The Personality Psychopathology-Five (PSY-5): Recent constructive replication and assessment literature review. *Psychological Assessment*, *24*(2), 432-443. https://doi.org/10.1037/a0025830

Harkness, A. R., McNulty, J. L., & Ben-Porath, Y. S. (1995). The Personality Psychopathology–Five (PSY-5): Constructs and MMPI-2 scales. *Psychological Assessment*, *7*(1), 104-114. https://doi.org/10.1037/1040-3590.7.1.104

Haslam, N., McGrath, M. J., Viechtbauer, W., & Kuppens, P. (2020). Dimensions over categories: A meta-analysis of taxometric research. *Psychological Medicine*, *50*(9), 1418-1432. https://doi.org/10.1017/S003329172000183X

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238-247. https://doi.org/10.1037/1040-3590.7.3.238

Hogan, R. T. (1983). A socioanalytic theory of personality. In M. Page (Ed.), *1982 Nebraska symposium on motivation* (pp. 55-89). University of Nebraska Press.

Holgado-Tello, F. P., Chacón–Moscoso, S., Barbero–García, I., & Vila-Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, *1*(44), 153-166. https://doi.org/10.1007/s11135-008-9190-y

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179-185. https://doi.org/10.1007/BF02289447

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*(3), 347-387. https://doi.org/10.1093/schbul/sbq024

Kotov, R., Chang, S. W., Fochtmann, L. J., Mojtabai, R., Carlson, G. A., Sedler, M. J., & Bromet, E. J. (2011). Schizophrenia in the internalizing-externalizing framework: A third dimension? *Schizophrenia Bulletin*, *37*(6), 1168-1178. https://doi.org/10.1037/abn0000258

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., Clark, L. A., Eaton, N. R., Forbes, M. K., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E., Ivanova, M. Y., Lynam, D. R., Markon, K., . . . Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, *4*(4), 454-477. https://doi.org/10.1037/abn0000258

Kotov, R., Krueger, R. F., Watson, D., Cicero, D. C., Conway, C. C., DeYoung, C. G., Eaton, N. R., Forbes, M. K., Hallquist, M. N., Latzman, R. D., Mullins-Sweatt, S. N., Ruggero, C. J., Simms, L. J., Waldman, I. D., Waszczuk, M. A., & Wright, A. G. C. (2021). The Hierarchical Taxonomy of Psychopathology (HiTOP): A quantitative nosology based on consensus of evidence. *Annual Review of Clinical Psychology*, *17*. Advance online publication. https://doi.org/10.1146/annurev-clinpsy-081219-093304

Krueger, R. F. (1999). The structure of common mental disorders. *Archives of General Psychiatry*, *56*(10), 921-926. https://doi.org/10.1001/archpsyc.56.10.921

Lahey, B. B., Applegate, B., Hakes, J. K., Zald, D. H., Hariri, A. R., & Rathouz, P. J. (2012). Is there a general factor of prevalent psychopathology during adulthood? *Journal of Abnormal Psychology*, *121*(4), 971-977. https://doi.org/10.1037/a0028355

Levin-Aspenson, H. F., Watson, D., Clark, L. A., & Zimmerman, M. (2020). What is the general factor of psychopathology? Consistency of the p Factor across samples. *Assessment*. Advance online publication. https://doi.org/10.1177/1073191120954921

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635-694. https://doi.org/10.2466/pr0.1957.3.3.635

Maples, J. L., Carter, N. T., Few, L. R., Crego, C., Gore, W. L., Samuel, D. B., Williamson, R. L., Lynam, D. R., Widiger, T. A., Markon, K. E., Krueger, R. F., & Miller, J. D. (2015). Testing whether the DSM-5 personality disorder trait model can be measured with a reduced set of items: An item response theory investigation of the Personality Inventory for DSM-5. *Psychological Assessment*, *27*(4), 1195–1210. https://doi.org/10.1037/pas0000120

Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological Bulletin*, *137*(5), 856-879. https://doi.org/10.1037/a0023678

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist*, *52*(5), 509-516. https://doi.org/10.1037/0003-066X.52.5.509

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, *41*(5), 582-592. https://doi.org/10.1097/01.MLR.0000062554.74615.4C

Regier, D. A., Kuhl, E. A., & Kupfer, D. J. (2013). The *DSM-5*: Classification and criteria changes. *World Psychiatry*, *12*(2), 92-98. https://doi.org/10.1002/wps.20050

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395-1411. https://doi.org/10.1037/pas0000754

Ruggero, C. J., Kotov, R., Hopwood, C. J., First, M., Clark, L. A., Skodol, A. E., Mullins-Sweatt, S. N., Patrick, C. J., Bach, B., Cicero, D. C., Docherty, A., Simms, L. J., Bagby, R. M., Krueger, R. F., Callahan, J. L., Chmielewski, M., Conway, C. C., De Clercq, B., Dornbach-Bender, A., . . . Zimmermann, J. (2019). Integrating the Hierarchical Taxonomy of Psychopathology (HiTOP) into clinical practice. *Journal of Consulting and Clinical Psychology*, *87*(12), 1069-1084. https://doi.org/10.1037/ccp0000452

Samuel, D. B. (2011). Assessing personality in *DSM-5*: The utility of bipolar constructs. *Journal of Personality Assessment*, *93*(4), 390-397. https://doi.org/10.1080/00223891.2011.577476

Sellbom, M. (2019). The MMPI-2-Restructured Form (MMPI-2-RF): Assessment of personality and psychopathology in the twenty-first century. *Annual Review of Clinical Psychology*, *15*, 149-177. https://doi.org/10.1146/annurev-clinpsy-050718-095701

Simms, L. J., & Watson, D. (2007). The construct validation approach to personality scale construction. In R. Robins, C. Fraley, & R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 240-258). Guilford Press.

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*(4), 557-566. https://doi.org/10.1037/pas0000648

Smith, G. T., Atkinson, E. A., Davis, H. A., Riley, E. N., & Oltmanns, J. R. (2020). The general factor of psychopathology. *Annual Review of Clinical Psychology*, *16*, 75-98. https://doi.org/10.1146/annurev-clinpsy-071119-115848

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1), 117-143. https://doi.org/10.1037/pspp0000096

Spratto, E. M. (2018). *In search of equality: Developing an equal interval Likert response scale*. https://commons.lib.jmu.edu/diss201019/172

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

Watson, D., Levin-Aspenson, H. F., Waszczuk, M. A., Conway, C. C., Dalgleish, T., Dretsch, M. N., . . . HiTOP Utility Workgroup. (in press). Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): III. Emotional dysfunction superspectrum. *World Psychiatry*.

Watson, D., O'Hara, M. W., Naragon-Gainey, K., Koffel, E., Chmielewski, M., Kotov, R., Stasik, S. M., & Ruggero, C. J. (2012). Development and validation of new anxiety and bipolar symptom scales for an expanded version of the IDAS (the IDAS-II). *Assessment*, *19*(4), 399-420. https://doi.org/10.1177/1073191112449857

Williams, T. F., & Simms, L. J. (2018). Personality traits and maladaptivity: Unipolarity vs. bipolarity. *Journal of Personality*, *86*(5), 888-901. https://doi.org/10.1111/jopy.12363

Williams, T. F., & Simms, L. J. (2020). The conceptual foundations of descriptive psychopathology. In A. Wright, & M. Hallquist (Eds.), *The Cambridge handbook of research methods in clinical psychology* (pp. 33-44). Cambridge University Press. https://doi.org/10.1017/9781316995808.006

Wright, A. G., & Simms, L. J. (2015). A metastructural model of mental disorders and pathological personality traits. *Psychological Medicine*, *45*(11), 2309-2319. https://doi.org/10.1017/S0033291715000252

Wright, A. G. C., Hallquist, M. N., Morse, J. Q., Scott, L. N., Stepp, S. D., Nolf, K. A., & Pilkonis, P. A. (2013). Clarifying interpersonal heterogeneity in borderline personality disorder using latent mixture modeling. *Journal of Personality Disorders*, *27*(2), 125-143. https://doi.org/10.1521/pedi.2013.27.2.125

World Health Organization. (2020). *International statistical classification of diseases and related health problems* (11th ed.). https://icd.who.int/