HHS Public Access

Author manuscript

Personal Disord. Author manuscript; available in PMC 2024 January 01.

Published in final edited form as:

Personal Disord. 2023 January ; 14(1): 105–117. doi:10.1037/per0000581.

Factor analysis in personality disorders research: Modern issues and illustrations of practical recommendations

Ashley L. Watts[1], **Ashley L. Greene**[2,3,*], **Whitney Ringwald**[4,*], **Miriam K. Forbes**[5,*], **Cassandra M. Brandes**[6,*], **Holly F. Levin-Aspenson**[7,*], **Colette Delawalla**[8,*]

[1]Department of Psychological Sciences, University of Missouri, Columbia, MO.

[2]VISN 2 Mental Illness Research, Education and Clinical Center, James J. Peters VA Medical Center, Bronx, NY.

[3]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY.

[4]Department of Psychology, University of Pittsburgh, Pittsburgh, PA.

[5]Centre for Emotional Health and School of Psychological Sciences, Macquarie University, Sydney, Australia.

[6]Department of Psychology, Northwestern University, Evanston, IL

[7]Department of Psychiatry and Human Behavior, Brown University Warren Alpert Medical School, Providence, RI.

[8]Department of Psychological Science, Ball State University, Muncie, IN.

Abstract

The development of factor analysis is uniquely situated within psychology, and the development of many psychological theories and measures are likewise tethered to the common use of factor analysis. In this paper, we review modern methodological controversies and developments of factor analytic techniques through concrete demonstrations that span the exploratory-confirmatory continuum. Also, we provide recommendations for working through common challenges in personality disorders research. To help researchers conduct riskier tests of their theory-implied models, we review what factor analysis is and is not, as well as some dos and don'ts for engaging in the process of model evaluation and selection. Throughout, we also emphasize the need for closer alignment between factor models and our theories, as well as clearer statements about which criteria would support or refute the theories being tested. Consideration of these themes appears promising in terms of advances in theory, research, and treatment surrounding the nature and impact of personality disorders.

Author correspondence: Correspondence should be addressed to Ashley L. Watts, 200 South Seventh Street, Columbia, MO, 65211. ashleylwatts@gmail.com.

*All authors following Watts share second authorship because they contributed to this manuscript equally.

Author note: All code and data are available on OSF (https://osf.io/dyth3/). All authors contributed to conceptualization and writing. M.K. Forbes (bassackwards analysis) and A.L. Watts (all other analyses) were responsible for data analysis. Thanks to Joshua D. Miller for contributing some of the data used in the demonstrations here.

Keywords

factor analysis; bassackwards; model equivalence; exploratory confirmatory continuum

Many renowned psychological theories and models were borne out of factor analysis. To name a few, Spearman's (1904) study of the interrelatedness of performance on varied intelligence measures generated the theory that a single dimension, "*g*", causes diverse types of intellectual ability; Allport's analysis of the language we use to describe people and their behavior led to the development of the Five-Factor Model of personality (Allport & Odbert, 1936); and factor analyses of various instruments designed to assess personality disorder (PD) inspired the DSM-5's shift towards a dimensional conceptualization of PDs (Krueger et al., 2012). In fact, no other statistical tool has been as integral to the creation, development, and refinement of psychological measurement and theories as factor analysis.

In this paper, we focus on modern issues and approaches to factor analysis in PD research. We place a particular emphasis on novel methodological developments and controversies surrounding the implementation and interpretation of factor analysis over the last decade or so. Our emphasis on recent developments and changing recommendations in the field extends many excellent reviews of factor analysis, basic and technical (Sellbom & Tellegen, 2019; Wright, 2017; Wright & Zimmermann, 2015).

Factor Analysis in PD Research

In the broadest sense, factor analysis is a form of latent variable modeling that aims to reduce a larger set of observed (manifest) variables into fewer unobserved (latent) variables called factors, a practice termed dimension reduction. Factors summarize the covariation among observed variables and partition variables into sets that tend to covary more strongly with each other than they do with other variables.

Worked examples.

PD researchers use factor analysis for many purposes, including to explore and validate the structure of an instrument, and to understand the latent structure of personality features or PD symptoms. For instance, factor analysis in PD research might be used to examine the interrelatedness of a broad swath of PD criteria or disorders, the structure of a single PD, or even the structure of a given measure of PD features or related traits (e.g., Conway et al., 2016; Sharp et al., 2015). Scattered around this paper, we provide worked examples of the types of factor analyses researchers might conduct, using data from 855 undergraduates from Emory University and the University of Georgia. Undergraduates self-reported DSM-IV-TR PD symptoms on the SCID-II PD questionnaire (First et al., 1995). In the spirit of accessibility, we rely on concrete demonstrations using PD data, with all analyses conducted in R and all data and code available on OSF (https://osf.io/dyth3/).

Exploratory factor analysis.

Exploratory factor analysis (EFA) was developed by Spearman (1904) to examine interrelatedness of intelligence tasks. The goal of EFA is to describe the number of latent

dimensions and to maximize the amount of variance explained in the data. EFA is referred to as “exploratory" because it does not require *a priori* hypotheses or specification surrounding the patterning of variables’ loadings (or strength of association with the factor) onto factors, and all variables are allowed to load onto all factors (Mulaik, 1987; Figure 1a). The researcher is not even required to have a hypothesis about the number of factors that best describe the data (though they may, and often do).

Because researchers can conduct an EFA with few expectations about its results, EFA requires the use of additional statistical methods to determine which model best describes the data. A critical issue surrounds how many factors one should extract. Historically, researchers relied on visual interpretation of the scree plot, which displays the eigenvalues (corresponding to the amount of variance explained in the variable set) for each factor, and selected the number of factors that appear prior to the elbow in the plot (i.e., 2 factors in Figure S1). Others have promoted the Kaiser criterion, arguing that one should extract the number of factors with an eigenvalue greater than 1. These practices are generally considered outdated (Preacher & MacCallum, 2003)[i] Modern methods include parallel analysis, Velicer’s minimum average partial (MAP) test, and the test of Very Simple Structure (VSS). Researchers should use multiple methods and compare their results, as there is no consensus on an optimal method. Because these methods often yield different factor structures (i.e., loading patterns and presence and/or strength of interfactor correlations), the final EFA solution is selected by weighting evidence from statistical methods to guide factor extraction, interpretability, theory, and compatibility with the existing literature.

In Example 1 (https://osf.io/dyth3/), we conducted an EFA with geomin (oblique, or correlated) rotation and a maximum likelihood estimator on the tetrachoric correlation matrix[ii] of 15 borderline PD symptoms. Factor rotation minimizes the complexity of factor loading matrices to get better differentiation among the factors, and to achieve a simpler structure for ease of interpretation. Parallel analysis suggested 7 factors or 2 principal components, and MAP and the VSS test indicated 2 factors.

Regarding the EFA solution of borderline PD symptoms, a 7-factor solution was clearly over-extracted (Table S1), such that one factor was indicated by a single symptom. In contrast, the 2-factor solution was clearly interpretable with moderately correlated (r = .54) factors, one reflecting Identity Disturbance (i.e., sense of self changed dramatically, depending on people with) and the other reflecting all other symptoms (the exception was frequent impulsive behaviors, which did not load strongly on either factor). By balancing interpretability and information from a variety of statistical methods to guide factor extraction, we settled on a 2-factor solution.

[i]Global fit assessments have historically supported hypothesis tests in the context of CFA (McNeish & Wolf, 2021). Some work argued that global fit statistics may be helpful for guiding decisions about the number of factors to retain in EFA (e.g., Clark & Bowles, 2018; Preacher et al., 2013), but there is growing evidence that fit indices perform too inconsistently to warrant use in this context (Montoya & Edwards, 2021).

[ii]Factor analysis of ordinal or binary items requires computing a polychoric or tetrachoric correlation matrix, respectively, and conducting the EFA on that, as opposed to using the raw data. Pearson correlations tend to produce inflated associations between categorical items. We have shown how to do this in R for EFA (Ex. 1) and CFA (Ex. 2).

As an aside, parallel analysis often provides results for both factor analysis and principal components analysis because they are, in many respects, viewed as complementary methods. Notwithstanding the fact that principal components analysis and factor analysis are dimension reduction techniques that produce linear combinations of variables to summarize a correlation matrix, they have several key differences that are worth mentioning. First, factor analysis corresponds to a measurement model of a latent variable, whereas principal components analysis does not. Second, factor analysis summarizes the shared variance among a variable set and principal components analysis summarizes both the shared and unshared variance. This difference arises because factor analysis assumes that variables have unique variances and principal components analysis does not. Third, principal components are always unrotated and orthogonal to one another, whereas factors can be rotated.

Confirmatory factor analysis.

More than 60 years after EFA was introduced, Jöreskog (1969) developed confirmatory factor analysis (CFA) to test more specific hypotheses regarding the hierarchical arrangement of performance on intelligence tasks. Over time, the use of CFA eclipsed EFA, as the field has increasingly prioritized theory testing about what was being measured over exploration of how many things might be measurable. In contrast with EFA, CFA requires researchers to make more decisions about model specification: the number of factors, whether those factors are correlated (oblique) or uncorrelated (orthogonal), and which items load onto which factors.

A major (often untenable) assumption in CFA is the notion of *simple structure*, that observed variables load onto one factor and one factor only (i.e., no cross-loadings; Figure 1b). CFA is referred to as “confirmatory” because model testing begins by assuming that a proposed factor structure is correct and then uses the data to examine whether that model provides an accurate summary of the data. To that end, we use global goodness-of-fit assessments (often referred to as fit indices or fit statistics) to conduct hypothesis tests (McNeish & Wolf, 2021). Fit indices depict the degree of alignment between the observed data and the hypothesized model. The better the model fits, the better the model reconstructs the data.

Researchers commonly use CFA to confirm whether a set of PD features conform to a unidimensional structure, or whether a 1-factor solution fits well (Ex. 1). As we might suspect from our EFA example, a 1-factorm unidimensional solution for borderline PD symptoms fits the data relatively well but not “good” in terms of conventional benchmarks, suggesting that a multidimensional structure might better describe the data (Table S2). If these symptoms’ dimensionality was of specific interest, we could further test for departures from unidimensionality using indices like explained common variance (Ex. 1; e.g., Rodriguez et al., 2016). Also, we might use CFA to examine the interrelatedness of a set of PD features (Ex. 2). Here, we examined a 3-factor solution of borderline, paranoid, and schizoid PD symptoms, given that reality testing in borderline PD is associated with psychotic-like experiences (e.g., Lenzenweger et al., 2001). This model fits the data well, although high factor intercorrelations (rs = .72-.75) suggest poor discrimination among them (Table S3).

Exploratory structural equation modeling.

Exploratory structural equation modeling (ESEM) was developed to address issues with overly restrictive CFA models, namely that simple structure is rarely observed in practice (Marsh et al., 2014). Instead, it is common for variables to cross-load such that they index both their intended construct as well as another (or multiple others; Morin et al., 2016). Because cross-loadings are typically forced to be zero in CFA, factor correlations are likely inflated and models tend to fit relatively poorly. Thus, even if the data resemble the hypothesized structure, a researcher might conclude that their conceptual model is refuted based on its poor fit. Examples of this problem are factor analyses of Big Five personality inventories, whose general theoretical structures are supported but fit is improved in an ESEM framework (compared with CFA; Marsh et al., 2010).

A defining characteristic of ESEM is the use of target rotation. In contrast with EFA, where the patterning of loadings is not specified, ESEM allows the researcher to specify the number of factors and marker variables for one or more of them. That is, the researcher selects variables that are thought to reflect “pure” indicators of the factor and the solution is rotated with that specification in mind, meaning that the program attempts to find a rotation that maximizes loadings from a pre-defined set of indicators and minimizes cross-loadings for those items onto other factors. Thus, ESEM reflects a balance between EFA and CFA because it allows tests of a hypothesized CFA structure that allows for cross-loadings (Figure 1c).

The issue of simple structure is particularly pertinent to PD research because the diagnostic criteria for many DSM PDs are themselves multidimensional, and there is a nontrivial amount of content overlap across PD categories. For instance, lack of close friends is a criterion of schizoid and schizotypal PDs, and anger/hostility is featured in paranoid, antisocial, and borderline PDs. Thus, an item assessing lack of close friends within the context of schizoid PD is likely to cross-load onto other PD-specific factors in a CFA. Indeed, PD features cross-load in metastructural models of PD symptoms or diagnoses (e.g., Forbes et al., 2017). Also, even traits that cohere to a strong 1-factor solution may cross-load onto other factors because they inherently blend multiple domains. For example, suspiciousness captured in paranoid PD (among others) reflects a blend of antagonism and detachment (e.g., Krueger et al., 2012).

With ESEM, we can examine whether the proposed 3-factor CFA of borderline, paranoid, and schizoid PD symptoms is tenable, in the sense that it is well supported in more exploratory analyses (see also Greene et al., 2022; Watts, Boness, et al., 2021), and whether model fit improves after relaxing the assumption of simple structure (Ex. 2). This model fits the data worse than the CFA solution (Table S2), and schizoid PD does not form a factor independent of borderline and paranoid PDs (Table S3). Rather, borderline PD symptoms make up a relatively independent factor, and paranoid and schizoid PD symptoms appear to form a factor. A third factor contains cross-loadings from identity disturbance symptoms of borderline PD and cross-loadings from schizoid PD symptoms. Because of the third factor’s limited interpretability, we might think of this ESEM solution as over-extracted, but the extent of cross-loadings highlights that simple structure is overly restrictive. Likewise, we see major reductions in factor intercorrelations (rs ranged from .23 to .68). All told, it is

relatively clear from this ESEM that the firm distinction between paranoid and schizoid PDs is not empirically supported (Mittal et al., 2007), and neither is the assumption of simple structure. Further, the fact that a target rotation did not support our CFA structure further reinforces the value of using more exploratory analyses to diagnose CFA solutions: We can extract 3 factors for each PD, but the data do not necessarily naturally conform to such a structure (Greene et al., 2022; Watts et al., 2020).

The bassackwards method.

Goldberg's (2006) bassackwards method can facilitate exploration of multiple levels of a hierarchy, which is useful given that the field has increasingly conceptualized personality, PDs, and other forms of psychopathology as hierarchically organized in nature (e.g., Kotov et al., 2017). Further, we anticipated that a hierarchical structure might be tenable in these data given that the results from parallel analysis, MAP, and VSS indicated different solutions. Solutions of varying numbers of factors are accommodated in hierarchical frameworks, like those derived from bassackwards methods.

The bassackwards method uses either principal components analysis or EFA with orthogonal or oblique rotation to extract a single component or factor at the highest level of the hierarchy, adding one more component or factor at each subsequent level (i.e., two on the second level, three on the third). Correlations between sequential levels of the structure characterize the hierarchical relationships among the components or factors. Forbes (2020) proposed several modifications to the bassackwards method to provide a simpler yet more complete picture of the hierarchical structure by examining the correlations between *all* levels of the structure, rather than only sequential levels.

Because our data are clearly multidimensional but there were not clear delineations between borderline, paranoid, and schizoid PD features, we used a bassackwards approach to explore the hierarchical structure of these features (see also Crowe et al., 2019, for an application to narcissism). The bassackwards analysis (Ex. 3) revealed a five-component model (Figure 1d).[iii] By default, a single component sits at the apex of the hierarchy. The first component split into (1) Detachment, which was dominated by schizoid PD symptoms but represented by subsets of paranoid PD (i.e., unforgiving, suspiciousness) and borderline PD (i.e., identity disturbance) symptoms, (2) and all other items. At the next level, dimensions reflecting tendencies toward Interpersonal Conflict, dominated by paranoid PD symptoms, and a block of borderline PD symptoms emerged. In turn, borderline PD split into Identity Disturbance and Volatility. Finally, at the lowest level, an Isolation component related to both Interpersonal Conflict and Detachment emerged. This analysis yields a conceptually rich hierarchical structure that broadly corresponds to the CFA and ESEM results. Some components corresponded to DSM-defined syndromes, whereas others reflected transdiagnostic tendencies.

[iii]Parallel analysis indicated up to 6 components, but the sixth component was Incoherent and not properly identified. We focused on orthogonal (varimax) principal components, but it is also possible to use oblique rotations and/or EFA. Goldberg (2006) highlighted several major advantages of PCA over EFA, which include computational economy and avoidance of negative residual variances (Heywood cases). These advantages are especially important with complex hierarchical structures that have many levels and components.

The exploratory-confirmatory continuum of factor analysis.

EFA, CFA, and ESEM occupy different regions of an exploratory-confirmatory continuum, with EFA and CFA occupying opposite poles and ESEM sitting somewhere in between them (Figure S2; Sass & Schmitt, 2020; Wright, 2017). Nevertheless, the extent to which EFA and CFA are sufficiently exploratory and confirmatory, respectively, depends on the way they are used. For instance, the modal EFA surrounds using one factor rotation and extraction method (e.g., promax with principal axis factoring). Realistically, however, researchers should use multiple factor rotation and extraction methods to examine whether factor structures are robust to the chosen method (Greene et al., 2022).

Additionally, researchers could use multiple random starts to ensure that their factor solution replicates. Factor analysis (and many other forms of quantitative methods) is not guaranteed to produce a single globally optimal solution, which means that the resulting model on display may not be the most interpretable or the only one that most closely describes the data. It is often computationally inefficient to estimate all possible solutions, so most statistical software considers a limited set of options when estimating the model parameters and displays one solution. Due to focusing on a more limited set of options, multiple locally (vs. globally) optimal solutions can result. Prior work suggests that orthogonal rotations and certain rotation criteria, such as geomin, are especially suspectable to the local optima problem (Greene et al., 2022; Mansolf & Reise, 2017). To increase the likelihood of settling on the best solution, researchers are encouraged to inspect factor loading patterns for all available optima (often fewer than 10 solutions) to see if there is agreement. When there is agreement across local optima, the researcher can feel more confident about their solution (see Nguyen & Waller, 2022, for a more detailed discussion).

As with EFA, CFA solutions require vetting because the implementation of CFA is often far less “confirmatory” than is implied by the name. As we mentioned earlier, the researcher is required to specify the number of factors included in the model, whether factors are correlated, and which variables indicate which factors. At the same time, it is rare for researchers to specify characteristics like the degree of factor loadings and factor intercorrelations in a CFA framework. Thus, the modal CFA specification is arguably fairly agnostic with respect to *a priori* hypotheses regarding the observed model, particularly aspects of the model that have the potential to adjudicate between competing factor explanations and theories. Moreover, model modification indices, which provide the improvement in the chi-squared fit statistic if the parameter were to be modeled, are often used to improve model fit. Nevertheless, the inclusion of model modifications (e.g., dropping items with low factor loadings), to improve fit are inherently post hoc. Thus, CFA solutions veer further and further towards the exploratory end of the exploratory-confirmatory continuum as a function of the number of modifications included.

In fact, in many circumstances, chasing model fit in CFA is arguably akin to *p*-hacking, or making methodological decisions to achieve statistically significant results. Within the context of conventional fit benchmarks, attaining “good” fit is analogous to a hypothesis test (McNeish & Wolf, 2021). If researchers rely exclusively on model fit to guide model selection, in absence of strong a priori rationale (and theory) for model adoption, they run the risk of hindsight bias, whereby they overestimate the likelihood of predicting a finding

that was not hypothesized *a priori* once it becomes clear that the modification improves fit. Of course, model misfit can indicate serious problems with model specification, so it should not be ignored completely. Likewise, some degree of data-driven decision-making is appropriate, particularly when unexpected features of the data (e.g., nonnormality, influential cases) render a hypothesized model inappropriate. The same is true for using EFA as a guide for future CFA specifications, or as a tool for probing problematic CFA results. In either case, EFA results should be presented to prevent the questionable research practice of framing CFA specifications as though they were hypothesis driven when they were actually guided by EFA results (i.e., the factor analytic analogue to hypothesizing after results are known or HARKing; Crede & Harms, 2019). Finally, researchers should aim for transparency in balancing confirmatory and data-driven decision-making, which may include preregistering their analyses so that the distinction between *a priori* and data-driven decisions is explicit.

Popular Types of Factor Models for PD Research

The unidimensional model posits that a single factor is responsible for the shared variance among a variable set (e.g., a 1-factor model of BPD symptoms). In contrast, the correlated factors model (Figure 1c) contains two or more distinct but related factors that summarize the shared variance among their variables, such as in our 3-factor CFA of borderline, paranoid, and schizoid PD symptoms. Because borderline, paranoid, and schizoid PD factors are highly correlated in these data, we might suspect that they share features *and* contain features that distinguish them, consistent with a hierarchical organization of PD symptoms. Such a hierarchical structure could be modeled with the bifactor model (Figure 1e), the higher-order model (Figure 1f), or, as we showed, with the bassackwards method (Figure 1d).

The bifactor model decomposes covariation among variables into two types of factors, general and specific (Figure 1e). General factors reflect a single source of covariation among all variables included in the model. Any remaining covariation among subsets of variables is captured with specific factors. To properly identify the model, the general factor is constrained to be uncorrelated with the specific factors. Traditionally, specific factors are constrained to be uncorrelated with one another, under the assumption that the general factor is the sole source of covariance among variables that load on different specific factors, but this constraint is not necessary to identify the model and is often conceptually untenable. In higher-order models, covariation among variables is captured with first-order factors, whose covariation is in turn captured with one or more superordinate second-order factors (Figure 1f). Like specific factors in a bifactor model, first-order factors represent residuals such that their contents reflect the shared variance of their constituent items net of the variance captured by the second-order factor(s).

What Factor Analysis Is Not

Factor analysis is often misinterpreted or confused with other methods. Now that we have defined what common forms of factor analysis are, we turn to describe what factor analysis is *not*.

1. Factor analysis is not a test of the common cause model.

Factor analysis is a summary of the covariance structure of variables between (or across) people. Historically, factor analysis was conceived of as a test of the common cause model. According to this model, a latent factor gives rise to (or causes) the shared variance of its items (Bollen, 1989; e.g., general intelligence is thought to cause correlations among intelligence tasks, [Spearman, 1904]). But the act of specifying a well-fitting factor model is an inadequate test of the common cause model. Instead, latent variables simply summarize the covariance of their items, without regard for the processes that explain the covariation. Simulation studies show that a 1-factor model can fit data produced by a single latent variable, multiple independent or correlated latent variables, and/or multiple causal loops among indicators (Hayduk, 2014; van Bork et al., 2021). Thus, it is possible that a single mechanism causes the covariance among a set of items, but it is also possible that items covary due to dynamic interactions among the items (consistent with a network conceptualization; Borsboom, 2017), that multiple mechanisms are responsible for different subsets of covariance among items, and so on (see "Do: Consider alternative explanations for your factors" for a more thorough discussion). For example, if impairments in empathy and intimacy reflect a single factor, it is also plausible (1) that these impairments are caused by a shared interpersonal dysfunction construct, (2) that empathy impairments cause intimacy problems, (3) that intimacy impairments cause empathy problems, or (4) that intimacy and empathy impairments reciprocally cause problems in one another.

2. Factor analysis is not a way to model personality "types" or "subtypes."

Factor analysis is a variable-centered, not person-centered, analysis. Variable-centered analyses examine relations among variables, whereas person-centered approaches identify subgroups of people based on their similarities on a set of variables (Milligan & Cooper, 1987).[iv] Thus, factor analysis is not a direct test of typologies. For instance, there is a storied history of various psychopathy "subtypes," in which primary psychopaths are thought to exhibit extreme lack of conscientiousness and empathy, and secondary psychopaths are thought to exhibit impulsivity and emotional distress (Hicks & Drislane, 2018). Factor analyses of various psychopathy measures do tend to identify at least two dimensions that broadly differentiate callousness (Factor 1) and impulsivity (Factor 2), but they cannot be interpreted as evidence of subtypes. That is, if we find that Factor 1 is associated with proactive aggression, it does not mean that *primary psychopaths* are more likely to engage in proactive aggression. More appropriate methods for extracting subtypes are cluster analysis or latent profile analysis (see Wright & Zimmerman, 2015, for an application to PDs).

3. Factor analysis is not necessarily a sufficient representation of within-person processes.

Factor analyses of cross-sectional data (e.g., PD symptoms from a clinical interview) do not tell us about the dynamic, within-person processes at the heart of most clinical theories, such as how and why our constructs (e.g., PD features) co-occur across situations and over

[iv]That said, cluster analysis can group observations (rather than variables) in terms of their similarity on a given set of variables, in turn revealing groupings of people that are alike on a given set of variables (see Milligan & Cooper, 1987). Approaches like agglomerative or divisive hierarchical clustering, for instance, can reveal fine-grained hierarchical structures of PD symptoms.

time. Moreover, if your construct of interest reflects dynamic processes, factor analyses may oversimplify (and even obscure detection of critical components of) such processes. There are, however, factor analytic approaches that characterize within-person processes, either idiographic (for a specific individual), nomothetic (for a population), or both. Such approaches include p-technique factor analysis (idiographic), multilevel structural equation modeling (nomothetic), and group iterative multiple model estimation (nomothetic and idiographic; see Wright et al., 2015, for an application to PD ratings).

Factor Analysis: Some Dos and Don'ts

Once a technique that took days if not weeks to implement, over the course of a century, factor analysis can now be completed in seconds. Notwithstanding the wonders of such a technological advance, factor analysis' ease of implementation is somewhat of a mixed blessing because its ease of implementation facilitates misapplication. In what follows, we provide some practical suggestions and considerations when conducting factor analysis in PD research. Because of the breadth of relevant considerations, we focus on issues that are especially salient to recent debates and novel methods put forth in the literature.

Do: Consider the limitations of model fit.

In recent years, the field has developed relative consensus around the limitations of using model fit statistics to guide model evaluation, comparison, and selection. In turn, we describe several limitations and their relevance to PD studies.

1. What constitutes "good" model fit depends on the model and data.— Conventional model fit benchmarks (or "cutoffs") are not one size fits all, meaning that they are influenced by numerous model characteristics, including your sample size, the number of indicators, the number of factors, as well as whether your indicators are categorical or continuous, have a non-linear association with the latent factors (latent factors assume linear relationships), or are skewed, among other characteristics of the data. In the case of the data used here, they were binary and positively skewed, sometimes heavily so. We took care to specify that our indicators are ordered categorical, and we used a robust estimator (weighted least squares mean and variance adjusted, or "WLSMV") because it is most appropriate for skewed, categorical data. Had we used an estimator that had not considered these characteristics (such as maximum likelihood or weighted least squares without means and variances adjusted), our models would have artificially lower fit (see Ex. 1 and Table S2, for an example with a 1-factor model of borderline PD symptoms). Thus, model fit is easily influenced by characteristics of the indicators and model, and it is critical to choose appropriate estimators. We direct interested readers to a broader, more thorough treatment of this topic, which we can address only briefly given the scope of this paper (Flora & Curran, 2004).

Additionally, model fit benchmarks must be catered to a model's specific characteristics (e.g., number of items, factors, latent factor reliability; Hu & Bentler, 1999; McNeish & Wolf, 2021). Additionally, model fit benchmarks do not generalize across different types of model misfit (e.g., latent variable relations, cross-loadings, correlated residuals; Heene et al., 2011; Marsh et al., 2004) and data characteristics (e.g., sample size, skew; Nye & Drasgow,

2011), such that any one source of misfit may have an undue influence on model fit. McNeish and Wolf (2021) recently developed dynamic fit indices (along with an R package [dynamic] and an interactive application to facilitate their use: https://dynamicfit.app) that produce benchmarks that are catered to the model and data at hand, generating fit indices that function as effect sizes that quantify degree of model misfit. To demonstrate, we computed dynamic fit indices for our correlated CFA model of borderline, schizoid, and schizotypal PD symptoms (Ex. 2). Although this model fit well per conventional fit benchmarks, dynamic fit index benchmarks were far more stringent (i.e., SRMR < .024, RMSEA < .013, CFI > .994). Thus, when model fit benchmarks are catered to our specific model, we would no longer accept our model as a "well-fitting" one.

2. **Some models fit better than others, regardless of the data.—**In confirmatory modeling scenarios, many studies estimate and compare increasingly complex measurement models (e.g., 1-factor, correlated factors, bifactor model). In these cases, model fit "contests" are of little use, because certain models are expected to provide superior fit, regardless of the input data. For instance, compared with correlated factors models, the bifactor model is expected to provide superior fit to *any* data (Greene et al., 2019, 2022, under review; Mansolf & Reise, 2017), despite relatively limited evidence of increased utility, structural fidelity, and replicability of its latent factors (e.g., Rodriguez et al., 2016; Watts et al., 2019, 2020). The bifactor model fits data better than other models because the correlated factor and higher-order models are often nested within the bifactor model, meaning that bifactor models can fit a wider range of data than the former. That is, the bifactor model is prone to fit well due to its general structure, so it is better at accommodating data with unmodeled complexities (e.g., cross-loadings, correlated residuals; Greene et al., 2022), which known as having high fitting propensity (Bonifay & Cai, 2017; Falk & Muthukrishna, 2021). Nevertheless, a model that best accommodates the data is not the same as a model that best characterizes the data-generating mechanism(s). This problem tends to be overlooked because people assume that fit indices penalize for this accommodational tendency, though relative fit indices penalize for parametric complexity (i.e., the number of freely estimated parameters), not configural complexity (i.e., the particular arrangement of variables in the model; Falk & Muthukrishna, 2021).

Complicating matters further, although goodness-of-fit is sometimes mistaken as an index of how well a model describes the data, the opposite can be true. As model fit increases, a model is *less likely* to approximate the true structure in the population due to overfitting (Bonifay, 2021). As a model fits increasingly well, it is more likely to capture sample-specific variation, even random noise (e.g., Reise et al., 2016), in turn reducing its generalizability (Bonifay, 2021; Greene et al., under review). That is, when a researcher selects a model that is especially prone to overfitting data over a worse fitting but equally viable alternative model, they run the risk of selecting the model that is less likely to replicate in another study (e.g., Watts, Lane, et al., 2020).[V] Thus, somewhat counterintuitively, the best-fitting model is less likely to be the truest, or even most useful, model.

[V]This problem reflects the broader tension between bias and variance (i.e., the bias-variance tradeoff) in the machine learning and structural equation modeling literatures.

3. Model fit statistics are useless when comparing equivalent models.— Equivalent models are those that represent different theoretical representations of the data, but share identical fit index and chi-square statistics, model-implied covariance matrices, and residual matrices (Hershberger & Marcoulides, 2013). Salient examples include that: a correlated three-factor model can be re-expressed as a fit-equivalent higher-order model with three lower-order factors (Table S2; Forbes et al., 2020)[vi]; an exploratory bifactor model can be equivalent to the bass-ackwards model (Ringwald et al., 2019); a 1-factor EFA/CFA can be equivalent to a network model of the same items (van Bork et al., 2021); and a 3-factor EFA and a 3-factor ESEM fit equally, despite the fact that they can have entirely different patterns of factor loadings and intercorrelations (Marsh et al., 2014). In fact, all EFA models, including bifactor and non-bifactor rotations, with the same number of factors will be fit-equivalent to one another, regardless of rotation orientation (e.g., orthogonal, oblique) and criterion (e.g., geomin, oblimin; Greene et al., 2022; Marsh et al., 2014).[vii]

Ultimately, model equivalence prevents researchers from using fit statistics to adjudicate between models with very different observed structures, and models that differ in terms of how theory-based they are (e.g., EFA vs. ESEM with target rotation). Though the issue of model equivalence may sound trivial, it has serious implications for how we interpret our models and how well they correspond with our hypotheses or theories. For instance, a 3-factor ESEM that more directly corresponds to one's working theoretical model fits equally as well as a model that may not conform to the theory at all. Also, we might conclude that there is a robust general factor of PD because we simply extracted such a dimension in a bifactor or bassackwards model even when these models are statistically equivalent to those that do not directly model a general factor (see also Watts et al., 2020). For any given dataset, there are sure to be a number of well-fitting, equivalent models, and multiple alternatives should be explored (Tomarken & Waller, 2003).

Do: Report factor reliability.

To address problems with relying on model fit indices alone when adjudicating model quality, a growing literature has emphasized the importance of also reporting model-based reliability coefficients. A variety of resources describe the estimation and interpretation of model-based reliability indices that vary in their relevance to models used in PD research (e.g., Brunner et al., 2012; Rodriguez et al., 2016). The omega family of indices (e.g., omega hierarchical, omega total) estimate the proportion of variance accounted for by each latent factor as an index of reliability, and explained common variance (ECV) represents one index to characterize the data's dimensionality to determine whether a unidimensional model is sufficient, or if additional factors are useful.

[vi]A higher-order model with three lower-order factors is just-identified, meaning that the number of free parameters is equivalent to the number of known parameters. Thus, this specific higher-order model is equivalent to a 3 correlated factors model, with the factor correlations from the latter being reexpressed as lower-order factors' loadings onto the higher-order factor. That is, factor loadings onto the higher-order factor are recapitulations of factor intercorrelations from a 3 correlated factors model. A higher-order model with four or more lower-order factors is identified and, thus, distinguishable from a correlated factors model.

[vii]This scenario further illustrates the well-known issue of rotational indeterminacy in exploratory modeling scenarios, which produces multiple equivalent factor structures (Greene et al., 2022), and is why the use of fit statistics to guide model selection in EFA is generally discouraged (Montoya & Edwards, 2020).

Model fit and model-based reliability indices do not always agree, which reflects a unique challenge known as the “reliability paradox” (McNeish et al., 2018): Many well-fitting models contain factors with poor reliability, and vice versa. Simply calculating and interpreting reliability indices has several important benefits, all under the umbrella of diagnosing a model’s appropriateness. First, some reliability indices facilitate plain language interpretations of factor strength (e.g., a factor explains only 10% of the variance in its indicators) and can help reveal weak (noisy) factors. Second, reliability indices contextualize global goodness-of-fit. A model with weak factors may fit well due to being heavily parameterized. Specifying greater numbers of factors with small factor loadings can often contribute to close model fit, but low factor reliability indicates that weak factors do not explain much variance in the indicator(s). When small to moderate loadings are also unreliable (e.g., with large standard errors), we are likely overfitting our data, leaving us with little useful information from our estimated model (Forbes et al., 2021).

Do: Be aware of assumptions your model is making.

As with any statistical method, factor analysis makes a variety of assumptions, some of which are more consequential than they first appear.

1. **Rotation methods.—**When using EFA, the first choice researchers have to make is selecting between oblique and orthogonal rotation methods, with most choosing an oblique rotation that allows for correlated as opposed to uncorrelated factors. From there, we can choose any number of factor rotations, though many researchers rely on the default method for their given program (e.g., oblimin for the R psych package, geomin for Mplus) or previously reported methods in the literature. Though the differences among rotation methods may seem trivial, which a researcher chooses can lead to different conclusions regarding structure of the data because each has slightly different sets of assumptions (see also Sass & Schmitt, 2010).

   To illustrate how conclusions can vary as a function of rotation method, we extracted 3-factor EFAs of schizotypal PD symptoms with five different rotation methods (i.e., geomin, oblimin, promax, simplimax, BentlerQ; Ex. 5). Factor congruence was strong (Φs > .99 to 1.00) among geomin, oblimin, promax, but simplimax tended to produce discrepant results. Simplimax produced a relatively strong general factor and two other weak factors, whereas the others produced dimensions for (a) positive symptoms, (b) suspiciousness, (c) supernatural experiences. Thus, considering consensus across several rotation methods may help ensure that the chosen model is capturing consistent patterns in the data.

2. **Local independence.—**In classical test theory, the concept of local independence pertains to the assumption that all items are related to one another by way of the latent variable, meaning that items should be uncorrelated with each other after accounting for the latent factor. This assumption goes hand in hand with the common cause assumption: what is responsible for the interrelatedness of the items is that latent factor and the latent factor only. Unfortunately, this is assumption is often unmet.

   To confirm that our 1-factor model of borderline PD symptoms did not violate the assumption of local independence, we examined modification indices (Ex. 1). Fit would

improve substantially if we allowed 3 items pertaining to suicidality to correlate with one another, and 2 items pertaining to identity disturbance to correlate. These residual correlations make conceptual sense, suggest that we have violated the assumption of local independence, and imply that an alternative, multidimensional model may be more appropriate.

Nevertheless, often, researchers will retain and even prioritize a 1-factor solution because it is compatible with their theory or goals (Watts, Boness, et al., 2021). In the case of modeling borderline PD symptoms, if a researcher is interested in borderline PD *per se*, adopting a 1-factor model may seem most appropriate, but the researcher runs the risk of creating several issues. First, if there is a robust multidimensional solution in the data, the researcher may not be studying the most meaningful level of abstraction. In construct validation efforts, the researcher may make a conclusion about borderline PD's correlates that may be specific or unique to a narrower set of items in the model (e.g., Vainik et al., 2015). Global borderline PD may be associated with increased likelihood of later divorce or relationship dissolution in a prospective examination, but that association may be driven largely by a specific subset of indicators pertaining to relationship problems (e.g., having relationships with lots of extreme ups and downs). If that were the case, one could also argue that the effect is largely driven by criterion contamination (i.e., relationship problems are being used to predict relationship problems).

Second, if aspects of borderline PD captured in multiple factors relate differentially with an external criterion, the associations for borderline PD *per se* may either water down the effects of certain subdimensions, or even approach zero in cases where subdimensions have associations with opposing signs (e.g., moderate positive and moderate negative). Third, if violations of local independence are not accounted for in the model, the meaning of the latent factor will be biased *towards* the source of the local dependence. If the source of local independence reflects method variance (such as item-keying), it is possible that the meaning of the latent factor is biased towards construct-irrelevant variance, including careless responding (e.g., Schmitt & Stuits, 1985).

Do: Consider alternative models.

Because factor models simply summarize the covariance structure of the data, they are relatively agnostic with respect to the data generating mechanisms (see also "*Model fit statistics are useless when adjudicating between equivalent models*" and "Do: Consider alternative explanations for your factors"). Within the realm of factor models, it is increasingly *uncommon* for researchers to test a thorough set of alternative factor models, which raises the possibility that our favored models are promoted due to confirmation bias.

One possible explanation for the field's decreased focus on testing a wide range of possible models is that many researchers place too much stock in the "C" in CFA. Consider a bifactor model with borderline, paranoid, and schizoid specific factors (Figure 1e, Table S4). Although the model fits well, other aspects of the model are worrisome. The borderline specific factor is either unrepresented or negatively indicated by most of its indicators; as such, it is highly unreliable and explains only 2% of the variance in its indicators. As a riskier test of the CFA model, we could apply bifactor EFA to see whether each specific

factor is retained in an exploratory context. Alas, a bifactor EFA does not reveal strong PD-specific factors (Ex. 4, Table S4). The model contained a strong general factor and weak specific factors reflecting: Identity Disturbance, Self-harm, and Detachment. With more thorough tests of alternative models, we are less likely to confuse good CFA model fit with good theoretical support (see also Watts et al., 2020, 2021).

Do: Remember that measurement affects your output.

The results of any model, including factor models, are highly influenced by their contents. If content related to a narrow construct is overrepresented relative to other indicators of a broader construct, an artifactual or "bloated specific" factor may emerge. Likewise, if content is underrepresented, a "true" specific factor may not emerge (Watts, Boness, et al., 2021). For instance, there are conflicting conclusions surrounding whether disinhibition and anankastia (rigid perfectionism) are separate domains or are opposite ends of a maladaptive conscientiousness dimension. Oltmanns and Widiger (2018) showed that when two rigid perfectionism items were added to an EFA of Big Five traits, the items formed a coherent factor with conscientiousness items. When six to ten rigid perfectionism items were added, they formed an independent factor. It is possible that the independent factor reflects a bloated specific factor, but it is also possible that the failure to recover a separate anankastia factor in much of the literature is due to lack of content coverage in prevailing maladaptive trait measures (e.g., PID-5; Krueger et al., 2012).

These competing possibilities highlight the need to balance measurement and theory. If the goal is to adequately assess a construct that is deemed highly relevant to PDs, then covering it thoroughly seems ideal (from the perspective of item response theory, this is compatible with the notion of covering a wide range of theta). At the same time, it is difficult to draw firm conclusions regarding the nature of a construct based on measurement coverage alone, particularly if you have prioritized measurement considerations that nearly guarantee support for your theory. For instance, if a researcher added 10 items assessing a preference for chocolate over vanilla to a PD inventory, they might find that the items form a homogenous independent factor, but this does not necessarily mean that chocolate preference is a robust dimension relevant to personality pathology. Moreover, inclusion of psychometrically redundant items can artificially boost factor reliability, at the cost of (1) representing a broad range of a construct and (2) misleading researchers into assuming their items reflect a latent construct. Unfortunately, no quantitative methods clearly and comprehensively determine the extent to which the effects of content representation on factor structure are artifactual or substantive, so researchers must again rely on theory to defend their inferences. Finally, the structure of our models is *highly sensitive* to the type of input, such as whether we are modeling diagnoses, symptom counts, or individual symptoms/traits. Dichotomous diagnoses are associated with attenuated reliability (Markon et al., 2011), and diagnostic covariance can be distorted by low base rates (often due to quasi-artificial diagnostic thresholds). Using symptom counts as opposed to dichotomous diagnoses lessens these concerns (Wright & Simms, 2015), but because PDs are plagued by within-category heterogeneity, modeling symptom counts still undermines the utility and interpretability of PD models (Williams et al., 2017).

Do: Consider alternative explanations for your factors.

Even though researchers often presume that a latent factor reflects their intended, substantive construct of interest (the reification fallacy), they can reflect construct-irrelevant variance from a variety of sources. First, as we mentioned just now, unmodeled sources of local dependence can bias the meaning of a factor. A major source of local dependence is shared method variance, that indicators from the same instrument covary due to construct-irrelevant attributes of the measure (e.g., item-keying). Unidimensional measures can produce 2-factor solutions, with different factors for positively and negatively keyed items, but these factors should not necessarily be interpreted as reflecting substantively different constructs. Also, in monomethod analyses of personality trait inventories, for instance, the consistent finding of a general factor of personality led many researchers to conclude that there is a *substantive trait* that explains overall personality functioning (Rushton & Irwing, 2011). Nevertheless, general factors across different personality inventories are only modestly correlated, and a general factor does not emerge when the inventories are analyzed together, suggesting the general factor of personality largely captures measure-specific idiosyncrasies (Hopwood et al., 2011).

Second, variables can correlate due to attributes of the rater (e.g., self-report, informant-report, clinical interviewer). One source of shared rater variance is evaluative consistency, which refers to the tendency to rate oneself or others more positively or negatively across a wide range of characteristics. The influence of evaluative consistency on a factor's meaning may be particularly relevant for PD research given the strong valence (i.e., relative social undesirability) of its constructs. Once again, evidence for the general factor of personality is much weaker when personality items are reworded to be less valenced (Bäckström et al., 2009), and fails to emerge entirely after accounting for rater-specific variance (Chang et al., 2012). Multi-method/rater study designs that directly model rater-specific variance are necessary to disentangle shared measure or rater variance from the intended construct (Campbell & Fiske, 1959).

Third, variables can correlate because they have similar distributions in a sample, resulting in a factor that reflects severity of impairment rather than (or in addition to) a cohesive construct (Bernstein & Teng, 1989). There is an especially high likelihood of recovering a "difficulty factor" in studies of psychopathology that comprise participants with relatively low degrees of psychopathology, because the negative skew of responses will be magnified. For example, in an EFA of 25 maladaptive trait scales, Ringwald and colleagues (2021) found a Psychoticism factor primarily marked by Unusual Beliefs, Unusual Experiences, and—unexpectedly—Self-Harm. Self-harm probably covaried with the other indicators because they had the lowest endorsement rates (i.e., highest severities/difficulties), not necessarily because self-harming tendencies are a core feature of Psychoticism. Of course, it is possible that some response process led to co-endorsement of both self-harm and psychosis. We raise this example as a potential illustration of spurious factors that can arise when subsets of indicators have similar distributions (see Bernstein & Teng, 1989, for more discussion).

Do: Develop construct validity.

Due to the combined limitations of model fit in formulating valid inferences and the fact that a latent factor can take on numerous meanings, researchers must rely on construct validation to inform the potential utility and meaning of a latent factor. Construct validity is never established definitively; rather, validation is an iterative process that involves articulation of *theoretically derived hypotheses* about how a construct should manifest, known as a "nomological net" (Cronbach & Meehl, 1955). Results from these hypothesis tests are then used to update theory. This validation process therefore elevates model interpretation from mere speculation to inference (Strauss & Smith, 2009).

Although construct validation requires multiple steps, a particularly important component is external validity, the practice of examining a factor's relations with external correlates. Often, researchers simply correlate their latent factors (e.g., impulsive psychopathy features) with measures of other, presumably relevant constructs (e.g., substance use, number of arrests, aggression) and conclude that their factor has real, psychological meaning, and that the factor is valid if the correlations are significant (at $\alpha = .05$) in their sample. This practice does not reflect the intention behind construct validity. Rather, researchers should articulate a nomological net clearly and in advance of analysis (e.g., Brandes et al., 2021).

Even the act of labeling a factor constitutes an implicit theory of a construct, but this theory becomes falsifiable only if it is elaborated concretely and only when external validators are carefully selected (Bollen, 1989). And our validity tests should be "risky" ones (Meehl, 1978): *What ought a factor correlate with if Theory 1, but not Theory 2, is supported? What validator would place Theory 1 at strong risk?* Thus, researchers must identify validators that our factor *should* correlate with (convergent validity), and validators that our factor *should not* correlate with (discriminant validity) if it indeed reflects our construct of interest. Better yet, a stronger test would articulate an expected range of effect sizes (e.g., Funder & Ozer, 2019). If either of these tests fail, the theory requires revisiting.

Finally, strong evidence of external validity requires multiple forms of evidence, including predictive validity (e.g., antagonism in young adulthood should predict divorce in midlife), known-groups validity (e.g., symptom means should differ between clinical and community samples), temporal validity (e.g., pathological states should fluctuate more than pathological traits over repeated measurements), among other forms (Clark & Watson, 2019; Flake et al., 2017). Still, even if a strong nomological network is established, our items may assess something other than the intended construct (Embretson, 1983).

Conclusion

Factor analysis is an immensely useful, but flexible and often misunderstood statistical technique. In contrast with more basic and technical primers of factor analysis, we focused here on modern applications, limitations, and misinterpretations of factor analysis with the aim of raising awareness of common issues and points of contention in the field and updating recommendations for best practice. One key theme was clarifying what factor analysis is and is not, to emphasize how essential it is to *match our models and methods to our theories* (Fried, 2020). In particular, it is essential to remember that factor analysis is not

an adequate test of the common cause model or of the dynamic unfolding of PD features within individuals over time. By acknowledging these limitations, we can better cater our conclusions to the statistical model we tested, and even determine whether factor analysis bears on our theory at all.

Also, we described several methods for interrogating the meaning of our latent factors and stressed the importance of recognizing that our latent factors can reflect construct-irrelevant variance. Awareness of these issues is expected to help researchers avoid the common reification fallacy, assuming that our latent factors reflect our causal construct of interest, without further inquiry into the factor's meaning. We also highlighted the value in approaching statistical modeling using varied methods to converge on shared, and thus robust, conclusions and to minimize the likelihood of reporting spurious effects. Along the way, we argued (1) that model evaluation should be approached with the knowledge that model results are highly sensitive to data and sampling characteristics, (2) that model selection is a process that should be conducted with *multiple* characteristics in mind (e.g., factor reliability, construct validity), rather than model fit alone, and (3) that exploratory analysis can and should be used to bear on the validity of confirmatory models.

Finally, the overarching principle of this paper is the importance of subjecting our methods and models to riskier tests (Meehl, 1978). With the concept of riskier tests in mind, we can ask ourselves: Does this model provide a *strong test* of my theory? What methods and criteria can I use to provide *stronger support* for my theory? And finally, what criteria can I use to *falsify* my theory? In our view, careful attention to these broad themes has the potential to markedly advance theory, research, and treatment surrounding the nature and impact of PDs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding sources:

ALW is funded through K99AA028306 (Principal Investigator: Watts). ALG is supported by the Office of Academic Affiliations, Advanced Fellowship Program in Mental Illness Research and Treatment, Department of Veterans Affairs. MKF is supported by a National Health and Medical Research Council Investigator Grant (APP1194292).

References

Allport GW, & Odbert HS (1936). Trait-names: A psycho-lexical study. Psychological Monographs, 47(1), i–171.

Bäckström M, Björklund F, & Larsson MR (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. Journal of Research in Personality, 43(3), 335–344.

Bernstein IH, & Teng G (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. Psychological Bulletin, 105(3), 467–477.

Bollen KA (1989). Structural equations with latent variables. Wiley.

Bonifay W (2021). Increasing generalizability via the principle of minimum description length [Preprint]. PsyArXiv.

Bonifay W, & Cai L (2017). On the complexity of item response theory models. Multivariate Behavioral Research, 52(4), 465–484. [PubMed: 28426237]

Borsboom D (2017). A network theory of mental disorders. World Psychiatry, 16(1), 5–13. [PubMed: 28127906]

Brandes CM, Reardon KW, Shields AN, & Tackett JL (2021). Towards construct validity of relational aggression: An examination of the Children's Social Behavior Scale. Psychological Assessment, 33(9), 855–870. [PubMed: 33956474]

Brunner M, Nagy G, & Wilhelm O (2012). A tutorial on hierarchically structured constructs. Journal of Personality, 80(4), 796–846. [PubMed: 22091867]

Campbell DT, & Fiske DW (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81–105. [PubMed: 13634291]

Chang L, Connelly BS, & Geeza AA (2012). Separating method factors and higher order traits of the Big Five: A meta-analytic multitrait–multimethod approach. Journal of Personality and Social Psychology, 102(2), 408–426. [PubMed: 21967007]

Clark DA, & Bowles RP (2018). Model fit and item factor analysis: Overfactoring, underfactoring, and a program to guide interpretation. Multivariate Behavioral Research, 53(4), 544–558. [PubMed: 29683723]

Clark LA, & Watson D (2019). Constructing validity: New developments in creating objective measuring instruments. Psychological Assessment, 31(12), 1412–1427. [PubMed: 30896212]

Conway CC, Hammen C, & Brennan PA (2016). Optimizing prediction of psychosocial and clinical outcomes with a transdiagnostic model of personality disorder. Journal of Personality Disorders, 30(4), 545–566. [PubMed: 26168327]

Crede M, & Harms P (2019). Questionable research practices when using confirmatory factor analysis. Journal of Managerial Psychology, 34(1), 18–30.

Cronbach LJ, & Meehl PE (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281–302. [PubMed: 13245896]

Crowe ML, Lynam DR, Campbell WK, & Miller JD (2019). Exploring the structure of narcissism: Toward an integrated solution. Journal of Personality, 87(6), 1151–1169. [PubMed: 30742713]

Embretson SE (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 179–197.

Falk CF, & Muthukrishna M (2021). Parsimony in model selection: Tools for assessing fit propensity. Psychological Methods.

Flora DB, & Curran PJ (2004). An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis With Ordinal Data. Psychological Methods, 9(4), 466–491. 10.1037/1082-989X.9.4.466 [PubMed: 15598100]

First MB, Spitzer RL, Gibbon M, & Williams JBW (1995). The Structured Clinical Interview for DSM-III-R Personality Disorders (SCID–II): Part I. Description. Journal of Personality Disorders, 9, 83–91.

Flake JK, Pek J, & Hehman E (2017). Construct validation in social and personality research: Current practice and recommendations. Social Psychological and Personality Science, 8(4), 370–378.

Forbes MK (2020). Improving hierarchical models of individual differences: An extension of Goldberg's bass-ackwards method [Preprint]. Open Science Framework.

Forbes MK, Greene AL, Levin-Aspenson H, Watts AL, Hallquist M, […] & Krueger RF (2020). A comparison of the reliability and validity of the predominant models used in research on the empirical structure of psychopathology. Journal of Abnormal Psychology.

Forbes MK, Kotov R, Ruggero CJ, Watson D, Zimmerman M, & Krueger RF (2017). Delineating the joint hierarchical structure of clinical and personality disorders in an outpatient psychiatric sample. Comprehensive Psychiatry, 79, 19–30. [PubMed: 28495022]

Fried EI (2020). Lack of theory building and testing impedes progress in the factor and network literature. Psychological Inquiry, 31(4), 271–288.

Funder DC, & Ozer DJ (2019). Evaluating effect size in psychological research: Sense and nonsense. Advances in Methods and Practices in Psychological Science, 2(2), 156–168.

Goldberg LR (2006). Doing it all Bass-Ackwards: The development of hierarchical factor structures from the top down. Journal of Research in Personality, 40(4), 347–358.

Greene AL, Eaton NR, Forbes MK, Fried EI, Watts AL, Kotov R, Krueger RF (under review). Model fit is a fallible indicator of model quality in quantitative psychopathology research: A reply to Bader and Moshagen. Journal of Psychopathology and Clinical Science.

Greene AL, Eaton NR, Li K, Forbes MK, Krueger RF, Markon KE, Waldman ID, […] & Kotov R (2019). Are fit indices used to test psychopathology structure biased? A simulation study. Journal of Abnormal Psychology, 128(7), 740–764. [PubMed: 31318246]

Greene AL, Watts AL, Forbes MK, Kotov R, Krueger RF, & Eaton NR (2022). Misbegotten methodologies and forgotten lessons from Tom Swift's Electric Factor Analysis Machine: A demonstration with competing structural models of psychopathology. Psychological Methods.

Hayduk L (2014). Seeing perfectly fitting factor models that are causally misspecified: Understanding that close-fitting models can be worse. Educational and Psychological Measurement, 74(6), 905–926.

Heene M, Hilbert S, Draxler C, Ziegler M, & Bühner M (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. Psychological Methods, 16(3), 319–336. [PubMed: 21843002]

Hershberger SL, & Marcoulides GA (2013). The problem of equivalent structural models. In Structural equation modeling: A second course, 2nd ed (pp. 3–39). IAP Information Age Publishing.

Hicks BM, & Drislane LE (2018). Variants ("subtypes") of psychopathy. In Handbook of psychopathy, 2nd ed (pp. 297–332). The Guilford Press.

Hopwood CJ, Wright AGC, & Brent Donnellan M (2011). Evaluating the evidence for the general factor of personality across multiple inventories. Journal of Research in Personality, 45(5), 468–478. [PubMed: 22879686]

Hu L, & Bentler PM (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.

Jöreskog KG (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34(2), 183–202.

Kotov R, Krueger RF, Watson D, Achenbach TM, Althoff RR, Bagby RM, … & Zimmerman M (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. Journal of Abnormal Psychology, 126(4), 454–477. [PubMed: 28333488]

Krueger RF, Derringer J, Markon KE, Watson D, & Skodol AE (2012). Initial construction of a maladaptive personality trait model and inventory for DSM-5. Psychological Medicine, 42(9), 1879–1890. [PubMed: 22153017]

Lenzenweger MF, Clarkin JF, Kernberg OF, & Foelsch PA (2001). The Inventory of Personality Organization: Psychometric properties, factorial composition, and criterion relations with affect, aggressive dyscontrol, psychosis proneness, and self-domains in a nonclinical sample. Psychological Assessment, 13(4), 577–591. [PubMed: 11793901]

Mansolf M, & Reise SP (2017). When and why the second-order and bifactor models are distinguishable. Intelligence, 61, 120–129.

Markon KE, Chmielewski M, & Miller CJ (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. Psychological Bulletin, 137(5), 856–879. [PubMed: 21574681]

Marsh HW, Hau K-T, & Wen Z (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. Structural Equation Modeling: A Multidisciplinary Journal, 11(3), 320–341.

Marsh HW, Lüdtke O, Muthén B, Asparouhov T, Morin AJS, Trautwein U, & Nagengast B (2010). A new look at the big five factor structure through exploratory structural equation modeling. Psychological Assessment, 22(3), 471–491. [PubMed: 20822261]

Marsh HW, Morin AJS, Parker PD, & Kaur G (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. Annual Review of Clinical Psychology, 10(1), 85–110.

McNeish D, An J, & Hancock GR (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. Journal of Personality Assessment, 100(1), 43–52. [PubMed: 28631976]

McNeish D, & Wolf MG (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. Psychological Methods.

Meehl PE (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46(4), 806–834.

Milligan GW, & Cooper MC (1987). Methodology review: Clustering methods. Applied Psychological Measurement, 11(4), 329–354.

Mittal VA, Kalus O, Bernstein DP, & Siever LJ (2007). Schizoid personality disorder. In Personality disorders: Toward the DSM-V. (pp. 63–79). Sage Publications, Inc.

Montoya AK, & Edwards MC (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. [PubMed: 33994558]

Morin AJS, Arens AK, Tran A, & Caci H (2016). Exploring sources of construct-relevant multidimensionality in psychiatric measurement: A tutorial and illustration using the Composite Scale of Morningness: Construct-Relevant Multidimensionality. International Journal of Methods in Psychiatric Research, 25(4), 277–288. [PubMed: 26265387]

Mulaik SA (1987). A brief history of the philosophical foundations of exploratory factor analysis. Multivariate Behavioral Research, 22(3), 267–305. [PubMed: 26776378]

Nguyen HV, & Waller NG (2022). Local minima and factor rotations in exploratory factor analysis. Psychological Methods.

Nye CD, & Drasgow F (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. Organizational Research Methods, 14(3), 548–570.

Preacher KJ, & MacCallum RC (2003). Repairing Tom Swift's electric factor analysis machine. Understanding Statistics, 2(1), 13–43.

Preacher KJ, Zhang G, Kim C, & Mels G (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. Multivariate Behavioral Research, 48(1), 28–56. [PubMed: 26789208]

Reise SP, Kim DS, Mansolf M, & Widaman KF (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg self-esteem scale. Multivariate Behavioral Research, 51(6), 818–838. [PubMed: 27834509]

Ringwald WR, Beeney JE, Pilkonis PA, & Wright AGC (2019). Comparing hierarchical models of personality pathology. Journal of Research in Personality, 81, 98–107. [PubMed: 31186592]

Ringwald WR, Emery L, Khoo S, Clark LA, Kotelnikova Y, Scalco MD, Watson D, Wright AGC, & Simms L (2021). Structure of Pathological Personality Traits through the Lens of the CAT-PD Model [Preprint]. PsyArXiv. 10.31234/osf.io/kuefm

Rodriguez A, Reise SP, & Haviland MG (2016). Applying bifactor statistical indices in the evaluation of psychological measures. Journal of Personality Assessment, 98(3), 223–237. [PubMed: 26514921]

Rushton P, & Irwing P (2011). The general factor of personality: Normal and abnormal. In The Wiley-Blackwell handbook of individual differences (pp. 132–161). Wiley Blackwell.

Sass DA, & Schmitt TA (2010). A comparative investigation of rotation criteria within exploratory factor analysis. Multivariate Behavioral Research, 45(1), 73–103. [PubMed: 26789085]

Schmitt N, & Stuits DM (1985). Factors defined by negatively keyed items: The result of careless respondents? Applied Psychological Measurement, 9(4), 367–373.

Sellbom M, & Tellegen A (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. Psychological Assessment, 31(12), 1428–1441. [PubMed: 31120298]

Sharp C, Wright AGC, Fowler JC, Frueh BC, Allen JG, Oldham J, & Clark LA (2015). The structure of personality pathology: Both general ('g') and specific ('s') factors? Journal of Abnormal Psychology, 124(2), 387–398. [PubMed: 25730515]

Spearman C (1904). “General intelligence,” objectively determined and measured. The American Journal of Psychology, 15(2), 201–293.

Strauss ME, & Smith GT (2009). Construct validity: Advances in theory and methodology. Annual Review of Clinical Psychology, 5(1), 1–25.

Tomarken AJ, & Waller NG (2003). Potential problems with “well fitting” models. Journal of Abnormal Psychology, 112(4), 578–598. [PubMed: 14674870]

Vainik U, Mõttus R, Allik J, Esko T, & Realo A (2015). Are trait–outcome associations caused by scales or particular items? Example analysis of personality facets and BMI. European Journal of Personality, 29(6), 622–634.

van Bork R, Rhemtulla M, Waldorp LJ, Kruis J, Rezvanifar S, & Borsboom D (2021). Latent variable models and networks: Statistical equivalence and testability. Multivariate Behavioral Research, 56(2), 175–198. [PubMed: 31617420]

Watts AL (2022, April 4). Factor analysis in personality disorders research: Modern issues and practical suggestions. Retrieved from osf.io/dyth3

Watts AL, Boness CL, Loeffelman JE, Steinley D, & Sher KJ (2021). Does crude measurement contribute to observed unidimensionality of psychological constructs? A demonstration with DSM–5 alcohol use disorder. Journal of Abnormal Psychology, 130(5), 512–524. [PubMed: 34472887]

Watts AL, Lane SP, Bonifay W, Steinley D, & Meyer FAC (2020). Building theories on top of, and not independent of, statistical models: The case of the p-factor. Psychological Inquiry, 31(4), 310–320. [PubMed: 33510565]

Watts AL, Poore HE, & Waldman ID (2019). Riskier tests of the validity of the bifactor model of psychopathology. Clinical Psychological Science, 7(6), 1285–1303.

Williams TF, Scalco MD, & Simms LJ (2018). The construct validity of general and specific dimensions of personality pathology. Psychological Medicine, 48(5), 834–848. [PubMed: 28826417]

Wright AGC (2017). The current state and future of factor analysis in personality disorder research. Personality Disorders: Theory, Research, and Treatment, 8(1), 14–25.

Wright AGC, & Zimmermann J (2015). At the nexus of science and practice: Answering basic clinical questions in personality disorder assessment and diagnosis with quantitative modeling techniques. In Huprich SK (Ed.), Personality disorders: Toward theoretical and empirical integration in diagnosis and assessment. (pp. 109–144). American Psychological Association.

Wright AGC, Beltz AM, Gates KM, Molenaar PCM, & Simms LJ (2015). Examining the dynamic structure of daily internalizing and externalizing behavior at multiple levels of analysis. Frontiers in Psychology, 6.

Wright AGC, & Simms LJ (2015). A metastructural model of mental disorders and pathological personality traits. Psychological Medicine, 45(11), 2309–2319. [PubMed: 25903065]

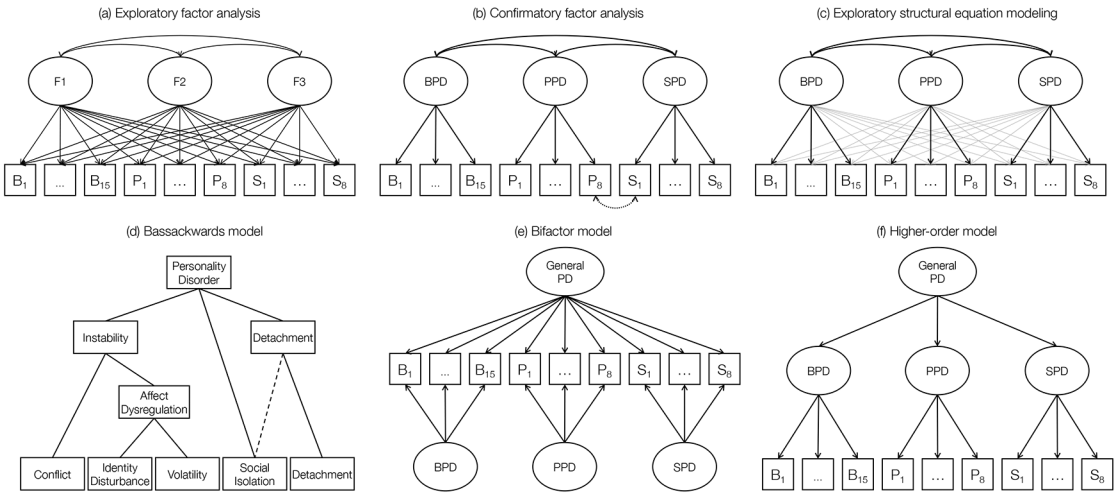

Figure 1. Common factor models used in the personality disorders literature.

Note. In latent variable modeling, the square boxes represent manifest variables, and the circles or ovals represent latent variables. Manifest variables are indicators of latent variables. The magnitude of the indicators' relations with various factors, depicted as a straight arrow from the latent factor to the indicator, are expressed as factor loadings, which are akin to a correlation coefficient. Curved arrows reflect correlations or covariances, either between latent factors or manifest indicators. We do not depict them here for the sake of simplicity, but straight arrows that point towards the manifest indicators are commonly referred to as error terms. In all factor models, it is assumed that each indicator has residual variance, or variance that is not explained by the latent factor. This is often mistakenly referred to as "error variance," because not all indicator-specific residual variance should not be interpreted as error. B=borderline; P=paranoid; PD=personality disorder; S=schizoid.