Model fit is a fallible indicator of model quality in quantitative psychopathology research: a reply to Bader and Moshagen

Greene, A.L.; Eaton, N.R.; Forbes, M.K.; Fried, E.I.; Watts, A.L.; Kotov, R.; Krueger, R.F.

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Psychopathology and Clinical Science

In the public domain
ISSN: 2769-7541

2022, Vol. 131, No. 6, 696–703
https://doi.org/10.1037/abn0000770

REPLY

Model Fit Is a Fallible Indicator of Model Quality in Quantitative Psychopathology Research: A Reply to Bader and Moshagen

Ashley L. Greene[1, 2], Nicholas R. Eaton[3], Miriam K. Forbes[4], Eiko I. Fried[5], Ashley L. Watts[6], Roman Kotov[7], and Robert F. Krueger[8]

[1] VISN 2 Mental Illness Research, Education and Clinical Center, James J. Peters VA Medical Center, Bronx, New York, United States
[2] Department of Psychiatry, Icahn School of Medicine at Mount Sinai
[3] Department of Psychology, Stony Brook University
[4] Centre for Emotional Health, School of Psychological Sciences, Macquarie University
[5] Unit of Clinical Psychology, Faculty of Social and Behavioral Sciences, Leiden University
[6] Department of Psychological Sciences, University of Missouri
[7] Department of Psychiatry, Stony Brook University
[8] Department of Psychology, University of Minnesota, Twin Cities

As evidenced by our exchange with Bader and Moshagen (2022), the degree to which model fit indices can and should be used for the purpose of model selection remains a contentious topic. Here, we make three core points. First, we discuss the common misconception about fit statistics' abilities to identify the "best model," arguing that mechanical application of model fit indices contributes to faulty inferences in the field of quantitative psychopathology. We illustrate the consequences of this practice through examples in the literature. Second, we highlight the parsimony-adjacent concept of fitting propensity, which is not accounted for by commonly used fit statistics. Finally, we present specific strategies to overcome interpretative bias and increase generalizability of study results and stress the importance of carefully balancing substantive and statistical criteria in model selection scenarios.

General Scientific Summary
In this piece, we review the limitations of model fit indices, which limit the validity of inferences that are based on them. In doing so, we encourage psychopathology researchers to think about model selection more broadly, and to approach model fit assessments more cautiously.

Keywords: model selection, fitting propensity, model complexity, bifactor model, factor analysis

Supplemental materials: https://doi.org/10.1037/abn0000770.supp

Motivated by the trend of using close fit as decisive evidence of structural fidelity in the *p*-factor literature, our target article extended prior simulation work on cognitive abilities (Morgan et al., 2015; Murray & Johnson, 2013) to psychopathology (Greene et al., 2019). The relative fit of the confirmatory correlated factor and bifactor models to the simulated data varied as a function of the presence or absence and magnitude or placement of unmodeled complexities in the data-generating model. When the population model did not contain unmodeled complexities, relative and information-theoretic fit indices that penalize for the number of free parameters (hereafter referred to as "parsimony-adjusted fit indices") tended to favor the fitted model with fewer parameters—the pure correlated factors model—because all else was equal between the fitted models (i.e., identical model-implied covariance structures). When the population model was more complex, the bifactor model was systematically better at accommodating the

Ashley L. Greene https://orcid.org/0000-0002-7110-2560
Miriam K. Forbes https://orcid.org/0000-0002-6954-3818
Robert F. Krueger https://orcid.org/0000-0001-9127-5509

The ideas in this article have not been posted or presented elsewhere. We have no conflicts of interest to disclose. Ashley L. Greene is supported by the Office of Academic Affiliations, Advanced Fellowship Program in Mental Illness Research and Treatment, Department of Veterans Affairs. Ashley L. Watts is funded through K99AA028306 (Principal Investigator: Ashley L. Watts).

Correspondence concerning this article should be addressed to Ashley L. Greene, VISN 2 Mental Illness Research, Education and Clinical Center, James J. Peters VA Medical Center, 130 West Kingsbridge Road, Bronx, NY 10468, United States. Email: ashleylaurengreene@gmail.com

data than the correlated factor model, although the latter was substantively correct. This is but one source of spurious confidence in the bifactor model of psychopathology.

Bader and Moshagen (2022) contextualize our findings by describing the impact of unmodeled complexities on the behavior of specific model fit statistics. They conclude that model fit indices are not characterized by probifactor bias because "they performed as expected—they identified the model that was (contingent on its parsimony) most closely aligned with the empirical data" (p. 4). Here, we elaborate on one broad point of disagreement: the mathematical correctness of omnibus fit statistics does not preclude *probifactor interpretative bias*, which is the tendency for researchers to be misled and, in turn, prone to make problematic inferences about a given model's scientific value and theoretical plausibility. In the online supplemental materials, we also address the authors' potentially misleading discussion of the nested versus equivalent relations between our population models.

Because much of our disagreement with Bader and Moshagen surrounds the term "bias," we offer our two-part definition: (a) fit statistics have a predetermined preference for the confirmatory bifactor model over nested alternatives when they are fit to data containing unmodeled complexities of adequate size (Mansolf & Reise, 2017), which are largely unknown in real data and may be extensive; and (b) the validity of inferences about the "best" model is questionable when they are based on the confirmatory bifactor model's superior fit to the data relative to competing nested models due to its high fitting propensity (Bonifay & Cai, 2017; Falk & Muthukrishna, 2021).

Parsimony Is More Than the Number of Free Parameters

One specific point of contention we have with Bader and Moshagen's argument concerns statements that risk conflating "predictability" with "validity," such as "valid inferences regarding the structural representation of psychopathology require an unbiased assessment of the correspondence between competing theoretical models and empirical data (i.e., goodness of fit)" (p. 3). This treatment of parsimony is incomplete, because it does not address an adjacent concept known as *fitting propensity*, the ability to accommodate a wide range of data patterns (Bonifay, 2021; Preacher, 2006).

We argue that the validity of fit-based inferences is debatable, given that they are limited by the elements in fit indices' mathematical formulae. Goodness-of-fit is influenced by two sets of model features, (1) *parametric complexity*, the number of freely estimated parameters; and (2) *structural complexity*, the model's functional form, or the placement and flexibility of free parameters (Markon, 2019; Preacher, 2006; Raykov & Marcoulides, 1999).[1] The distinction between parametric and structural complexity is important because two models with the same parametric complexity (i.e., the same degrees of freedom) can have different functional forms and, thus, different fitting propensities. Consider a single-factor model and an orthogonal two-factor model (see Figure 1). The former fits well to numerous types of data, whereas the latter only fits well to data generated by two uncorrelated factors with the pattern of loadings closely aligned. Focusing on model fit is misleading because parsimony-adjusted fit indices penalize parametric complexity, *not* structural complexity. In fact, parsimony-adjusted fit indices favor structural complexity: "in a nested model selection setting, if each added parameter actually improves the model fit, both AIC and BIC ultimately select the most complex model" (Huang, 2017, p. 413). Thus, goodness-of-fit (F_0) improves as a function of both parametric and structural complexity.

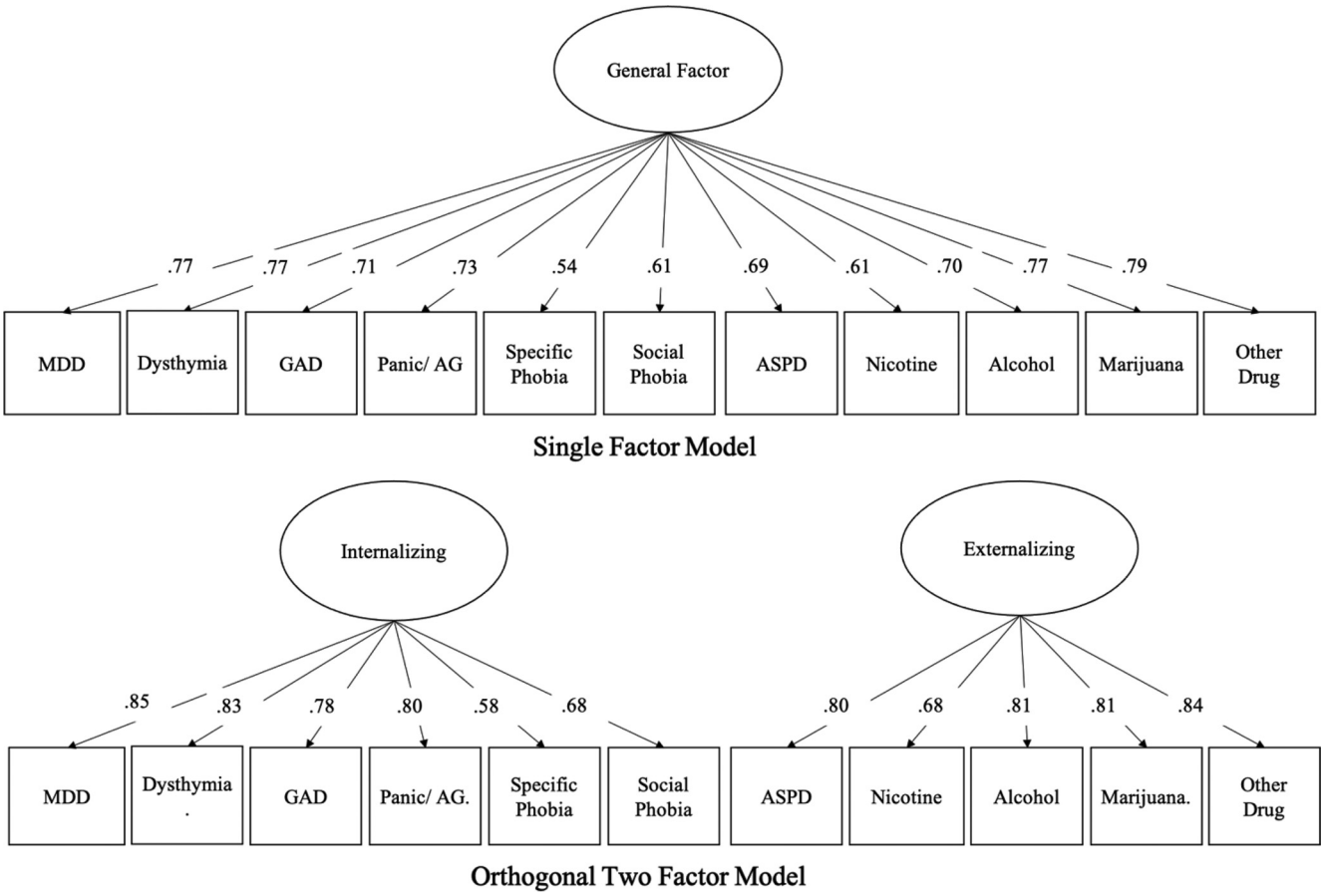
The misspecified bifactor model's advantage also increases as information accumulates through increased sample size (MacCallum, 2003; Marsh et al., 2004; Preacher, 2006), whereas the imperfectly specified correlated factor model will show greater misfit (i.e., higher F_0 values). Bader and Moshagen frame this moderating effect of sample size as a *positive* improvement for the performance of parsimony-adjusted fit indices, but we interpret this effect differently. The bifactor model is parametrically and structurally complex, so it is mistaken as a "factually" better model by fit statistics that have a preference for complexity as sample size increases. Therefore, while Bader and Moshagen are certainly correct that the bifactor model was more closely aligned with the data—as judged by the maximum likelihood discrepancy function (F_0)—it has the undesirable tendency to overfit data regardless of whether the population model has a bifactor structure (Bonifay & Cai, 2017; Markon, 2019). Thus, the bifactor model cannot readily distinguish between arbitrary noise and meaningful patterns of interest. It easily accommodates both (Bonifay & Cai, 2017; Reise et al., 2016; Watts et al., 2020).

We encourage readers to approach global fit assessments cautiously, despite Bader and Moshagen's emphasis on results derived from the discrepancy function (F_0): "the only reasonable statistical approach to choose the best model among a set of factually wrong candidate models is to base this decision on the degree of actual discrepancy, as one would clearly favor the model that is most closely aligned with the data—even if it might be wrong in some respects—over a model that is farther off" (p. 9).[2] While it is helpful to show that the data with unmolded complexities were less discrepant with the bifactor model (i.e., smaller F_0 values), we believe that Bader and Moshagen overstate the power of this approach. In light of the bifactor model's fitting propensity (Bonifay & Cai, 2017; Mansolf & Reise, 2017; Reise et al., 2016), its closer alignment with our simulated data are insufficient support for the claim that (a) parsimony-adjusted fit indices are able to identify the "best model" or (b) that the pure bifactor structure

[1] Parsimonious models constrain possible outcomes (Popper, 1959), which limits the number of datasets that they can fit well (Preacher, 2006; Roberts & Pashler, 2000). This view of parsimony illustrates why the higher-order model with three lower-order factors is more restrictive than the bifactor model: the unique constraints implied by the higher-order model results in less flexible parameters (i.e., factor loadings), whereas the flexible bifactor can fit a wider range of data, "regardless of truth or plausibility" (e.g., data that has no structure; Markon, 2019). Fit-based support for the bifactor model is weak, because fit indices cannot account for its higher flexibility compared with nested alternatives (i.e., similar to an exploratory factor model; Bonifay & Cai, 2017; Greene et al., 2022). Highly flexible parameters are characterized by "accommodational plasticity," leading to "predictive impotence" or the tendency to make inaccurate predictions due to overfitting (Hitchcock & Sober, 2004, p. 22).

[2] Another possible response to this point hinges on the fact that F_0 does not adjust for lack of parsimony. For instance, an exploratory factor model with k-1 factors (10 factors in this case) would consistently outperform every competing model herein on F_0, but it would not provide any new insights. The model would simply reconstruct the data without making it more interpretable. While traditional bifactor models are not as counterproductive as this example, they suffer from the same basic problem.

Figure 1
Two Models With Different Functional Forms, But the Same Degrees of Freedom



Note. Both models had 44 degrees of freedom. Model fit statistics for the single-factor model: root mean square error of approximation (RMSEA) = .049, comparative fit index (CFI) = .871, Tucker-Lewis index (TLI) = .839. Model fit statistics for the orthogonal two-factor model: RMSEA = .064, CFI = .782, TLI = .728. Parameter values were derived from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) dataset using weighted least square mean and variance adjusted (WLSMV) estimation.

provided the "best description of the data" compared with the other models. "Accommodation" is not the same as "description." The bifactor model only *reconstructed* the data better, which is not the same as *recovering* the true structure of our data (i.e., the true organization of latent variables).

Consequences of Probifactor Interpretative Bias

Retaining a model on the sole basis of superior fit blurs the distinctions between "optimal" and "best fitting" (Marsh et al., 2004; Roberts & Pashler, 2000; Sellbom & Tellegen, 2019). There are at least three important consequences of this practice, which we will situate within the context of research on the general factor of psychopathology (*p*-factor).

Difficulty Synthesizing Results

Models with high fitting propensity are less likely to generalize because they overfit data, meaning they accommodate unmodeled complexities in a correlation matrix that are likely to be sample- and study-specific (Bonifay, 2021). Models that overaccommodate noise or error often fit well due to the estimation of unnecessary parameters, which reduces measurement precision (e.g., low specific factor loadings; Hancock & Mueller, 2011; McNeish et al., 2018), verisimilitude (plausibility), and replicability (usefulness; Schmitt et al., 2018).

A natural consequence of this problem is that it is difficult to synthesize results from studies on bifactor models of psychopathology. Substantive interpretations of the bifactor model vary widely across studies with different samples, measures of similar constructs, and factor analytic methods (Greene et al., 2022; Watts et al., 2020). Indeed, the general factor's meaning routinely shifts across studies, and specific factors are often ill-defined (i.e., unreliable and inconsistent constructs; Watts et al., 2020). Moreover, Forbes and colleagues (Forbes et al., 2021) found that confirmatory bifactor models of psychopathology reveal their flexibility to fit any data through their imprecise parameter estimates (i.e., large standard errors) and unreliable specific factors (see also Watts et al., 2020). These trends demonstrate the bifactor's ability to accommodate components of error (e.g., correlated residuals, disturbance factors

or variances) that are included in the latent factor models being tested (Beauducel, 2013; Tomarken & Waller, 2003).

Unstable and Uninterpretable Latent Factors

Another consequence of solely relying on model fit to adjudicate models is that a model with high fitting propensity can fit well even if it is incompatible with the conceptual model. In bifactor models, incompatibilities often arise through weak specific factors that not statistically or conceptually isomorphic with their intended constructs—typically factors in the correlated factor model. Weak specific factors in bifactor models are often caused by the presence of pure indicators of the general factor (loadings that approach 1; Robertson, 2019), which leave no remaining variance to be explained by the specific factor. In turn, pure indicators define the general factor and convert it into an unintentional approximation of narrower content domains in the correlated factors model (Burns et al., 2020; Greene et al., 2022; Heinrich et al., 2021).

Seminal studies of the general factor of psychopathology selected optimal models of the structure of psychopathology with reference to model fit that included pure indicators of the general factor (Lahey et al., 2012) and specific factor collapse (Caspi et al., 2014).[3] Thus, latent factors in bifactor models of psychopathology were likely labeled prematurely given that these general factors were unplanned proxies for distress and thought disorder, respectively. Other interpretative consequences also followed, such as largely tautological inferences (i.e., the *p*-factor reflecting disordered thought; Caspi et al., 2014) that went on to inspire many subsequent studies. Thus, a potentially misguided choice between competing models based on fit indices can change a single study's discussion section and even shape the discourse within the field (Caspi & Moffitt, 2018).

Reifying Latent Factors

Confirmatory factor analysis contributes to the hyper-focus on model fit and the subsequent urge to reify latent factors, which conflates statistical and theoretical constructs. Statistical general factors are conflated with the *p*-factor as a substantive theoretical construct (Fried, 2020; Watts et al., 2020). As a *statistical construct*, a general factor simply summarizes the shared variance among the model's observed indicators—sometimes contributing little more than a sum score (Fried et al., 2021)—and may oversimplify multidimensional data (Forbes et al., 2021). As a *theoretical* construct, the *p*-factor's many faces are due to the derivation of post hoc theories by means of factor analysis (Greene et al., 2022). The *p*-factor literature parallels the long history of critiques of Spearman's theory of general intelligence: what is "general" to one set of variables rarely translates to another (Cattell, 1952; Horn & McArdle, 2007). Thus, numerous, potentially sample-specific, *p*-factor theories have been developed (for a review of competing *p*-factor definitions see Caspi & Moffitt, 2018; Smith et al., 2020). Systematizing measurement across studies and placing greater emphasis on replicating patterns of parameters, will improve generalizability and aid in the development of trustworthy constructs and theories.

Overcoming Interpretative Bias

It is imperative to carefully balance both statistical and conceptual justifications for preferring one model over another. The subjective nature of conducting factor analysis can interfere with this balance, though there are well-established guidelines for making logically, theoretically, clinically, and statistically informed determinations about the relative *quality* of models. In Table 1, we list recommended practices for addressing probifactor interpretative bias in applied psychopathology modeling scenarios, which threatens valid inference (Marsh et al., 2004; Tomarken & Waller, 2003, 2005). Even if a model meets each of these criteria, acknowledging the imperfection of all models is vital for understanding and testing latent structures (MacCallum, 2003).

Likewise, statistical and conceptual criteria require greater specificity because neither is useful absent a proper theory. When constructing a confirmatory model, researchers should explicitly state what types of evidence would support or refute their model/theory, and what evidence might rule out competing models (Fried, 2020; Watts et al., 2020). For instance, sensitivity analyses are useful for determining whether general factors of psychopathology are truly general. If so, they should be indifferent to their indicators, or consistently defined no matter which subset of indicators are included in the model (i.e., indicator invariance; Reise, 2012). To test for indicator invariance, Watts and colleagues (Watts et al., 2020) extracted bifactor models of psychopathology and dropped one of 15 indicators from the general factor, one at a time. They found poor congruence across the resulting general factors despite only dropping single indicators (convergent *r*s ranged from −.9 to .9). Thus, quantifications of generalizability (AIC/BIC [Akaike's information criterion/Bayesian information criterion]) are inadequate assessments of the assumption that a chosen structural model can withstand minor changes to a variable set, which is critical for defending a bifactor model's validity and reliability.

Another common practice is to disregard fit when adjudicating between models, selecting one because it is most "interpretable." Without a clear definition of "interpretability," this is akin to providing no justification at all. Instead, one might construct a formal or computational model that closely follows a theory, simulate data from that model, compare the simulated (i.e., theory-implied) data to real data, and then update the theory and the corresponding formal model based on discrepancies between the two data sets (Borsboom et al., 2021; Robinaugh et al., 2021).

Lastly, a critical flaw of many simulation studies, including ours, is that data are often generated from a model with perfect, or near perfect, simple structure. This is unrealistic, if not impossible, in applied settings, which limits ecological validity. Simple structure population models are unrealistic proxies for real-world psychopathology data-generating mechanisms. In fact, all models will always misfit real data; they are imperfect representations of the complex phenomena we seek to understand. Because the assumption that a latent factor model holds exactly in the population is untenable, we must test hypotheses using data that cannot be

[3] The collapse of the thought disorder specific factor in the Caspi et al. (2014) model resulted in an early example of the bifactor S-1 model, which has been proposed as one solution to some of the limitations to symmetrical bifactor models (Burns et al., 2020; Eid, 2020; Eid et al., 2017; Heinrich et al., 2021).

Table 1
Comparable Aspects of Model Quality

	General
Are analytic goals aligned with a general factor model's assumption?	• Orthogonalized bifactor applications are best suited for item-level analysis to assess what *indicators* have in common, while higher-order models assess what latent *factors* have in common (Decker, 2021) • If a higher-order model's second-order factor is only meaningful in the context of their associations with lower-order factors, a bifactor model should be avoided in nested model comparisons.
How interpretable is my latent factor model?	• Substantive interpretability is judged by the placement, magnitude, and direction of values for factor loadings and factor intercorrelations, which should be aligned with theoretical expectations (Eid, 2020; Watts et al., 2020).
Are the intended constructs adequately defined?	• Latent constructs are well defined when the set of observed variables adequately describes the hypothesized latent factor's features (Clark & Watson, 2019).
	Reliability of latent factors
How well do latent factors capture systematic sources of variance?	• Model-based reliability indices characterize the degree to which a latent factor captures a meaningful amount of systematic variance (Rodriguez et al., 2016a, 2016b). • The magnitudes of standard errors for factor loadings are an indication of the precision of these parameters and should be compared across models (Waldman et al., 2017).
Paradoxical results between model fit and model-based reliability indices?	• Models with high measurement quality tend to fit worse than those with low measurement quality (Hancock & Mueller, 2011).
	Dimensionality
Does a set of indicators reflect a unidimensional or multidimensional construct?	• Bifactor model-based reliability coefficients may be used to quantify the average parameter bias that is introduced when a single factor model it applied to data with some multidimensionality (Reise et al., 2013) • Evaluate whether specific factors have incremental validity in predicting important external criteria over and above the general factor (Ferrando & Lorenzo-Seva, 2019).
Spurious support for construct's dimensionality?	• Support for unidimensionality may increase as a function of the number of cases without a diagnosis, or as a function of indicator skewness (Watts et al., 2021). • The bifactor model can arise in the presence of population heterogeneity (Raykov et al., 2019).
	Overfitting
Are there large discrepancies between results derived from exploratory and confirmatory methods?	• A strong test of the validity of a confirmatory model is to examine whether it is detected using exploratory methods (Greene et al., 2022).
How generalizable is a well-fitting model?	• Models should be fit to multiple datasets as standard tests of whether they replicate out of-sample. This can be done by freezing model parameters from "discovery" to "replication" sample (also see Hitchcock & Sober, 2004; Preacher, 2006).
Misuse of modification indices.	• Added parameters, including correlated residuals, may not generalize across samples (MacCallum et al., 1992). The same applies to parameters that are dropped (e.g., a general factor loading, a specific factor).
How falsifiable is a well-fitting model?	• Bayes Factors imposes an Occam's razor criterion that balances fit to the data and model complexity, which allows for an accounting of differences in fitting propensity that goes beyond the number of free parameters (Mulder, 2014).[a] • A better test of model fit is to fit that same model to random data, and to see if it fits well regardless of the data (ockhamSEM package in R, Falk & Muthukrishna, 2021; see also Bonifay & Cai, 2017).
	Global fit
Global fit indices assess overall fit to the data.	• Global fit provides no information about the presence/absence of model misspecifications, nor the adequacy of a model's structure (Hayduk, 2014a, 2014b).
A well-fitting model can be misspecified.	• Causal processes may be misspecified, such as relations between latent variables (Raykov, 2000). • The chi-square test of exact fit and other common fit indices are often insufficiently sensitive to detect model misspecifications.
A poor-fitting but correctly specified model?	• Model fit can decrease due to minor discrepancies between the observed and implied covariance matrices (Browne et al., 2002).
	Local fit
Model misspecifications?	• Attend to individual model parameters (Tomarken & Waller, 2003, 2005).
Diagnostic investigations of local misfit?	• Inspect residual output (Maydeu-Olivares, 2017; Tomarken & Waller, 2003), including individual residual cases, which may help adjudicate equivalent models (Raykov & Penev, 2014). • Nonsensical individual response patterns may be easily masked by overly complex models (Reise et al., 2016).

[a] While this approach is being applied in *SEM* contexts that pertain to models with inequality constraints, it is not clear whether the current state of the art is such that this overall approach can be applied to the specific nested model comparisons described here.

perfectly accommodated by any fitted model (see Cudeck & Browne, 1992; Montoya & Edwards, 2021). Researchers should also consider population structures that are not latent factor models to explore degrees of misfit for common factor models and assess the validity of factor model results.

Additional simulation studies are needed to probe psychopathology structures that feature greater numbers of factors and indicator variables, as well as other outcomes like confidence interval coverage and distributions of factor loadings, to overcome the limitations posed by the nested models investigated in our study. Recommended cut-offs for model fit indices are not fixed, but context-dependent (e.g., type of misspecification, estimation method, factor loading magnitude, model complexity, and strength of associations between observed variables, etc.; McNeish & Wolf, 2021; Xia & Yang, 2019). In the context of quantitative nosology research, the conditions where fit index values might be less likely to identify a good-fitting model remain unclear.

Conclusion

Evaluating and retaining models solely based on good model fit does not qualify as model selection. Instead, model selection is much broader: it is "the practice of evaluating theory-implied models relative to one another rather than to a fixed criterion" (Preacher, 2006, p. 254), with the aim of identifying the model that best balances conflicting goals—generalizability, plausibility, parsimony, and fidelity to the data (Myung et al., 2000; Pitt et al., 2002). Model selection is complex because (a) interpreting and defending a chosen model is a subjective practice (Browne & Cudeck, 1993; Cudeck & Henly, 1991; Marsh et al., 2004) and (b) as one prioritizes one goal over all others, there is an increased likelihood that the chosen model will not be the same one that best aligns with all goals (Montoya & Edwards, 2021; Preacher, 2006; Preacher et al., 2013). In fact, as we demonstrated here, prioritizing model fit comes with critical limitations. Moving forward, to improve the reliability and validity of psychopathology classification, we recommend that researchers prioritize *plausible* models that directly inform specific hypotheses, fit the data reasonably well, reliably capture the construct(s) of interest, and have lower fit propensity. Doing so will enhance *usefulness*—model replicability, construct validity, and research synthesis.

References

Asparouhov, T., & Muthén, B. (2019). Nesting and equivalence testing for structural equation models. *Structural Equation Modeling*, *26*(2), 302–309. https://doi.org/10.1080/10705511.2018.1513795

Bader, M., & Moshagen, M. (2022). No probifactor model fit index bias, but a propensity toward selecting the best model. *Journal of Psychopathology and Clinical Science*, *131*(6), 689–695. https://doi.org/10.1037/abn0000685

Beauducel, A. (2013). *The factor paradox: Common factors can be correlated with the variance not accounted for by the common factors!* arXiv. https://arxiv.org/abs/1308.4178

Bonifay, W. (2021). *Increasing generalizability via the principle of minimum description length*. PsyArXiv. https://doi.org/10.31234/osf.io/arpc7

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*(4), 465–484. https://doi.org/10.1080/00273171.2017.1309262

Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, *16*(4), 756–766. https://doi.org/10.1177/1745691620969647

Browne, M. W., MacCallum, R. C., Kim, C.-T., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, *7*(4), 403–421. https://doi.org/10.1037/1082-989X.7.4.403

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (Vol. 154, pp. 136–162). Sage Focus Editions.

Burns, G. L., Geiser, C., Servera, M., Becker, S. P., & Beauchaine, T. P. (2020). Application of the bifactor S–1 model to multisource ratings of ADHD/ODD symptoms: An appropriate bifactor model for symptom ratings. *Journal of Abnormal Child Psychology*, *48*(7), 881–894. https://doi.org/10.1007/s10802-019-00608-4

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

Caspi, A., & Moffitt, T. E. (2018). All for one and one for all: Mental disorders in one dimension. *The American Journal of Psychiatry*, *175*(9), 831–844. https://doi.org/10.1176/appi.ajp.2018.17121383

Cattell, R. B. (1952). *Factor analysis: An introduction and manual for the psychologist and social scientist*. Harper & Brothers.

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

Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika*, *57*(3), 357–369. https://doi.org/10.1007/BF02295424

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, *109*(3), 512–519. https://doi.org/10.1037/0033-2909.109.3.512

Decker, S. L. (2021). Don't use a bifactor model unless you believe the true structure is bifactor. *Journal of Psychoeducational Assessment*, *39*(1), 39–49. https://doi.org/10.1177/0734282920977718

Eid, M. (2020). Multi-faceted constructs in abnormal psychology: Implications of the bifactor S - 1 model for individual clinical assessment. *Journal of Abnormal Child Psychology*, *48*(7), 895–900. https://doi.org/10.1007/s10802-020-00624-9

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, *22*(3), 541–562. https://doi.org/10.1037/met0000083

Falk, C., & Muthukrishna, M. (2021). Parsimony in model selection: Tools for assessing fit propensity. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000422

Ferrando, P. J., & Lorenzo-Seva, U. (2019). An external validity approach for assessing essential unidimensionality in correlated-factor models. *Educational and Psychological Measurement*, *79*(3), 437–461. https://doi.org/10.1177/0013164418824755

Forbes, M. K., Greene, A. L., Levin-Aspenson, H. F., Watts, A. L., Hallquist, M., Lahey, B. B., Markon, K. E., Patrick, C. J., Tackett, J. L., Waldman, I. D., Wright, A. G. C., Caspi, A., Ivanova, M., Kotov, R., Samuel, D. B., Eaton, N. R., & Krueger, R. F. (2021). Three recommendations based on a comparison of the reliability and validity of the predominant models used in research on the empirical structure of psychopathology. *Journal of Abnormal Psychology*, *130*(3), 297–317. https://doi.org/10.1037/abn0000533

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Fried, E. I., Greene, A. L., & Eaton, N. R. (2021). The p factor is the sum of its parts, for now. *World Psychiatry*, *20*(1), 69–70. https://doi.org/10.1002/wps.20814

Gignac, G. E. (2016). The higher-order model imposes a proportionality constraint: That is why the bifactor model tends to fit better. *Intelligence*, *55*, 57–68. https://doi.org/10.1016/j.intell.2016.01.006

Giordano, C., & Waller, N. G. (2020). Recovering bifactor models: A comparison of seven methods. *Psychological Methods*, *25*(2), 143–156. https://doi.org/10.1037/met0000227

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. https://doi.org/10.1037/abn0000434

Greene, A. L., Watts, A. L., Forbes, M. K., Kotov, R., Krueger, R. F., & 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://doi.org/10.1037/met0000465

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, *71*(2), 306–324. https://doi.org/10.1177/0013164410384856

Hayduk, L. A. (2014a). 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. https://doi.org/10.1177/0013164414527449

Hayduk, L. A. (2014b). Shame for disrespecting evidence: The personal consequences of insufficient respect for structural equation model testing. *BMC Medical Research Methodology*, *14*(1), 124. https://doi.org/10.1186/1471-2288-14-124

Heinrich, M., Geiser, C., Zagorscak, P., Burns, G. L., Bohn, J., Becker, S. P., Eid, M., Beauchaine, T. P., & Knaevelsrud, C. (2021). On the meaning of the "p factor" in symmetrical bifactor models of psychopathology: Recommendations for future research from the bifactor-(S-1) perspective. *Assessment*. Advance online publication. https://doi.org/10.1177/10731911211060298

Hershberger, S. L., & Marcoulides, G. A. (2006). The problem of equivalent structural models. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 13–41). Information Age Publishing Inc.

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *The British Journal for the Philosophy of Science*, *55*(1), 1–34. https://doi.org/10.1093/bjps/55.1.1

Horn, J. L., & McArdle, J. J. (2007). Understanding human intelligence since Spearman. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100* (pp. 219–262). Routledge.

Huang, P.-H. (2017). Asymptotics of AIC, BIC, and RMSEA for model selection in structural equation modeling. *Psychometrika*, *82*(2), 407–426.

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

MacCallum, R. C. (2003). 2001 Presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *38*(1), 113–139. https://doi.org/10.1207/S15327906MBR3801_5

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Mansolf, M., & Reise, S. P. (2016). Exploratory bifactor analysis: The Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. *Multivariate Behavioral Research*, *51*(5), 698–717. https://doi.org/10.1080/00273171.2016.1215898

Mansolf, M., & Reise, S. P. (2017). When and why the second-order and bifactor models are distinguishable. *Intelligence*, *61*, 120–129. https://doi.org/10.1016/j.intell.2017.01.012

Markon, K. E. (2019). Bifactor and hierarchical models: Specification, inference, and interpretation. *Annual Review of Clinical Psychology*, *15*(1), 51–69. https://doi.org/10.1146/annurev-clinpsy-050718-095522

Marsh, H. W., 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*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika*, *82*(3), 533–558. https://doi.org/10.1007/s11336-016-9552-7

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment*, *100*(1), 43–52. https://doi.org/10.1080/00223891.2017.1281286

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

Molenaar, D. (2016). On the distortion of model fit in comparing the bifactor model and the higher-order factor model. *Intelligence*, *57*, 60–63. https://doi.org/10.1016/j.intell.2016.03.007

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. https://doi.org/10.1177/0013164420942899

Morgan, G. B., Hodge, K. J., Wells, K. E., & Watkins, M. W. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence*, *3*(1), 2–20. https://doi.org/10.3390/jintelligence3010002

Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in) equality constrained hypotheses. *Computational Statistics & Data Analysis*, *71*, 448–463. https://doi.org/10.1016/j.csda.2013.07.017

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, *41*(5), 407–422. https://doi.org/10.1016/j.intell.2013.06.004

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(21), 11170–11175. https://doi.org/10.1073/pnas.170283897

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*(3), 472–491. https://doi.org/10.1037/0033-295X.109.3.472

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson & Co.

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, *41*(3), 227–259. https://doi.org/10.1207/s15327906mbr4103_1

Preacher, K. J., 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. https://doi.org/10.1080/00273171.2012.710386

Raykov, T. (2000). On sensitivity of structural equation modeling to latent relation misspecifications. *Structural Equation Modeling*, *7*(4), 596–607. https://doi.org/10.1207/S15328007SEM0704_4

Raykov, T., Marcoulides, G. A., Menold, N., & Harrison, M. (2019). Revisiting the bi-factor model: Can mixture modeling help assess its applicability? *Structural Equation Modeling*, *26*(1), 110–118. https://doi.org/10.1080/10705511.2018.1436441

Raykov, T., & Marcoulides, G. A. (1999). On desirability of parsimony in structural equation model selection. *Structural Equation Modeling*, *6*(3), 292–300. https://doi.org/10.1080/10705519909540135

Raykov, T., & Penev, S. (2014). Exploring structural equation model misspecifications via latent individual residuals. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 133–146). Psychology Press.

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

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (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. https://doi.org/10.1080/00273171.2016.1243461

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

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, *73*(1), 5–26. https://doi.org/10.1177/0013164412449831

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

Robertson, S. (2019). *Bifactor models and factor collapse: A Monte Carlo study* [Doctoral dissertation]. Clemson University Institutional Repository.

Robinaugh, D., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. (2021). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, *16*(4), 725–743. https://doi.org/10.1177/1745691620974697

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53–61. https://doi.org/10.1007/BF02289209

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*(4), 345–362. https://doi.org/10.1080/00223891.2018.1449116

Sellbom, M., & Tellegen, A. (2019). Factor analysis in psychological assessment research: Common pitfalls and recommendations. *Psychological Assessment*, *31*(12), 1428–1441. https://doi.org/10.1037/pas0000623

Smith, G. T., Atkinson, E. A., Davis, H. A., Riley, E. N., & Oltmanns, J. R. (2020). The general factor of psychopathology. *Annual Review of Clinical Psychology*, *16*(1), 75–98. https://doi.org/10.1146/annurev-clinpsy-071119-115848

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with "well fitting" models. *Journal of Abnormal Psychology*, *112*(4), 578–598. https://doi.org/10.1037/0021-843X.112.4.578

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, *1*(1), 31–65. https://doi.org/10.1146/annurev.clinpsy.1.102803.144239

Waldman, I., Poore, H., Watts, A., Rathouz, P., Van Hulle, C., Zald, D., & Lahey, B. (2017). Issues in the validation of the general factor of psychopathology. *Behavior Genetics*, *47*(6), 676–676.

Watts, A. L., Lane, S. P., Bonifay, W., Steinley, D., & Meyer, F. A. C. (2020). Building theories on top of, and not independent of, statistical models: The case of the p-factor. *Psychological Inquiry*, *31*(4), 310–320. https://doi.org/10.1080/1047840X.2020.1853476

Watts, A. L., Meyer, F. A. C., Greene, A. L., Wood, P. K., Trull, T. J., Steinley, D., & Sher, K. J. (2021). *Spurious empirical support for the p-factor arises with the inclusion of undiagnosed cases*. PsyArXiv. https://doi.org/10.31234/osf.io/4tazx

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*(2), 113–128. https://doi.org/10.1007/BF02294531

Received June 8, 2021
Revision received April 30, 2022
Accepted May 16, 2022 ■