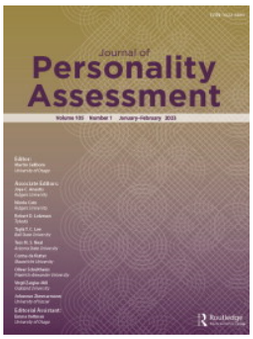

Routledge
Taylor & Francis Group

Journal of Personality Assessment

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hjpa20

Focusing Narrowly on Model Fit in Factor Analysis Can Mask Construct Heterogeneity and Model Misspecification: Applied Demonstrations across Sample and Assessment Types

Kasey Stanton, Ashley L. Watts, Holly F. Levin-Aspenson, Ryan W. Carpenter, Noah N. Emery & Mark Zimmerman

To cite this article: Kasey Stanton, Ashley L. Watts, Holly F. Levin-Aspenson, Ryan W. Carpenter, Noah N. Emery & Mark Zimmerman (2023) Focusing Narrowly on Model Fit in Factor Analysis Can Mask Construct Heterogeneity and Model Misspecification: Applied Demonstrations across Sample and Assessment Types, Journal of Personality Assessment, 105:1, 1-13, DOI: 10.1080/00223891.2022.2047060

To link to this article: https://doi.org/10.1080/00223891.2022.2047060

JOURNAL OF PERSONALITY ASSESSMENT
2023, VOL. 105, NO. 1, 1–13
https://doi.org/10.1080/00223891.2022.2047060

Routledge
Taylor & Francis Group

STATISTICAL DEVELOPMENTS AND APPLICATIONS

Focusing Narrowly on Model Fit in Factor Analysis Can Mask Construct Heterogeneity and Model Misspecification: Applied Demonstrations across Sample and Assessment Types

Kasey Stanton[1], Ashley L. Watts[2], Holly F. Levin-Aspenson[3], Ryan W. Carpenter[4], Noah N. Emery[5], and Mark Zimmerman[6]

[1]Department of Psychology, University of Wyoming; [2]Department of Psychological Sciences, University of Missouri; [3]Department of Psychiatry and Human Behavior, Brown University; [4]Department of Psychological Sciences, University of Missouri-St. Louis; [5]Department of Psychology, Colorado State University; [6]Rhode Island Hospital

ABSTRACT

This study builds upon research indicating that focusing narrowly on model fit when evaluating factor analytic models can lead to problematic inferences regarding the nature of item sets, as well as how models should be applied to inform measure development and validation. To advance research in this area, we present concrete examples relevant to researchers in clinical, personality, and related subfields highlighting two specific scenarios when an overreliance on model fit may be problematic. Specifically, we present data analytic examples showing that focusing narrowly on model fit may lead to (a) incorrect conclusions that heterogeneous item sets reflect narrower homogeneous constructs and (b) the retention of potentially problematic items when developing assessment measures. We use both interview data from adult outpatients ($N = 2{,}149$) and self-report data from adults recruited online ($N = 547$) to demonstrate the importance of these issues across sample types and assessment methods. Following demonstrations with these data, we make recommendations focusing on how other model characteristics (e.g., factor loading patterns; carefully considering the content and nature of factor indicators) should be considered in addition to information provided by model fit indices when evaluating factor analytic models.

ARTICLE HISTORY
Received 5 August 2021
Accepted 9 February 2022

Exploratory factor analytic (EFA), confirmatory factor analytic (CFA), and "hybrid" exploratory structural equation modeling (ESEM) approaches feature prominently in measure development efforts and examinations of personality and psychopathology structure (Greiff & Heene, 2017; Sellbom & Tellegen, 2019; Wright, 2017). Many crucial decision points arise when using factor analysis, including how many latent factors to extract in analyses and how to best specify item loadings a priori with use of CFA, or in some cases, ESEM models (Wright, 2017). When using CFA, ESEM, and, to a lesser extent, EFA, researchers also examine model fit indices to guide model selection. These indices typically are compared against "benchmarks" presumed to reflect an "acceptable" or "well-fitting" model (e.g., interpreting fit based on root mean square error of approximation [RMSEA] values; Hopwood & Donnellan, 2010).

Information derived from model fit indices can be useful for guiding model interpretation and selection by providing a general sense of the degree to which models align with observed data. That being said, concerns have been raised about an overreliance on model fit as an indicator of model validity (Barrett, 2007; Gignac, 2007; Sellbom & Tellegen, 2019). Sellbom and Tellegen (2019) describe that it is often the case that "limited theoretical consideration […] goes into decision making when selecting an 'optimal' model" when using factor analysis, such that "researchers seem to allow themselves to be dictated by model fit indices" (p. 1431). Various simulation efforts also have demonstrated limitations of fit indices for identifying the true ("population") structural model, as model fit may be stronger for a model other than the true or known structure (e.g., Bonifay & Cai, 2017; Greene et al., 2019).

This research indicates the need to consider additional model characteristics to determine model validity and viability, as ignoring other model characteristics (e.g., factor loadings, external factor associations) can result in a focus on problematic models that serve as guiding frameworks for literatures (Roberts & Pashler, 2000; Watts et al., 2019). For example, measure validation efforts for some measures of social-cognitive vulnerabilities (e.g., intolerance of uncertainty; Carleton et al., 2007) have focused heavily on identifying well-fitting models to guide measure development, with less attention given to other aspects of the measure development process. As a result of focusing on identifying

CONTACT Kasey Stanton kaseyjstanton@gmail.com Department of Psychology, University of Wyoming, 1000 E. University Avenue, Laramie, WY 82071.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/00223891.2022.2047060

well-fitting models without thoroughly considering key issues such as discriminant validity, some social-cognitive vulnerability measures have later been shown to include item content that is difficult to differentiate from general distress (Naragon-Gainey & Watson, 2018; Stanton, 2020).

Herein, we focus on two problems that can arise from an overreliance on model fit indices when identifying models. First, researchers may fail to consider valid, alternative, multifactor solutions when single-factor models fit well. Second, even when multifactor solutions are considered, model fit indices may indicate acceptable to good fit even when item loadings on specific factors are misassigned. Previous theoretical articles provide general recommendations regarding the need to consider model characteristics other than fit as reviewed (e.g., Greiff & Heene, 2017; Roberts & Pashler, 2000; Sellbom & Tellegen, 2019), and other studies offer insights into interpretations of model fit in other specific data analytic scenarios (e.g., when evaluating higher-order factor models; Gignac, 2007). However, the two specific issues of focus here have received relatively little attention in psychometric research to date, and we provide practical demonstrations relevant to clinical, personality, and other subfields to facilitate researchers' awareness of these issues when applying factor analysis in their own work.

Issues with failing to recognize heterogeneity in item sets

Researchers may mistakenly interpret heterogeneous item sets as being homogeneous in nature if model fit is "good" or "acceptable" for single-factor models (Watts et al., 2021). Similarly, researchers sometimes conflate "good" internal consistency (e.g., as determined using coefficient alpha) with scale homogeneity (Dunn et al., 2014). As a result, an overreliance on these indices or their misinterpretation can result in the conclusion that an item set reflects a single, narrow construct even when it does not (Chmielewski et al., 2011). In such cases, heterogeneous item sets may then be summed to create global composite scores, even though more homogeneous item sets within broader composites may associate differentially with external criteria (Jackson et al., 1976; Smith & McCarthy, 1995; Smith et al., 2009).

For example, different psychopathy dimensions (e.g., disinhibition, callousness) show distinct neural correlates that may be obscured when focusing analyses solely on global psychopathy scores (Latzman et al., 2020). Similarly, the borderline personality disorder (PD) criteria from the *Diagnostic and Statistical Manual of Mental Disorders* (currently fifth edition; *DSM–5;* American Psychiatric Association, 2013) are heterogenous in nature, and ratings of individual borderline PD criteria show distinctive associations with other variables (e.g., inappropriate anger shows some specificity with aggression; Chmielewski et al., 2011; Sharp et al., 2015). However, item sets used to assess borderline PD criteria sometimes have been interpreted as being homogeneous because single-factor CFA models of borderline PD ratings show good fit, with fit for single-factor models appearing better than multidimensional model configurations in some studies (e.g., Clifton & Pilkonis, 2007; Feske et al., 2007; Johansen et al., 2004). Although illustrative, these examples are not limited to the personality pathology literature. In fact, Greiff and Heene (2017) noted that these issues persist across substantive research areas (e.g., substance use assessment; see Watts et al., 2021), as they describe that when single-factor CFA structures fit well, one might "conclude, probably just as 99% of other researchers working in assessment would, that you found support for the unidimensional structure" (p. 313) without carefully considering other model characteristics.

Issues with failing to recognize misassigned or problematic items

As a second related issue, if researchers focus narrowly on model fit, they may retain potentially problematic items for scale/subscale scoring. For example, if model fit indices suggest acceptable fit, researchers may overlook items that are problematic to include in scales due to loading weakly on their targeted factors (e.g., items assessing emptiness may load much less strongly on a latent borderline PD factor than affective instability items; Johansen et al., 2004). Related issues include the possibility that well-fitting multifactor confirmatory models could have items that (a) show strong cross-loadings on other factors on which they are not specified to load or (b) are assigned to load onto factors other than those on which they load most strongly. However, such aspects of model misspecification may go undetected without careful consideration of alternative structures (Greene et al., 2022), which may hinder measure development efforts and the application of structural models for informing assessment more generally (Loevinger, 1957; Jackson, 1970; Sellbom & Tellegen, 2019).

The issues of items being misassigned to factors or failing to be clear indicators of a single factor may be particularly salient when examining structural models of symptoms, given that different symptom experiences often are closely interrelated (Clark & Watson, 2019; Kotov et al., 2017). For example, items assessing "having thoughts that don't make sense to others" often are used to score thought disorder scales, but individuals with high levels of internalizing psychopathology may have worries that seem irrational to others and may strongly endorse these items as a result (Samuel et al., 2018). In such cases then, a two-factor model consisting of thought disorder and internalizing dimensions potentially could fit well if an item assessing "thoughts that don't make sense to others" is allowed to load only on thought disorder, even when such an item could have a loading of equal or stronger magnitude on internalizing.

Study aims and demonstrations

In this study, we show how focusing narrowly on the interpretation of model fit indices when evaluating factor models may lead to the previously described issues of (a) incorrectly concluding that heterogeneous item sets reflect narrow homogeneous constructs and (b) the retention of potentially

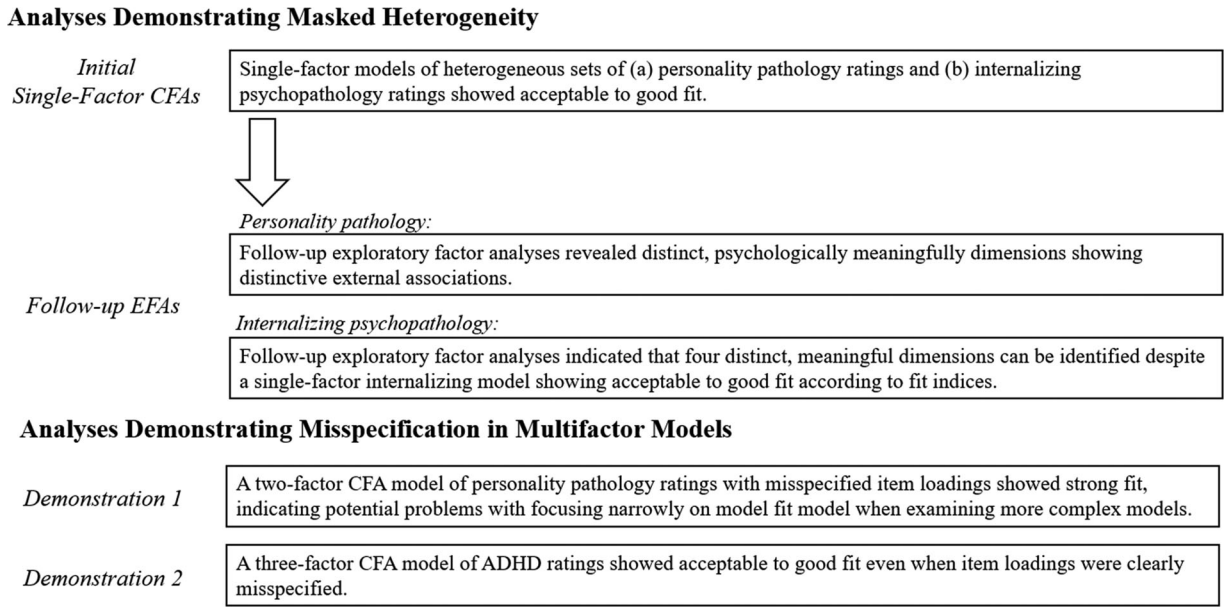

Figure 1. Overview of study data analytic demonstrations.
Note. Personality disorder data are from the outpatient sample ($N = 2,149$), and other analyses are based on the online community sample data ($N = 547$).

problematic items for scoring scales/subscales. We address these issues through a series of demonstrations using data from multiple samples and assessment methods as shown in Figure 1, which provides an overview of the specific analyses presented subsequently.

We present results from both EFA and CFA models illustrating these issues regarding model fit interpretation, beginning with an initial focus on single-factor CFA models. We focus on CFA models first because researchers often rely more heavily on making inferences based on model fit when using more confirmatory modeling approaches to (a) evaluate the acceptability of models and (b) inform interpretations of the degree to which item sets are homogeneous when examining single-factor structures (acknowledging parallels for using CFA, ESEM, and EFA when examining single factor solutions; Greene et al., 2022; Greiff & Heene, 2017).

After reporting the results of single-factor CFA models, we proceed to conduct follow-up EFAs to show that additional interpretable factors can be extracted even when model fit indices indicate acceptable to good fit for single-factor CFA structures. We then score subscales based on the results of these follow-up EFAs and examine these subscales' external correlates with measures of other psychopathology, personality, and psychosocial functioning. These analyses mimic how factor analytic approaches often are used to inform scale/subscale scoring (Clark & Watson, 2019) and illustrate that subscales corresponding with distinct dimensions from our EFAs show differential patterns of external correlates that would be obscured by focusing analyses solely on total scores reflecting a single dimension. Following that, we demonstrate that factor interpretability is important to consider in addition to model fit even when examining multifactorial CFA structures.

When reviewing the subsequent demonstrations, it is important to recognize that factor analytic approaches range in the degree to which they are exploratory to confirmatory rather than representing an exploratory versus confirmatory dichotomy (Chabrol et al., 2005; Greene et al., 2022; Schmitt et al., 2018; Wright, 2017). Although we use follow-up EFAs to highlight aspects of model misspecification for CFA models showing acceptable to good fit according to traditional interpretative cutoffs, we do not intend to suggest that more exploratory analyses are superior to applications of more confirmatory approaches in all situations and contexts. Indeed, other recent demonstrations highlight how use of more exploratory and confirmatory approaches in tandem can be useful for advancing knowledge of personality and psychopathology structure (Greene et al., 2022; Schmitt et al., 2018). Furthermore, in addition to EFA, we could have used other data analytic approaches such as comparing multiple CFA model configurations. We chose to use EFA after examining initial CFA models because it represented a straightforward option for demonstrating that models can have adequate to good fit according to traditional benchmarks even when item loadings are clearly misspecified and/or other model features are problematic.

Many of the issues discussed here also are applicable to measure development efforts and maximizing measures' construct validity (also see Clark & Watson, 2019 & Loevinger, 1957 for discussion of the distinction between the reliability and validity of "measures" versus "measurements" of constructs). More exploratory approaches may be particularly useful in earlier stages of the measure development process (e.g., EFAs with no loadings specified) to determine the extent to which emergent structures are consistent with general theoretical expectations (e.g., is each factor well-defined when extracting a specific number of factors based on theoretical considerations; Clark & Watson, 2019; Greene et al., 2022). Approaches traditionally described as confirmatory are useful for evaluating models with an increasing number of constraints. At minimum, CFA requires specifying which indicators are allowed to load onto which factors. However,

even with use of CFA, other model features (e.g., the magnitude of factor loadings, the magnitude of interfactor correlations) typically are not specified a priori, such that these models rarely are entirely confirmatory (Greene et al., 2022). Keeping these issues in mind, our EFAs following our initial CFAs demonstrate potential pitfalls with focusing narrowly on model fit when evaluating model suitability through a series of basic illustrative examples spanning different personality and psychopathology domains.

Method

Sample one: Interview data from the adult outpatient sample

Sample description

All participants across samples provided their informed consent for participation, and all research procedures received institutional ethics board approval. We did not receive ethics board permission to share data from either of these samples to open source repositories. However, frequencies and descriptive statistics for all items included in subsequent factor analyses is provided on the Open Science Framework (see https://tinyurl.com/f9j6mrc2). Data analytic syntax and output for these factor analyses and polychoric correlation matrices for all items used in these analyses also are available at this link. Full study datasets are available upon request from the first author.

Regarding the first sample, participants were 2,149 adult outpatients (mean age = 38.5, $SD = 12.5$) who completed interview assessments as part of the Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) Project (Zimmerman, 2016). Data were collected at treatment intake, and we included data from all participants regardless of specific diagnosis or diagnoses. The majority of participants were female (61.0%). Most participants identified as being White or European American (90.7%), 4.4% identified as Black or African American, with remaining participants endorsing other identities. Highest level of education was as follows: 40.6% with an associate degree or some college; 21.9% high school degree or equivalent; 15.1% with a 4-year college degree; 14.2% with some graduate school or a graduate/professional degree; and 8.3% with less than a high school or equivalent level of education.

Interview measures and procedure

Our factor analyses focus on nine item-level ratings of borderline, avoidant, schizotypal, and paranoid PD traits drawn from the Structured Interview for *DSM-IV* Personality (SIDP-IV; Pfohl et al., 1997). Items were rated on a 0 (*not present*) to 3 (*strongly present*) scale, and frequencies for each rating are shown in Online Supplemental Table S1. Borderline (ratings for criteria 1, 2, 3, and 7) and avoidant PD ratings (criterion 2 rating; fear of not being liked) were used to assess identity and relationship disturbance. Schizotypal (criterion 9 rating; social anxiety due to paranoia) and paranoid PD ratings (ratings for criteria 1, 4, and 5) were used to assess mistrust. These ratings were selected because they (a) all reflect various aspects of interpersonal dysfunction and/or (b) load strongly onto a general PD factor (Sharp et al., 2015). We focused on limited item sets in this and subsequent examples to provide straightforward demonstrations of issues related to focusing narrowly on model fit, not because we intended to conduct comprehensive examinations of psychopathology structure.

Other variables used for our analyses focused on determining external correlates of PD factors included (a) antisocial PD ratings obtained via the SIDP-IV, (b) lifetime internalizing and externalizing disorder ratings based on the Structured Clinical Interview for *DSM-IV* Axis I Disorders (SCID-IV; First et al., 1995), and (c) other clinically-relevant ratings (e.g., suicide attempt history) from the Schedule for Affective Disorders and Schizophrenia (SADS; Endicott & Spitzer, 1978; see Online Supplemental Table S2 for item descriptive statistics). Regarding antisocial PD ratings, complete data for specific conduct disorder ratings were not available for many participants. As a result, participants were rated as meeting criteria for antisocial PD based on having three or more antisocial PD traits present (4.3% sample prevalence).

Internalizing disorders included major depressive disorder (MDD; lifetime prevalence 64%), social anxiety disorder (29.8%), panic disorder (25.8%), generalized anxiety disorder (19.7%), and persistent depressive disorder (10.3%). Externalizing disorders included alcohol use disorder (40.9%), tobacco use disorder (20.6%), cannabis use disorder (16.1%), intermittent explosive disorder (6.1%), and antisocial PD (4.3%). We also summed scores on these dichotomous diagnostic ratings (e.g., if MDD history was rated as present, then a score of "1" was computed toward the composite total) to create internalizing and externalizing composite variables; thus, internalizing and externalizing scores each ranged from 0 to 5. Extensive information about the training process for interviewers (PhD-level psychologists or bachelor's level research assistants) and interrater reliability data described in other articles (Stanton et al., 2018; Zimmerman, 2016). For example, interrater reliability analyses for all internalizing diagnoses in these data exceed .80, with similar estimates reported for externalizing ratings.

Sample two: Self-report data from the online community sample

Sample description

Participants were 547 US adults (mean age = 38.0 years, $SD = 12.3$) recruited using Amazon Mechanical Turk. Most participants were women (59%), 40% were men, and 0.5% were nonbinary (the small remaining percentage of participants did not provide this information). A small percentage of participants (1.2%) were transgender. Most participants identified as White or European American (74.2%; 9.7% Asian American; 7.9% Black or African American; remaining participants endorsed other identities); 5.3% reported being Hispanic or Latino/Latina. The most common responses for highest level of education were 38.6% having bachelor's degree and 37.7% having completed some college

or an associate's degree. Finally, 18.1% of participants reported receiving medication to treat psychiatric issues, and 9.1% reported currently receiving psychotherapy.

Self-report measures

Our first set of factor analyses using data from this sample focused on the 27 items from the Dysphoria (10 items), Lassitude (6 items), Social Anxiety (6 items), and Ill-Temper (5 items) scales of the Expanded Version of the Inventory of Depression and Anxiety Symptoms (IDAS-II; Watson et al., 2012). Items from these scales were selected to assess negative emotion and cognitive patterns characteristic of the internalizing domain. Participants responded to the IDAS-II items in reference to the past 2 weeks using a 5-point scale ranging from 1 (*not at all*) to 5 (*extremely*). Additional factor analyses focused on item-level data from the 18-item Adult ADHD Self-Report Scale (ASRS; Kessler et al., 2005), which assesses inattentiveness, motor hyperactivity/impulsivity, and verbal hyperactivity/impulsivity symptoms (Stanton et al., 2018). Participants responded to these items using a scale ranging from 0 (*never*) to 4 (*very often*) in reference to the past 6 months.

We also examined how subscales created based on factor analyses of the IDAS-II items associated with scores from both (a) the second edition of the Big Five Inventory (BFI-2; Soto & John, 2017) and (b) the Short Dark Triad (SD3; Jones & Paulhus, 2014). The BFI-2 assesses five-factor model personality traits, and the SD3 assesses Machiavellianism, psychopathy, and grandiose narcissism using 9 items to assess each construct. Participants responded to the BFI-2 and SD3 using a scale ranging from 1 (*disagree strongly*) to 5 (*agree strongly*). Finally, we included trait scores from the Broad Autism Phenotype Questionnaire (BAPQ; Hurley et al., 2007). The BAPQ assesses three phenotypic dimensions relevant to autism: Aloofness (12 items), Pragmatic Language Difficulties (e.g., "out of sync in conversations"), and Rigidity (12 items; "very set in my ways"). Participants responded to the BAPQ items using a scale ranging from 1 (*very rarely*) to 6 (*very often*). Descriptive statistics and coefficient omega estimates for all measures are provided in Online Supplemental Table S3.

Data analytic overview

An overview of study analyses are presented in Figure 1 as reviewed. First, we present results from single-factor CFA models in both datasets. Second, we conduct EFAs showing that distinct dimensions can be identified using these same item sets. The magnitude of factor loadings in CFA models were not specified even though all items were specified to load onto a single factor in all initial CFA models. We created scales to represent dimensions from both single-factor CFA and multifactor EFA models for subsequent analyses examining external correlates. When creating scales to model single-factor solutions, scores for items loading $> |.40|$ on factors were included in scale scoring. For multifactor solutions, we scored subscales using items with (a) absolute loadings $\geq .40$ on their primary factor and (b) absolute cross-loadings $\leq .30$ on other factors, consistent with measure development recommendations (Clark & Watson, 2019). Where relevant, we compared subscale correlations using a Fisher's r-to-z transformation and conducted two-tailed within-sample z-tests of their differences. We used a threshold of $p < .001$ for evaluating all difference tests, acknowledging that relatively small differences in correlations still may have been likely to be significant at this level due to our use of large sample sizes for analyses. Finally, we present results from multifactor CFA models to show that model fit indices can indicate good fit even when item loading specifications are problematic. We made no specifications when estimating these CFA models other than specifying which item indicators were allowed to load on specific factors, with items being allowed to load on a single factor only across CFA analyses.

All factor analyses were conducted using Mplus Version 8 (Muthén & Muthén, 2017). We estimated all factor models using a weighted least squares mean and variance adjusted (WLSMV) estimator given our focus on analyzing item-level data, and all EFA analyses were conducted using a promax rotation. To evaluate fit for CFA models, we considered fit indices that commonly are reported when evaluating models, including the confirmatory fit index (CFI), the Tucker-Lewis Index (TLI), RMSEA, and the standardized root mean squared residual (SRMR). For RMSEA and SRMR, values $\leq$ .08 often are interpreted as indicating acceptable fit, and CFI and TLI values $\geq .95$ commonly are interpreted as indicating adequate to good fit (Hu & Bentler, 1999). Nevertheless, there is not clear consensus on specific cutoff values (McNeish & Wolf, 2021). For example, CFI and TLI values $\geq .90$ commonly are interpreted as indicating acceptable fit across literatures, as are RMSEA values of $\leq .10$ (Hopwood & Donnellan, 2010). We also present McDonald's omega (ω) and coefficient alpha (α) estimates where relevant.

Results

Single-factor confirmatory factor analytic models and masked item heterogeneity

Sample one: Interview ratings in adult outpatients

Single-factor CFA solution. The single-factor CFA model of the PD ratings yielded acceptable to good model fit (RMSEA = .063; CFI = .941; TLI = .921; SRMR = .060; χ^2 = 258.994, $df = 27$). Additionally, as shown in Table 1, all standardized factor loadings were $\geq .48$ on this factor, which we labeled *Interpersonal Dysfunction*.

Follow-up exploratory factor analysis and external validation. Although a single-factor CFA model was viable, we conducted a follow-up EFA to show that distinct dimensions could be identified in these data. Prior studies have not examined the factor structure and multidimensionality of the nine specific PD ratings used here to our knowledge. Therefore, we conducted a parallel analysis to help inform

Table 1. Factor loadings of personality pathology ratings from exploratory and confirmatory models in the outpatient sample.

	One-factor confirmatory	Exploratory model		Misspecified confirmatory	
Personality rating	Interpersonal dysfunction	Identity disturbance	Suspiciousness	Identity disturbance	Suspiciousness
Identity Disturbance Subscale Items					
BPD 3: Unstable sense of self	**.75**	**.79**	−.01	**.77**	–
BPD 2: Unstable/intense relationships	**.72**	**.59**	.18	**.74**	–
BPD 7: Chronic feelings of emptiness	**.66**	**.77**	−.08	**.68**	–
BPD 1: Frantically avoids abandonment	**.63**	**.64**	.02	**.65**	–
AVPD 2: Afraid of not being liked	**.48**	**.41**	.10	**.49**	–
Suspiciousness Subscale Items					
PARPD 1: Suspicious of being exploited	**.59**	−.14	**.85**	–	**.67**
PARPD 4: Suspicious of others' remarks	**.66**	.15	**.61**	–	**.78**
PARPD 5: Persistently bears grudges	**.57**	.09	**.56**	–	**.65**
STYPD 9: Social anxiety due to paranoia	**.53**	.16	**.43**	**.54**	–

$N = 2{,}149$. All ratings shown are from the Structured Interview for *DSM-IV* Personality (SIDP-IV), and factor loadings > .40 are **bolded**. Factors from the misspecified two-factor confirmatory correlated .73, and factors from the exploratory factor analytic model correlated .65. The label and number before each rating represent the respective disorder and criterion that rating is used to assess. AVPD = Avoidant personality disorder; BPD = Borderline personality disorder; PARPD = Paranoid personality disorder; STYPD = Schizotypal personality disorder.

Table 2. Correlations for the personality pathology scales with other interview ratings.

Diagnosis/Variable	Interpersonal Dysfunction	Identity Disturbance	Suspiciousness
Internalizing Composite	**.34**	**.36**	.17
Social Anxiety Disorder	**.39**	**.43**	.18
Persistent Depressive Disorder	.19	**.22**	.07
Generalized Anxiety Disorder	**.22**	**.21**	.16
Panic Disorder	.16	.18	.08
Major Depressive Disorder	.15	.16	.08
Externalizing Composite	**.23**	**.20**	.18
Antisocial Personality Disorder	**.32**	**.29**	**.25**
Intermittent Explosive Disorder	.14	.08	.17
Alcohol Use Disorder	.18	.18	.11
Cannabis Use Disorder	.19	.18	.13
Tobacco Use Disorder	.15	.13	.14
Other Clinically-Relevant Ratings			
Clinician global functioning rating	**−.36**	**−.36**	−.23
5-year psychosocial functioning impairment	**.32**	**.32**	.21
Worked missed in the past 5 years	.23	.23	.14
Lifetime number suicide attempts	.18	.19	.09
Lifetime number inpatient hospitalizations	.15	.16	.07

$N = 2{,}149$ for ratings with the Internalizing and Externalizing Composites and all diagnostic ratings; $N = 2{,}141$ for all other clinical ratings. Correlations with individual diagnoses are polyserial correlations; all other correlations are Pearson correlations. Correlations ≥ |.20| are **bolded**, and all correlations ≥ |.09| were significant at a $p < .001$ level. Underlined correlations were significantly different for Identity Disturbance and Suspiciousness at a $p < .001$ level.

how many dimensions should be extracted in these data. Parallel analysis indicated that up to two factors could be extracted (sample eigenvalues = 2.82, 1.11, 1.02; random data eigenvalues = 1.10, 1.07, 1.04). Consistent with this, two interpretable dimensions emerged when extracting two factors, as shown in the middle columns of Table 1. Items assessing a negative, unclear self-image (e.g., unstable sense of self) and needing validation from others (e.g., fearing abandonment) loaded strongly on Factor I, which we labeled *Identity Disturbance*. Items assessing mistrust of others (e.g., suspicious of being exploited) loaded strongly onto Factor II, which we labeled *Suspiciousness* (correlation between Factor I and II = .65).

We scored Identity Disturbance (5 items; α= .66; ω = .67; $M = 2.0$, $SD = 2.5$) and Suspiciousness (4 items; α= .55; ω = .59; $M = 1.0$, $SD = 1.5$) subscales based on these results. Estimates for α and ω values for these subscales were slightly lower than cutoff values (e.g., ≥ .70) commonly interpreted as indicating acceptable internal consistency; this likely was due at least in part to these subscales' brevity, as indicators of internal consistency often increase as a function of item number (Clark & Watson, 2019; Dunn et al., 2014). We also scored a 9-item scale reflecting the broader Interpersonal Dysfunction factor (α = .72; ω = .73; $M = 3.0$, $SD = 3.4$) to evaluate the degree to which subscale correlations differed from those for this composite.

Table 2 provides associations for this composite scale and these two subscales. The Identity Disturbance and Suspiciousness subscales showed similar, weak associations with externalizing ratings. However, Identity Disturbance showed correlates that were significantly stronger in magnitude than those for Suspiciousness with all clinically relevant ratings and nearly all internalizing ratings when examining *z*-test correlation differences. Several differences for Identity Disturbance and Suspiciousness were pronounced, including differences in associations with social anxiety (*r*s = .43 and .18, respectively) and the internalizing composite rating (*r*s = .36 and .17, respectively). When considering subscale versus composite correlations, patterns of associations for the general Interpersonal Dysfunction composite paralleled associations for the Identity Disturbance subscale very closely, such that scores on this subscale appear to have driven

Table 3. Factor loadings for internalizing symptom ratings in the online community sample.

Item	General Internalizing	Dysphoria	Lassitude	Ill-temper	Social Anxiety
Felt inadequate[a]	**.87**	**.73**	.13	−.06	.20
Felt discouraged[a]	**.89**	**.70**	.19	.06	.07
Felt depressed[a]	**.86**	**.62**	.15	.24	−.01
Little interest in usual activities	**.83**	**.44**	**.41**	.13	−.02
Blamed myself for things[a]	**.85**	**.58**	.06	.12	.24
Worried all the time[a]	**.86**	**.44**	.22	.19	.17
Felt drowsy[b]	**.76**	−.03	**.84**	.01	.03
Felt exhausted[b]	**.80**	.20	**.75**	.04	−.09
Trouble waking up[b]	**.75**	−.02	**.73**	−.02	.15
Felt worse in the morning[b]	**.78**	.04	**.70**	.08	.07
Trouble concentrating[b]	**.83**	.18	**.57**	.04	.16
Took effort to get going	**.87**	.38	**.56**	−.06	.10
Felt fidgety, restless[b]	**.75**	.23	**.52**	.13	−.01
Trouble making up my mind[b]	**.82**	.25	**.50**	.09	.11
Slept more than usual[b]	**.64**	−.03	**.41**	.11	.24
Felt enraged[c]	**.80**	.04	−.02	**.89**	.02
Was furious[c]	**.81**	.15	.03	**.81**	−.05
Lost my temper[c]	**.72**	−.04	.04	**.73**	.11
Little things made me mad[c]	**.85**	.04	.20	**.69**	.06
Felt like breaking things[c]	**.83**	.15	.07	**.63**	.12
Talked more slowly	**.75**	−.18	.26	**.42**	.38
Anxious about speaking[d]	**.79**	.00	.08	−.01	**.82**
Difficulty talking with others[d]	**.76**	−.07	.17	.01	**.77**
Worried about embarrassment[d]	**.78**	.19	−.03	.00	**.76**
Became anxious in crowds[d]	**.81**	.15	.13	−.01	**.68**
Felt self-conscious[d]	**.79**	.17	−.03	.17	**.61**
Difficulty with eye contact[d]	**.85**	.24	.00	.18	**.59**

$N = 547$. All items are from the Expanded Version of the Inventory of Depression and Anxiety Symptoms. All loadings $\geq .40$ are **bolded**. [a] = item scored in the Dysphoria scale for subsequent analyses; [b] = item scored for Lassitude; [c] = item scored for Ill-Temper; [d] = item scored for Social Anxiety. The Dysphoria factor correlated .66, .61, and .61 with the Lassitude, Ill-Temper, and Social Anxiety factors, respectively. The Lassitude factor correlated .67 and .68 with the Ill-Temper and Social Anxiety factors, respectively, and the Ill-Temper and Social Anxiety factors correlated .66.

observed associations for the overall composite. Collectively then, Identity Disturbance and Suspiciousness showed key, distinctive associations that would be masked when focusing analyses solely on total scores reflecting a more general PD dimension.

Sample two: Self-rated internalizing symptoms in the online community sample

Single-factor CFA. Next, we examined these same issues using item-level data drawn from the four different IDAS-II scales. A single-factor model on which all items were specified to load generally fit well (CFI = .956; TLI = .952; SRMR = .052; RMSEA = .088; model χ^2 = 1684.412, $df = 324$). The RMSEA value for this model slightly exceeded some cutoffs values for determining acceptable fit (e.g., .08; Hu & Bentler, 1999), but we anticipate that many researchers would deem this model acceptable given other index values and because this RMSEA value still was lower than other commonly used cutoffs (i.e., ≤.10; see Hopwood & Donnellan, 2010). Additionally, as shown in Table 3, all items had standardized loadings ≥.70 on this single factor, with one exception (i.e., "slept more than usual"; loading = .64).

Follow-up exploratory factor analysis and external validation. Next, we examined a four-factor EFA model. We extracted four factors given that the items used for these analyses were drawn from four IDAS-II scales that have been validated extensively in prior factor analytic research across sample types (e.g., patient, community, undergraduate; Watson et al., 2012). This four-factor EFA model is shown in Table 3. All factors were interpretable, and each factor had at least five items with primary loadings ≥ .40 on them and with cross-loadings ≤ .30 on other factors. Factors from this model were strongly intercorrelated as noted in Table 3 (all interfactor correlations ≥ .60), but no interfactor correlation exceeded .70.

We scored Dysphoria (5 items; both α and ω = .93; $M = 10.8$, $SD = 5.8$), Lassitude (8 items; α and ω = .90; $M = 16.0$, $SD = 7.5$), Ill-Temper (5 items; α and ω = .90; $M = 8.4$, $SD = 4.5$), and Social Anxiety (6 items; α and ω = .91; $M = 10.9$, $SD = 5.9$) subscales for examining external correlates. Table 3 provides subscale scoring compositions, noting that three items were not included in subscale scoring because they were not clear indicators of any single factor. We also created a General Internalizing composite scale including scores from all 27 IDAS-II items (both α and ω = .97; $M = 51.4$, $SD = 23.7$).

Table 4 presents correlations for these internalizing measures with various trait measures. The General Internalizing composite and internalizing subscales had similar correlates in many ways (e.g., robust negative associations with BFI-2 Conscientiousness). Still, individual subscales showed some correlates that were distinctive from those for other subscales and the General Internalizing composite. For example, all subscales associated robustly with BFI-2 Negative Emotionality, but the correlation for Dysphoria (r = .71) was significantly stronger than that for any other subscale based on z-test correlation comparisons. Ill-Temper also showed some distinctive associations with BFI-2

Table 4. Correlations for internalizing symptom scales with personality and other trait dimensions.

Personality Measure	General Internalizing	Ill-temper	Dysphoria	Lassitude	Social Anxiety
Five-Factor Model Traits					
BFI-2 Negative Emotionality	**.68**	**.49**	**.71**	**.62**	**.58**
BFI-2 Agreeableness	**−.45**	**−.47**	−.38	−.38	**−.41**
BFI-2 Conscientiousness	**−.50**	**−.42**	**−.45**	**−.47**	**−.41**
BFI-2 Extraversion	**−.43**	−.25	**−.43**	−.37	**−.46**
BFI-2 Open-Mindedness	−.12	−.14	−.07	−.08	−.15
Other Trait Dimensions					
BAPQ Pragmatic Language	**.50**	**.42**	**.41**	**.44**	**.49**
SD3 Psychopathy	.37	**.47**	.26	.30	.35
BAPQ Aloofness	**.41**	.28	**.40**	.34	**.44**
SD3 Machiavellianism	.34	.34	.28	.29	.31
BAPQ Rigidity	.28	.23	.26	.23	.30
SD3 Narcissism	−.00	.12	−.10	.01	−.04

$N = 547$. All correlations shown are Pearson correlations. Correlations $\geq |.40|$ are **bolded**, and all correlations $\geq |.14|$ were significant at a $p < .001$ level. Correlations that are underlined are significantly different from those for all other subscales in the same row at a $p < .001$ level. BAPQ = Broad Autism Phenotype Questionnaire; BFI-2 = Big Five Inventory-2; SD3 = Short Dark Triad.

Extraversion ($r = -.25$) and SD3 Psychopathy ($r = .47$; both correlations significantly different from all other subscale correlations).

Comparing correlations for two specific subscales also was informative. For instance, Dysphoria and Ill-Temper had significantly different correlates in the opposite direction with SD3 Narcissism ($rs = .12$ and $-.10$, respectively; $p < .001$ for difference). As another example, Social Anxiety correlated significantly more strongly with BAPQ Aloofness than did Ill-Temper ($rs = .44$ and $.28$; $p < .001$ for difference). Many of these correlations are not novel or surprising (e.g., social anxiety with aloofness), and differences in correlational patterns for subscales were less pronounced than in our example focused on PD interview ratings. However, examining these associations indicates that in addition to be differentiable on content grounds, these four internalizing subscales showed specificity in their associations in some ways.

Acceptable model fit in multifactor models but potential item misspecification

Sample one: Interview ratings in adult outpatients

Next, we demonstrate that even if multifactor solutions are examined, problematic item misspecification can be missed by focusing too heavily on model fit. Returning to the PD outpatient interview data, the far-right columns of Table 1 present standardized loadings from a misspecified two-factor CFA model of Identity Disturbance and Suspiciousness. All items showed strong loadings (i.e., > .45) on their assigned factors. Model factors correlated strongly (.73), though at a level that may be deemed acceptable by many researchers given that CFA can inflate interfactor correlations (Hopwood & Donnellan, 2010). Furthermore, this model fit well (RMSEA = .046; CFI = .970; TLI = .959; SRMR = .051; model $\chi^2 = 142.824$, $df = 26$), such that it likely would be deemed sufficient to guide subscale scoring when developing measures.[1]

Nevertheless, comparing the two-factor CFA and EFA solutions in Table 1 reveals possible aspects of CFA model misspecification that could be overlooked. Specifically, the schizotypal PD criterion rating reflecting "social anxiety due to paranoid fears" is assigned to and loads strongly on the Identity Disturbance factor in the CFA model, even though this item loads much more strongly on the Suspiciousness factor in the EFA model (loadings = .16 and .43, respectively in the two-factor EFA). Assigning this schizotypal PD rating to Identity Disturbance could be plausible theoretically given that it assesses social anxiousness, as do some of the specific avoidant and borderline PD items.

Sample two: Self-rated ADHD symptoms in the online community sample

Next, we tested a misspecified variation of a three-factor model of ADHD Inattentiveness, Verbal Hyperactivity/Impulsivity, and Motor Hyperactivity/Impulsivity identified in prior examinations of the ASRS's factor structure (e.g., Stanton et al., 2018) to demonstrate these same issues with item misspecification using self-report data. Note that we also could have demonstrated item misspecification for well-fitting models with the previously reviewed IDAS-II internalizing symptom data (e.g., all model fit indices indicated good fit when select items assessing irritability were assigned to load onto a Lassitude rather than Ill-Temper factor); however, we leverage the range of symptom measures included in this dataset to highlight the relevance of these issues across measures assessing a range of constructs.

The three-factor CFA structure of ADHD symptoms shown in Table 5 showed acceptable to good fit based on information provided from different fit indices (RMSEA = .078; CFI = .960; TLI = .954; SRMR = .038; model $\chi 2 = 571.904$, $df = 132$). All items also loaded strongly on their assigned factors (i.e., > .60).[2] Similar to other examples, interfactor correlations were strong in magnitude (i.e.,

[1]Other well-fitting but misspecified models also could be identified by assigning items to load onto factors in a manner different from the example here, such that this is only a one demonstrative example of misspecification. For example, model fit also was good (e.g., RMSEA = .047, CFI = .986, TLI = .956) for another model where the avoidant PD item assessing fear of not being liked was specified to load onto the Suspiciousness factor.

[2]Once again, other configurations of assigned factor loadings could have been used to demonstrate model fit indices indicating good fit even in the context of model misspecification (e.g., well-fitting models can be identified when specific items assessing motor hyperactivity/impulsivity are specified to load onto the Verbal Hyperactivity/Impulsivity factor).

Table 5. Factor loadings of ADHD symptom ratings from three-factor EFA and CFA models in the online community sample.

	Exploratory factor model			Misspecified confirmatory model		
ADHD Rating	Inattentiveness	Verbal Hyp	Motor Hyp	Inattentiveness	Verbal Hyp	Motor Hyp
2. Have difficulty getting things in order	**.89**	−.03	−.01	**.84**	.-	.-
8. Difficulty sustaining attention for boring work	**.74**	−.07	.17	**.80**	.-	.-
4. Avoid or delay starting hard tasks	**.81**	−.02	−.05	**.74**	.-	.-
1. Have trouble wrapping up projects	**.78**	.07	−.09	**.75**	.-	**.-**
10. Misplace or have difficulty finding things	**.70**	.08	.01	**.77**	.-	.-
11. Often distracted by activity and noises	**.65**	.03	.14	**.78**	.-	.-
7. Make careless mistakes on difficult projects	**.64**	.15	.07	**.81**	.-	.-
3. Trouble remembering appointments	**.61**	.12	.08	**.75**	.-	.-
9. Difficulty concentrating when listening	**.50**	.13	.23	**.-**	.-	**.82**
18. Interrupt others when they are busy	.05	**.84**	−.10	**.-**	**.78**	.-
17. Difficulty waiting turn in different situations	.01	**.82**	.00	**.-**	**.82**	.-
16. Trouble waiting to respond in conversations	−.01	**.74**	−.05	**.-**	**.68**	.-
15. Talk too much in social situations	.03	**.59**	.13	**.-**	**.75**	.-
12. Leave seat when expected to remain seated	.16	.30	.30	**.-**	.-	**.68**
13. Often feel restless or fidgety	.06	−.03	**.86**	**.-**	.-	**.81**
5. Fidget or squirm often	.10	−.12	**.85**	**.-**	.-	**.77**
6. Feel overly active or compelled to do things	−.13	.17	**.60**	**.-**	.-	**.56**
14. Have difficulty unwinding and relaxing	.21	.12	**.50**	**.-**	.-	**.76**

N = 547. All items shown are from the Adult ADHD Self-Report Scale, and items are paraphrased versions of the originals. Hyp = Hyperactivity/Impulsivity. In the exploratory model, Inattentiveness correlated .62 and .64 with Verbal and Motor Hyperactivity/Impulsivity, respectively; Verbal and Motor Hyperactivity/Impulsivity correlated .63 in this model. In the misspecified confirmatory model, Inattentiveness correlated .68 and .82 with Verbal and Motor Hyperactivity/Impulsivity, respectively; Verbal and Motor Hyperactivity/Impulsivity correlated .73 in this model.

rs for Inattentiveness with Verbal and Motor Hyperactivity/Impulsivity = .68 and .82, respectively, *r* for Verbal and Motor Hyperactivity/Impulsivity = .73), but this model still could be deemed suitable for guiding ASRS subscale scoring on the basis of these item loadings and this model generally fitting well. Importantly, however, this model still fit well despite ASRS item 9 ("difficulty concentrating), a clear inattentiveness indicator, being assigned to load onto the Motor Hyperactivity/Impulsivity factor. In fact, this item had the *strongest* loading on Motor/Hyperactivity of any item.

Based on existing structural research indicating that the ASRS items may reflect defined as many as three distinct factors (Gibbins et al., 2012; Stanton et al., 2018), we examined a three-factor EFA structure, which also is presented in Table 5. As shown in Table 5, item 9 loaded most strongly onto Inattentiveness as anticipated when extracting three factors using EFA (loading = .50, loadings < .25 on other factors). This was the case even though this item had a very strong loading on Motor Hyperactivity/Impulsivity when specified to load onto that factor in the CFA model reviewed previously.

A more subtle aspect related to this specific example with ADHD ratings focuses on item 12 (i.e., "leaves seat"), which loaded strongly (loading = .68) onto its assigned Motor Hyperactivity/Impulsivity factor in the CFA model. However, in the EFA model, it showed loadings of the same magnitude (.30) on Verbal Hyperactivity/Impulsivity and Motor Hyperactivity/Impulsivity. Consequently, if these analyses were conducted to inform subscale scoring, this item would be identified as a "splitter" and may not be included in subscale scoring as a result (Clark & Watson, 2019). Thus, had the original misspecified three-factor CFA model been identified as suitable, both items 9 and 12 potentially would have been inappropriately scored for subscales representing factors. Finally, interfactor correlations for the three-factor EFA model were weaker than those from the CFA model (e.g., *rs* for Verbal and Motor Hyperactivity/Impulsivity were .63 in the EFA but .73 in the CFA model; see Table 5 for all factor intercorrelations).

Discussion

These examples using data from multiple samples and methods demonstrate potential problems with failing to (a) consider valid, alternative multifactor solutions when single-factor models fit well and (b) identify items with misassigned loadings on specific factors in multifactor models. Regarding this first issue, we showed that well-fitting models could lead to potentially problematic interpretations that heterogeneous item sets are instead homogeneous. First, our analyses using a heterogenous set of interview ratings assessing various criteria from four PDs demonstrated that even when a single-factor model fits well, distinguishable factors showing meaningfully different patterns of external correlates also could be identified. Additionally, our example focusing on self-rated internalizing symptoms also indicated that a single-factor model generally fit well according to most indices examined, even though these analyses were based on 27 items drawn from four different symptom scales. Taken together then, these demonstrations illustrate how focusing narrowly on model fit could lead researchers astray without careful consideration of item content and alternative multidimensional structures.

At the same time, we recognize that a narrow focus on model fit can lead to overfactoring and prioritizing models that are *overly complex* representations of data, even though our demonstrations did not focus explicitly on this issue. Model fit will almost invariably improve with each additional factor extracted, and narrowly focusing on model fit may result in researchers extracting additional factors that are poorly defined or difficult to interpret regardless of whether more exploratory or confirmatory approaches are applied (Montoya & Edwards, 2021). Thus, the goal of our

demonstrations was simply to show that researchers may make problematic inferences when fit indices for single-factor models satisfy widely used cutoff values. For instance, borderline PD may be reified as a unidimensional construct based on single-factor models of borderline PD ratings showing good fit as noted (e.g., Clifton & Pilkonis, 2007). Along with other research, our results underscore the importance of considering other model characteristics when adjudicating model appropriateness, including factor loading patterns, item content alignment with construct definitions, the extent to which prioritizing more complex models meaningfully increments clinical assessment and prediction, and so on (Greiff & Heene, 2017; Sellbom & Tellegen, 2019).

Another important point related to balancing parsimony with capturing heterogeneity is that the same item scores may be valid indicators of both broad and narrow constructs (Clark & Watson, 2019). This is especially relevant to the application of hierarchical, dimensional frameworks such as the Hierarchical Taxonomy of Psychopathology (HiTOP; Kotov et al., 2017), which classifies symptom dimensions at both broad (e.g., the internalizing spectrum) and narrow (e.g., a dysphoria facet within internalizing) levels of abstraction. For instance, consider our examples focused on PD ratings. In this case, items reflecting both (a) Identity Disturbance and (b) Suspiciousness also were clear indicators of a general Interpersonal Dysfunction factor. As a result, they could be validly used to assess their respective specific factors *and* a more general factor depending on the goals of an analysis. Measures such as the Spectra: Indices of Psychopathology (Blais & Sinclair, 2018) that can be used to assess both specific dimensions (e.g., Aggression) and broad factors (e.g., Externalizing) demonstrate this approach. According to this perspective then, assessment focused on broad levels of abstraction (e.g., assessing broad spectra such as internalizing) are not "competing" with assessment of more specific dimensions (e.g., worry), as they could be complementary approaches for assessing psychopathology (e.g., see Blais & Sinclair, 2018; Stanton et al., 2020).

Identifying misspecification in the presence of acceptable model fit

As our other examples with the PD interview ratings and self-ratings of ADHD demonstrate, overemphasizing model fit can result in failure to detect other problematic aspects of models even when examining multifactor models. Although both examples presented reinforce these points, our demonstration with ADHD is particularly striking. In this example, even when items clearly assessing inattentiveness (e.g., concentration difficulties) were assigned to load onto hyperactivity/impulsivity factors when using CFA, model fit was acceptable to good when making comparisons against widely used interpretive benchmarks. There are ongoing debates in many substantive areas of personality and psychopathology research regarding the nature of specific model constellations that should be used as frameworks for guiding measurement (e.g., different conceptualizations of psychopathy structure; posttraumatic stress disorder structure; Schmitt et al., 2018; Veal et al., 2021). Keeping in mind that statistical models are imperfect representations of complex systems of traits and processes, assigning concentration items to load onto hyperactivity/impulsivity factors when examining ADHD symptom models represents an *obvious* error theoretically and based on prior research (Kessler et al., 2005; Martel et al., 2010; Stanton & Watson, 2016). However, it illustrates very well that models can fit well *even when* item misspecifications clearly are theoretically inconsistent. In many other cases, item misassignments often can be much more difficult to detect without careful consideration.

Recommendations for factor analytic and psychometric research

In what follows, we provide concrete recommendations to reduce the likelihood of the issues described occurring in measure development contexts and more generally. First, foundational steps include (a) clearly articulating construct definitions and (b) generating homogeneous item composites (HICs) representing each dimension predicted to underlie an item set when developing measures. These recommendations likely are familiar to researchers with expertise in psychometrics and have long been recognized as a fundamental in the measure development process (e.g., Jackson, 1970; Loevinger, 1957) but often remain neglected in measure validation and other research areas (Greiff & Heene, 2017; Sellbom & Tellegen, 2019).

We recognize that researchers hold differing viewpoints about the utility of more exploratory versus confirmatory factor analytic approaches at various phases of the measure development process and when examining personality and psychopathology structure (e.g., Veal et al., 2021). Acknowledging this, first considering more exploratory approaches may be particularly useful in early phases of the measure development process, rather than applying more confirmatory structures in initial development phases to "prove" that a structure is sufficient if fit indices indicate acceptable to good fit (Greiff & Heene, 2017; Sellbom & Tellegen, 2019). We also encourage researchers to consider the application of more exploratory approaches in cases where relatively little is known about factor structures a priori and/or when factor structures to be examined are likely to be complex in nature (Greene et al., 2022).

For example, the classification of dimensions traditionally defining *DSM* neurodevelopmental disorders within dimensional models such as the HiTOP remains unclear in some ways (Michelini et al., 2019). As a result, specifying more confirmatory models would be challenging and/or likely would requiring adjudicating amongst a very large number of possible model configurations. Model fit increasingly has become used when applying more exploratory factor analytic approaches as well, such that we caution against an overreliance on model fit regardless of whether more exploratory or confirmatory approaches are used (see Montoya & Edwards, 2021 for discussion of issues that may arise when using model fit interpretation to adjudicate amongst exploratory structures).

In addition to being useful for obtaining an initial understanding of indicator structure, use of more exploratory approaches early in the measure development process can improve assessment efficiency when developing measures. In our examples with the ADHD and internalizing item sets, exploratory approaches were useful for identifying items showing loadings of equal (or roughly equal) magnitudes on multiple factors when examining multifactor solutions. As described in contemporary measure development guidelines (Clark & Watson, 2019), it can be helpful to examine patterns of loadings across samples, to ensure that pattern loadings are not due to sample-specific idiosyncrasies. If these factor analytic results for internalizing and ADHD symptoms were used to guide measure development and were consistent with results from other samples, items that were not clear indicators of a single factor may not be included when scoring scales, allowing constructs to be assessed with fewer items.

Lastly, we have stressed the importance of careful scrutiny of item sets. However, we are not suggesting that researchers generate a theoretical mapping of anticipated dimensions and adhere to it rigidly when provided with contrary evidence. It could be problematic, for instance, to rigidly adhere to a theoretical model by ignoring important contradictory information provided by EFA. As an example of an alternative, more appropriate, data-driven approach, Watson et al. (2012) hypothesized that multiple dimensions assessing different aspects of social anxiety would emerge as distinct when developing the IDAS-II. However, results across samples indicated that multiple, well-defined dimensions could not be identified when analyzing social anxiety item sets, such that the IDAS-II assesses social anxiety using a single scale.

Future directions, limitations, and conclusion

Several limitations and related future directions would be useful for advancing understanding of these issues related to factor analytic model interpretation. First, we did not examine other related issues such as how the same items may function differently across sample types. Examining cross-sample issues such as these has been informative both in measure development research and other work focused on interpreting the replicability and substantive nature of factor structures (e.g., the *p* factor of psychopathology; Greene et al., 2022; Levin-Aspenson et al., 2021). Considering the relevance of this to our study, analyses involving the PD ratings included items assessing low-base rate symptoms such as paranoia, such that model fit and patterns of factor loadings could have varied in other samples (e.g., inpatient samples).

Other extensions of this research also would be interesting. For example, recent research indicates that the number of items used to assess the criteria for specific *DSM-5* disorders influences model fit and the extent to which criteria defining disorders such as alcohol use disorder appear unidimensional (Watts et al., 2021). However, our self-report dataset did not include assessment of the criteria for specific disorders, and our interview data included only single-item ratings for different PD criteria and no item level data for other diagnoses, which precluded direct examination of these issues across samples. Fit for all CFA models was good according to relaxed interpretative guidelines (e.g., CFI and TLI $\geq$ .90) and often exceeded more stringent cutoffs (e.g., CFI $\geq$ .960), such that we believe that it is plausible that researchers would deem these models acceptable in many cases. We again would like to acknowledge debate regarding the use of stringent versus more relaxed cutoffs (see Hopwood & Donnellan, 2010 for discussion), and we anticipate continued discussion and investigations into these issues. Future research in this area also would be useful for determining the utility of different interpretative cutoffs based on varying study design characteristics (e.g., fit varying based on the number of items used as reviewed; also see McNeish & Wolf, 2021). We anticipate ongoing debate regarding the suitability of more exploratory versus confirmatory factor analytic approaches in specific contexts (e.g., Veal et al., 2021; Hopwood & Donnellan, 2010; Perry et al., 2015). We agree with sentiments expressed by Hoekstra and Vazire (2021) that an openness to differing perspectives as evidence accumulates will be useful for improving measure development efforts and understanding of personality and psychopathology structure.

Acknowledging these future directions, our demonstrations illustrate that care is needed when interpreting factor analytic results in clinical and personality research. Although model fit indices provide important information and should not be disregarded, failing to consider other model characteristics can lead to misinterpretations regarding the nature of item sets. We hope that researchers will integrate these considerations when conducting factor analytic research and that other studies will extend investigation on these topics in the ways described.

Acknowledgments

We would like to thank all study participants for their time and effort in completing this study. We would like to thank Dr. Christina McDonnell (University of Wyoming; formerly University of Notre Dame) for her assistance in obtaining ethics approval for research used to obtain some study data presented here. We have no other acknowledgements to make.

Declaration of interest

The authors have no disclosures (financial or otherwise) to report.

Ethics approval

Research ethics committee approval was obtained for this research, and all individual research participants provided informed consent for their participation. Research ethics committee approval for the interview data and self-report data presented was obtained from the Rhode Island Hospital and University of Notre Dame Institutional Review Board.

Data availability statement

These studies and the associated analyses were based on data from pre-existing datasets and were not preregistered. Ethics approval was not

explicitly sought to post data presented in this manuscript to open access, online repositories, and thus, these data are not provided on such a forum. Please contact the lead author (Kasey Stanton; kaseyjstanton@gmail.com) should you have any questions or wish to access these data. Descriptive item information and data analytic output for all factor analyses conducted for this study are available on the open science framework at the following link: https://tinyurl.com/f9j6mrc2

References

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

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences*, *42*(5), 815–824. https://doi.org/10.1016/j.paid.2006.09.018

Blais, M., & Sinclair, S. J. (2018). *Spectra: Indices of psychopathology*. Psychological Assessment Resources.

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*(4), 465–484.

Carleton, R. N., Norton, P. J., & Asmundson, G. J. G. (2007). Fearing the unknown: A short version of the intolerance of uncertainty scale. *Journal of Anxiety Disorders*, *21*(1), 105–117. https://doi.org/10.1016/j.janxdis.2006.03.014

Chabrol, H., Ducongé, E., Casas, C., Roura, C., & Carey, K. B. (2005). Relations between cannabis use and dependence, motives for cannabis use and anxious, depressive and borderline symptomatology. *Addictive Behaviors*, *30*(4), 829–840.

Chmielewski, M., Bagby, R. M., Quilty, L. C., Paxton, R., & McGee Ng, S. A. (2011). A (re)-evaluation of the symptom structure of borderline personality disorder. *The Canadian Journal of Psychiatry*, *56*(9), 530–539. https://doi.org/10.1177/070674371105600904

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427.

Clifton, A., & Pilkonis, P. A. (2007). Evidence for a single latent class of diagnostic and statistical manual of mental disorders borderline personality pathology. *Comprehensive Psychiatry*, *48*(1), 70–78.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412. https://doi.org/10.1111/bjop.12046

Endicott, J., & Spitzer, R. L. (1978). A diagnostic interview: The schedule for affective disorders and schizophrenia. *Archives of General Psychiatry*, *35*(7), 837–844.

Feske, U., Kirisci, L., Tarter, R. E., & Pilkonis, P. A. (2007). An application of item response theory to the *DSM-III-R* criteria for borderline personality disorder. *Journal of Personality Disorders*, *21*(4), 418–433. https://doi.org/10.1521/pedi.2007.21.4.418

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1995). *Structured Clinical Interview for DSM-IV Axis I Disorders—Patient Edition (SCID-I/P, version 2.0)*. New York, NY: Department of Biometrics Research, State Psychiatric Institute.

Gibbins, C., Toplak, M. E., Flora, D. B., Weiss, M. D., & Tannock, R. (2012). Evidence for a general factor model of ADHD in adults. *Journal of Attention Disorders*, *16*(8), 635–644.

Gignac, G. E. (2007). Multi-factor modeling in individual differences research: Some recommendations and suggestions. *Personality and Individual Differences*, *42*(1), 37–48. https://doi.org/10.1016/j.paid.2006.06.019

Greene, A. L., Eaton, N. R., Li, K., Forbes, M. K., Krueger, R. F., Markon, K. E., Waldman, I. D., Cicero, D. C., Conway, C. C., Docherty, A. R., Fried, E. I., Ivanova, M. Y., Jonas, K. G., Latzman, R. D., Patrick, C. J., Reininghaus, U., Tackett, J. L., Wright, A. G. C., & Kotov, R. (2019). Are fit indices used to test psychopathology structure biased? A simulation study. *Journal of Abnormal Psychology*, *128*(7), 740–764.

Greene, A. L., Watts, A. L., Forbes, M. K., Kotov, R., Krueger, R., & Eaton, N. R. (2022). Misbegotten methodologies and forgotten lessons from Tom Swift's electric factor analysis machine: A demonstration with competing structural models of psychopathology. *Psychological Methods*. Advance online publication. https://psycnet.apa.org/record/2022-18281-001

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment*, *33*(5), 313–317. https://doi.org/10.1027/1015-5759/a000450

Hoekstra, R., & Vazire, S. (2021). Aspiring to greater intellectual humility in science. *Nature Human Behaviour*, *5*(12), 1602–1607. https://doi.org/10.1038/s41562-021-01203-8

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *14*(3), 332–346.

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

Hurley, R. S., Losh, M., Parlier, M., Reznick, J. S., & Piven, J. (2007). The Broad Autism Phenotype Questionnaire. *Journal of Autism and Developmental Disorders*, *37*(9), 1679–1690. https://doi.org/10.1007/s10803-006-0299-3

Jackson, D. N. (1970). A sequential system for personality scale development. *Current Topics in Clinical and Community Psychology*, *2*, 61–96.

Jackson, D. N., Ahmed, S. A., & Heapy, N. A. (1976). Is achievement a unitary construct? *Journal of Research in Personality*, *10*(1), 1–21. https://doi.org/10.1016/0092-6566(76)90079-9

Johansen, M., Karterud, S., Pedersen, G., Gude, T., & Falkum, E. (2004). An investigation of the prototype validity of the borderline *DSM-IV* construct. *Acta Psychiatrica Scandinavica*, *109*(4), 289–298.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the Short Dark Triad (SD3): A brief measure of dark personality traits. *Assessment*, *21*(1), 28–41. https://doi.org/10.1177/1073191113514105

Kessler, R. C., Adler, L., Ames, M., Demler, O., Faraone, S., Hiripi, E. V., Howes, M. J., Jin, R., Secnik, K., Spencer, T., Ustun, T. B., & Walters, E. E. (2005). The World Health Organization Adult ADHD Self-Report Scale (ASRS): A short screening scale for use in the general population. *Psychological Medicine*, *35*(2), 245–556. https://doi.org/10.1017/s0033291704002892

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*, *126*(4), 454–477. https://doi.org/10.1037/abn0000258

Latzman, R. D., Patrick, C. J., & Lilienfeld, S. O. (2020). Heterogeneity matters: Implications for Poeppl et al.'s (2019) meta-analysis and future neuroimaging research on psychopathy. *Molecular Psychiatry*, *25*(12), 3123–3124.

Levin-Aspenson, H. F., Watson, D., Clark, L. A., & Zimmerman, M. (2021). What is the general factor of psychopathology? Consistency of the p-factor across samples. *Assessment*, *28*(4), 1035–1049.

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

Martel, M. M., von Eye, A., & Nigg, J. T. (2010). Revisiting the latent structure of ADHD: is there a 'g' factor. *Journal of Child Psychology and Psychiatry*, *51*(8), 905–914. https://doi.org/10.1111/j.1469-7610.2010.02232.x

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*. https://doi.org/10.1037/met0000425

Michelini, G., Barch, D. M., Tian, Y., Watson, D., Klein, D. N., & Kotov, R. (2019). Delineating and validating higher-order dimensions of psychopathology in the Adolescent Brain Cognitive Development (ABCD) study. *Translational Psychiatry*, *9*(1), 261.

Montoya, A. K., & Edwards, M. C. (2021). The poor fit of model fit for selecting number of factors in exploratory factor analysis for scale evaluation. *Educational and Psychological Measurement*, *81*(3), 413–440.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Naragon-Gainey, K., & Watson, D. (2018). What lies beyond neuroticism? An examination of the unique contributions of social-cognitive vulnerabilities to internalizing disorders. *Assessment*, *25*(2), 143–158.

Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing model fit: Caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement in Physical Education and Exercise Science*, *19*(1), 12–21. https://doi.org/10.1080/1091367X.2014.952370

Pfohl, B., Blum, N., & Zimmerman, M. (1997). *Structured interview for DSM-IV personality*. American Psychiatric Press.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367. https://doi.org/10.1037/0033-295X.107.2.358

Samuel, D. B., Suzuki, T., Bucher, M. A., & Griffin, S. A. (2018). The agreement between clients' and their therapists' ratings of personality disorder traits. *Journal of Consulting and Clinical Psychology*, *86*(6), 546–555.

Schmitt, T. A., Sass, D. A., Chappelle, W., & Thompson, W. (2018). Selecting the "best" factor structure and moving measurement validation forward: An illustration. *Journal of Personality Assessment*, *100*, 345–362.

Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, *31*(12), 1428–1441.

Sharp, C., Wright, A. G. C., Fowler, J. C., Frueh, B. C., Allen, J. G., Oldham, J., & Clark, L. A. (2015). The structure of personality pathology: Both general ('g') and specific ('s') factors? *Journal of Abnormal Psychology*, *124*(2), 387–398. https://doi.org/10.1037/abn0000033

Smith, G. T., & McCarthy, D. M. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 300–308. https://doi.org/10.1037/1040-3590.7.3.300

Smith, G. T., McCarthy, D. M., & Zapolski, T. C. B. (2009). On the value of homogeneous constructs for construct validation, theory description, and description of psychopathology. *Psychological Assessment*, *21*(3), 272–284.

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

Stanton, K. (2020). Increasing diagnostic emphasis on negative affective dysfunction: Potentially negative consequences for differential diagnosis. *Clinical Psychological Science*, *8*(3), 584–589. https://doi.org/10.1177/2167702620906147

Stanton, K., & Watson, D. (2016). Adult ADHD: Associations with personality and other psychopathology. *Journal of Psychopathology and Behavioral Assessment*, *38*(2), 195–208. https://doi.org/10.1007/s10862-015-9519-5

Stanton, K., Forbes, M. K., & Zimmerman, M. (2018). Distinct dimensions defining the Adult ADHD Self-Report Scale: Implications for assessing inattentive and hyperactive/impulsive symptoms. *Psychological Assessment*, *30*(12), 1549–1559. https://doi.org/10.1037/pas0000604

Stanton, K., McDonnell, C. G., Hayden, E. P., & Watson, D. (2020). Transdiagnostic approaches to psychopathology measurement: Recommendations for measure selection, data analysis, and participant recruitment. *Journal of Abnormal Psychology*, *129*(1), 21–28.

Veal, R., Critchley, C., Luebbers, S., Cossar, R., & Ogloff, J. R. P. (2021). Factor structure of the Psychopathy Checklist: Screening Version (PCL:SV): A systematic review using narrative synthesis. *Journal of Psychopathology and Behavioral Assessment*, *43*(3), 565–582. https://doi.org/10.1007/s10862-021-09877-0

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

Watts, A. L., Boness, C. L., Loeffelman, J. E., Steinley, D., & Sher, K. J. (2021). Does crude measurement contribute to observed unidimensionality of psychological constructs? An example with *DSM-5* alcohol use disorder. *Journal of Abnormal Psychology*, *130*(5), 512–524. https://doi.org/10.1037/abn0000678

Watts, A. L., Poore, H. E., & Waldman, I. D. (2019). Riskier tests of the validity of the bifactor model of psychopathology. *Clinical Psychological Science*, *7*(6), 1285–1303. https://doi.org/10.1177/2167702619855035

Wright, A. G. C. (2017). The current state and future of factor analysis in personality disorder research. *Personality Disorders*, *8*(1), 14–25. https://doi.org/10.1037/per0000216

Zimmerman, M. (2016). A review of 20 years of research on overdiagnosis and underdiagnosis in the Rhode Island Methods to Improve Diagnostic Assessment and Services (MIDAS) project. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *61*(2), 71–79. https://doi.org/10.1177/0706743715625935