American Psychological Association

Journal of Abnormal Psychology

© 2019 American Psychological Association
0021-843X/19/$12.00

2019, Vol. 128, No. 7, 740–764
http://dx.doi.org/10.1037/abn0000434

Are Fit Indices Used to Test Psychopathology Structure Biased? A Simulation Study

Ashley L. Greene, Nicholas R. Eaton, and Kaiqiao Li
Stony Brook University

Miriam K. Forbes
Macquarie University

Robert F. Krueger
University of Minnesota

Kristian E. Markon
University of Iowa

Irwin D. Waldman
Emory University

David C. Cicero
University of Hawaii at Manoa

Christopher C. Conway
College of William and Mary

Anna R. Docherty
University of Utah School of Medicine

Eiko I. Fried
University of Amsterdam

Masha Y. Ivanova
University of Vermont

Katherine G. Jonas
Stony Brook University

Robert D. Latzman
Georgia State University

Christopher J. Patrick
Florida State University

Ulrich Reininghaus
Maastricht University and King's College London

Jennifer L. Tackett
Northwestern University

Aidan G. C. Wright
University of Pittsburgh

Roman Kotov
Stony Brook University

Structural models of psychopathology provide dimensional alternatives to traditional categorical classification systems. Competing models, such as the bifactor and correlated factors models, are typically

Editor's Note. Gregory A. Miller served as the action editor for this article. —AM

This article was published Online First July 18, 2019.

Ashley L. Greene and Nicholas R. Eaton, Department of Psychology, Stony Brook University; Kaiqiao Li, Department of Psychiatry, Stony Brook University; Miriam K. Forbes, Centre for Emotional Health, Department of Psychology, Macquarie University; Robert F. Krueger, Department of Psychology, University of Minnesota; Kristian E. Markon, Department of Psychological and Brain Sciences, University of Iowa; Irwin D. Waldman, Department of Psychology, Emory University; David C. Cicero, Department of Psychology, University of Hawaii at Manoa; Christopher C. Conway, Department of Psychological Sciences, College of William and Mary; Anna R. Docherty, Department of Psychiatry, University of Utah School of Medicine; Eiko I. Fried, Department of Psychology, University of Amsterdam; Masha Y. Ivanova, Department of Psychiatry, University of Vermont; Katherine G. Jonas, Department of Psychiatry, Stony Brook University; Robert D. Latzman, Department of Psychology, Georgia State University; Christopher J. Patrick, Department of Psychology, Florida State University; Ulrich Reininghaus, Department of Psychiatry and Neuropsychology, Maastricht University, and Centre for Epidemiology and Public Health, Health Service and Population Research Department, Institute of Psychiatry, Psychology, and Neuroscience, King's College London; Jennifer L. Tackett, Department of Psychology, Northwestern University; Aidan G. C. Wright, Department of Psychology, University of Pittsburgh; Roman Kotov, Department of Psychiatry, Stony Brook University.

Daniel Klein and Stacey Scott are acknowledged for their thoughtful feedback during the early stages of this project. Special thanks is also offered to Christian Luhmann for his valuable insight and constructive critiques of this work. Limited previews of this study's hypotheses and results were presented on two occasions. First, preliminary findings were discussed with HiTOP coauthors at the 2017 HiTOP meeting in Denver, Colorado. Second, some preliminary results were discussed in an APS conference presentation led by Roman Kotov. None of the results reported herein have been posted on any websites, listserves, or manuscript publications.

Correspondence concerning this article should be addressed to Ashley L. Greene, Department of Psychology, Stony Brook University, Psychology B Building, Stony Brook, NY 11794-2500. E-mail: ashleylaurengreene@gmail.com

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

compared via statistical indices to assess how well each model fits the same data. However, simulation studies have found evidence for probifactor fit index bias in several psychological research domains. The present study sought to extend this research to models of psychopathology, wherein the bifactor model has received much attention, but its susceptibility to bias is not well characterized. We used Monte Carlo simulations to examine how various model misspecifications produced fit index bias for 2 commonly used estimators, WLSMV and MLR. We simulated binary indicators to represent psychiatric diagnoses and positively skewed continuous indicators to represent symptom counts. Across combinations of estimators, indicator distributions, and misspecifications, complex patterns of bias emerged, with fit indices more often than not failing to correctly identify the correlated factors model as the data-generating model. No fit index emerged as reliably unbiased across all misspecification scenarios. Although, tests of model equivalence indicated that in one instance fit indices were *not biased*—they favored the bifactor model, albeit not unfairly. Overall, results suggest that comparisons of bifactor models to alternatives using fit indices may be misleading and call into question the evidentiary meaning of previous studies that identified the bifactor model as superior based on fit. We highlight the importance of comparing models based on substantive interpretability and their utility for addressing study aims, the methodological significance of model equivalence, as well as the need for implementation of statistical metrics that evaluate model quality.

General Scientific Summary
Latent variable models of psychopathology provide dimensional alternatives to traditional categorical classification systems (e.g., *DSM–5* and ICD-11), with the two most popular being the bifactor and correlated factors models. These competing structural models of psychopathology are often compared via statistical indices to assess how well each model fits the same data. The results of our simulation study suggest that bifactor models are often erroneously favored over correlated factor models when the simulated data were generated by a correlated factors model with minor misspecifications. Findings from tests of model equivalence also clarified the conditions under which fit indices' favoring of the bifactor model was characterized by bias. This calls into question the common practice of relying on common fit statistics when comparing structural models of psychopathology.

Keywords: bifactor model, factor analysis, fit index bias, model evaluation, Monte Carlo simulation

Supplemental materials: http://dx.doi.org/10.1037/abn0000434.supp

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Current mental disorder classification systems (e.g., *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition [*DSM–5*]; American Psychiatric Association, 2013) have significant limitations as organizational frameworks for clinical research and intervention efforts. As a case in point, these taxonomies postulate that mental disorders are independent, categorical entities. However, these diagnostic categories exhibit markedly heterogeneous presentations within individuals (Olbert, Gala, & Tupler, 2014), have poor reliability (e.g., Fried et al., 2016), and display high rates of comorbidity (for reviews of these issues see Krueger & Markon, 2006; Trull & Durrett, 2005). Such issues point to a notable mismatch between the model (*DSM* diagnoses) and the data (signs and symptoms as they manifest in patients; Kotov et al., 2017; Krueger & Eaton, 2015), which highlights the importance of investigating structural conceptualizations of mental disorders (Krueger, 1999; Loevinger, 1957; Meehl, 2001, 2004).

Attempts to address these issues have led to the proliferation of new quantitative approaches for conceptualizing psychopathology in a data-driven way, which have highlighted a set of core transdiagnostic dimensions. For instance, studies have found robust evidence for two major transdiagnostic factors, internalizing (accounting for associations among mood and anxiety disorders) and externalizing (accounting for associations among disorders of antisociality, impulsivity, substance use, etc.; Eaton et al., 2012; Eaton, Krueger, & Oltmanns, 2011; Forbush & Watson, 2013; Kramer, Krueger, & Hicks, 2008; Krueger, 1999), as well as the bifurcation of the internalizing factor into distress and fear subfactors (see Figure 1; Eaton et al., 2013; Krueger, 1999; Slade & Watson, 2006; Watson, 2009). Beginning with the factor analytic work of Achenbach and colleagues on dimensional syndromes (Achenbach, 1966; Achenbach, Conners, Quay, Verhulst, & Howell, 1989; Achenbach, Ivanova, & Rescorla, 2017), structural analyses have revealed transdiagnostic factors underlying a wide range of mental disorders in children, adolescents, and adults (Achenbach, Krukowski, Dumenci, & Ivanova, 2005; Kotov et al., 2011; Lahey et al., 2008; Slade & Watson, 2006; Vollebergh et al., 2001; Wright et al., 2013).

Such structural models of psychopathology provide parsimonious summaries of observed patterns of psychiatric comorbidity, and as a result have gained a great deal of traction, including improved reliability and validity, demonstrated utility, and clinical applications (Andrews et al., 2009; Eaton, Rodriguez-Seijas, Carragher, & Krueger, 2015; Kim & Eaton, 2017; Rodriguez-Seijas, Eaton, & Krueger, 2015; Rodriguez-Seijas, Eaton, Stohl, Mauro, & Hasin, 2017; Waszczuk et al., 2017). An increasing number of findings in the recent literature also suggest that various models can be situated into an overarching hierarchy (Farmer, Seeley, Kosty, Olino, & Lewinsohn, 2013; Kim & Eaton, 2015; Kotov et

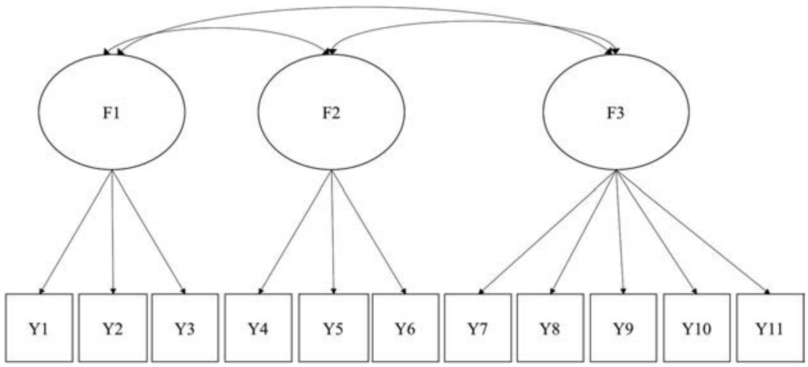

Figure 1. Three-factor oblique confirmatory factor analytic.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

al., 2017; Markon, Krueger, & Watson, 2005; Wright & Simms, 2015). Indeed, such findings have culminated in the development of a recently proposed framework—the Hierarchical Taxonomy of Psychopathology (HiTOP)—that organizes internalizing, externalizing, and other transdiagnostic dimensions (e.g., thought disorder, somatic problems, sexual dysfunctions) into an multilevel hierarchy (Kotov et al., 2017). This allows for an investigation of *hierarchy as construct* (Forbes et al., 2017; Kim & Eaton, 2015; Seeley, Kosty, Farmer, & Lewinsohn, 2011), which challenges researchers to think about how these factors can be integrated into comprehensive hierarchical structures (e.g., Kotov et al., 2017), as well as issues of breadth and specificity (Krueger, Tackett, & MacDonald, 2016). Thus, progress is being made with regard to using structural approaches to delineate a quantitative taxonomy of psychopathology.

Transdiagnostic Model Comparisons

Although evidence from structural research has converged on transdiagnostic reconceptualizations of mental disorder classification, fundamental questions of how best to model these constructs remain unclear. For instance, there is a great deal of support for two distinct transdiagnostic factors, internalizing and externalizing, which tend to be sizably correlated (e.g., ranging from $r = .4$ to .7; for a review and discussion of psychopathology factor interrelations, see Eaton, South, & Krueger, 2010), but not so highly that these constructs are conceptually indistinct (i.e., factor correlations $\geq$.80 or .85 are indicative of poor discriminative validity; Brown, 2015). Even so, the sizable correlations among transdiagnostic factors in structural models have led some researchers to posit a *general factor of psychopathology* (Caspi et al., 2014; Lahey et al., 2012, 2015; Patalay et al., 2015; Simms, Grös, Watson, & O'Hara, 2008; Snyder, Young, & Hankin, 2017). In an effort to investigate the possible presence of such a general factor of psychopathology, multiple studies have used a *bifactor* modeling approach, which specifies a general factor of psychopathology that saturates *all* mental disorders, along with specific factors, such as internalizing and externalizing, to capture residual covariation among indicators and reduce between-factor correlations (see Figure 2). In this modeling approach, the general factor is parameterized to be orthogonal to (i.e., uncorrelated with) the specific factors, and, most commonly, the specific factors are also parameterized to be orthogonal to one another (Brown, 2015; Holzinger & Swineford, 1937; Reise, 2012)—although bifactor models with correlated specific factors have sometimes been used (Carragher et al., 2016; Caspi et al., 2014; Laceulle, Vollebergh, & Ormel, 2015; Lahey, Krueger, Rathouz, Waldman, & Zald, 2017; Olino, McMakin, & Forbes, 2018; Patalay et al., 2015; Waldman, Poore, van Hulle, Rathouz, & Lahey, 2016). Regardless of the exact parameterization, the bifactor and correlated transdiagnostic factors models imply very different conceptualizations of the latent structure of mental disorders and how transdiagnostic factors (and thus mental disorders) relate to one another (for a detailed discussion on the different theoretical implications of these models see van Bork, Epskamp, Rhemtulla, Borsboom, & van der Maas, 2017). In particular, the correlated factors model estimates dimensions of psychopathology from the total shared variance among subsets of observed indicators (e.g., fear is defined by the variance shared across phobias). In contrast, the bifactor model estimates both general and specific dimensions of psychopathology, where the general dimension represents what is shared across all indicators, and the specific dimensions reflect more circumscribed patterns of shared residual variance apart from general psychopathology (e.g., features unique to fear once the general factor is taken into account).

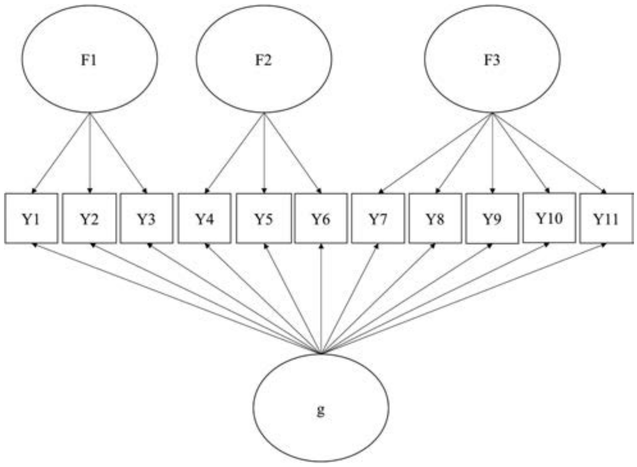

Figure 2. Four-factor orthogonal bifactor confirmatory factor analytic.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Multiple studies have compared the correlated-factor and bifactor modeling approaches to characterizing psychiatric comorbidity (e.g., Carragher et al., 2016; Caspi et al., 2014; Laceulle et al., 2015; Lahey et al., 2012; Lahey, Zald, et al., 2017; Olino, Dougherty, Bufferd, Carlson, & Klein, 2014). These studies typically have adjudicated these competing models via the comparison of model fit indices (e.g., Bayesian information criterion [BIC] values). That is, researchers fit several competing statistical models to a given dataset, and determine via fit indices which model is superior. In Figure 3, we depict the number of studies published per year, between 2010 to October 2017, that compared correlated-factor and bifactor models of psychopathology data ($N = 56$; see Supplemental Materials Appendix for details). The notable increase in such studies after 2014, from about three per year from 2010 through 2014, to 14 per year from 2015 to 2017, is indicative of the bifactor model's rising popularity—and extrapolating from the figure, this trend seems likely to continue. Most strikingly, the bifactor model was deemed superior to the correlated-factor model in 95% of the studies we reviewed. This may largely account for the notable proliferation of bifactor models in recent structural psychopathology research.

Bias in Model Fit Indices

In recent years, a body of work has emerged in the modeling literature, particularly in research on cognitive abilities, suggesting that traditional fit indices are biased[1] in favor of the bifactor model (Bonifay & Cai, 2017; Gignac, 2016; Mansolf & Reise, 2017; Maydeu-Olivares & Coffman, 2006; McFarland, 2016; Molenaar, 2016; Morgan, Hodge, Wells, & Watkins, 2015; Murray & Johnson, 2013; Reise, Kim, Mansolf, & Widaman, 2016; Yu, 2002). Two findings are particularly relevant. First, simulation studies have indicated that, even when data are generated from a known population-level correlated factors model without misspecifications, comparative fit indices (i.e., CFI/TLI) tend to favor the bifactor model rather than the correlated factors model, with BIC only performing well in larger sample sizes (e.g., when $N = 800$ as opposed to when $N = 200$; Morgan et al., 2015). Second, when model misspecifications are added (i.e., a parameter included in the population-level data generation model is not included in the simple structure models fit to the simulated data, such as small correlated residuals between indicators, resulting in misfit), these misspecifications often distort the fit indices' performance toward the bifactor model (Murray & Johnson, 2013). These findings are concerning, given that researchers in psychopathology commonly choose models primarily based on fit,[2] a feeble practice when all candidate models tend to fit the data well.

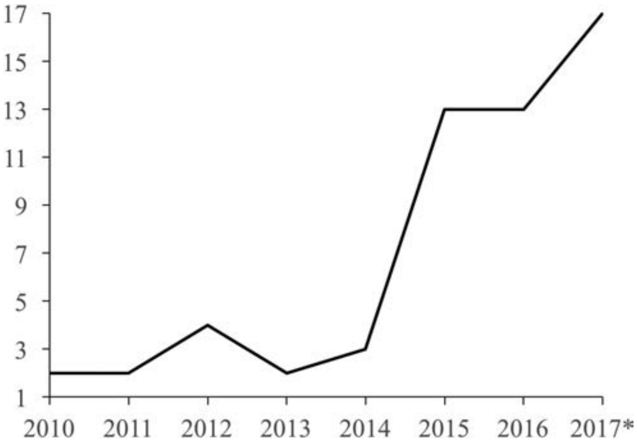

Figure 3. Number of studies comparing correlated factors and bifactor models of psychopathology data per year. *As of October 27, 2017.

There are several reasons for this general insensitivity of traditional fit indices when comparing latent variable models. Most important are considerations of what differentiates these models, such as the unique rank constraints that common measurement models imply for the data (i.e., different latent variable models entail distinguishable patterns of constraints on the observed covariance matrix; Mansolf & Reise, 2017; Silva, Scheines, Glymour, & Spirtes, 2006) and differences in fitting propensity (i.e., a model's average capacity to fit a variety of data patterns; Preacher, 2006). For example, the bifactor model has more built-in flexibility because of its extra dimension (i.e., p) and larger number of parameters (i.e., increased model complexity), which can accommodate minor misspecifications with fewer penalties to fit indices than the correlated-factor model, such as correlated residuals between indicators that are too small to justify inclusion in the model (Murray & Johnson, 2013). Further, there is evidence that the bifactor model risks overfitting data by capturing random noise (Bonifay & Cai, 2017) and/or capitalizing on fluctuations in sampling error that give rise to chance intercorrelations (Murray & Johnson, 2013), as opposed to valid variability that researchers intend to model (Reise et al., 2016). These properties increase the likelihood that a bifactor model will provide superior fit to data relative to the less complex correlated-factor model (Reise, 2012). Such evidence from other fields supports the possibility that findings from structural psychopathology studies in which a bifactor model was identified as superior to a correlated factors model by examination of fit indices may be the result of fit index bias. This apparent vulnerability of fit indices warrants a direct examination in a simulation study reflective of common scenarios in modeling the latent structure of mental disorders.

Unresolved Questions

Whereas prior simulation studies in the field of cognitive abilities have examined the limitations of using fit indices for model comparisons, this issue has not yet been thoroughly studied in the psychopathology literature. This is relevant because data typically encountered in structural studies of psychopathology differ from that in cognitive modeling in two aspects. First, cognitive ability models usually feature continuous data from considerably smaller sample sizes in the range of 200 to 2,000 (Chen, West, & Sousa, 2006; Gignac, 2016; Maydeu-Olivares & Coffman, 2006; Molenaar, 2016; Morgan et al., 2015; Murray & Johnson, 2013), with most studies focusing on comparisons between the higher-order

[1] Throughout this article, the term *bias* is referred to in a broad sense. That is, *bias* connotes a systematic distortion of fit statistics attributable to specific properties of the models they are used to evaluate.

[2] A wide range of methods and statistics exist for assessing models fitted to data. Within the context of this article, we refer to the term *fit* as conceptualized through the use of traditional fit indices and information criteria to maintain consistency with similar lines of previous research.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

and bifactor models. By contrast, psychopathology structural studies typically use either dichotomous indicators (e.g., present/absent diagnoses or criteria; Caspi et al., 2014; Greene & Eaton, 2016; Greene & Eaton, 2017; Laceulle et al., 2015) or positively skewed symptom count variables (Eaton et al., 2011; Olino et al., 2014). Second, these differences in indicator distributions warrant different estimators (cognitive: maximum likelihood; clinical: weighted least squares with adjusted means and variances [WLSMV], or robust maximum likelihood [MLR]). Thus, various model misspecifications particularly germane to structural psychopathology modeling scenarios (e.g., cross-loadings and correlated residuals among indicators between and across factors) have not been jointly examined in previous simulation research using more than one estimator and sample sizes greater than 2,000. In particular, no prior simulation study has accounted for these types of model errors when generating sample data sets from a known correlated factors population-level structure, despite previous demonstrations that these common characteristics of data have the potential to affect fit indices with both continuous and categorical variables (Morgan et al., 2015; Murray & Johnson, 2013; Yu, 2002). Only the one study by Morgan and colleagues (2015) directly assessed the performance of both the correlated factors and bifactor models when data were generated by a correlated factors model, but did not include misspecifications in their population model. Although Murray and Johnson (2013) provide preliminary work on the topic using a higher-order structure as the data generating model, they limited the size of correlated residual and cross-loading parameters to values of .10 to .20 as they were interested in *minor* unmodeled complexity (i.e., the criterion for meaningful cross-loadings is typically >.30; Schmitt & Sass, 2011), which might be too small for structural models of psychopathology (e.g., Greene & Eaton [2016] found panic with agoraphobia to cross-load on distress at .29 and fear at .45).

The Present Study

To address these questions, we simulated data from a known population-level correlated factors latent structure, and conducted separate analyses to include various model misspecifications. These data were simulated based on model parameters from one of the most seminal structural studies to date (Lahey et al., 2012), which showed superiority of the bifactor model over the correlated factors model via fit index comparisons. In simulating data, we created data-generating correlated three-factor models with (a) no misspecifications, (b) a cross-loading where one item loaded on two factors, (c) a correlated residual between two indicators loading on different factors, and (d) a correlated residual between two indicators loading on the same factor (see Figure 4). Because of the interplay between sample size and fit statistics/indices (Marsh, Hau, & Grayson, 2005), each misspecification was tested with different sample sizes and strengths of misspecification. To mimic characteristics of frequently modeled psychopathology data, we examined sets of positively skewed indicators (representing symptom counts) and sets of dichotomous indicators (representing diagnostic indicators) for which we used MLR and WLSMV estimators, respectively. Thus, in seeking to extend previous lines of this research to psychopathology, this study is novel because of its inclusion of both MLR & WLSMV estimators, an expanded the range of sample sizes (500 to 40,000), as well as an increased range of types and magnitude of population model misspecifications (.1, .3, and .5) that are based on values in prior clinical research (e.g., Greene & Eaton, 2016). Lastly, we assessed the extent to which our two competing models are statistically distinguishable by conducting tests of model equivalence between the bifactor model and each of the four data-generating correlated factor models (Hershberger & Marcoulides, 2006).

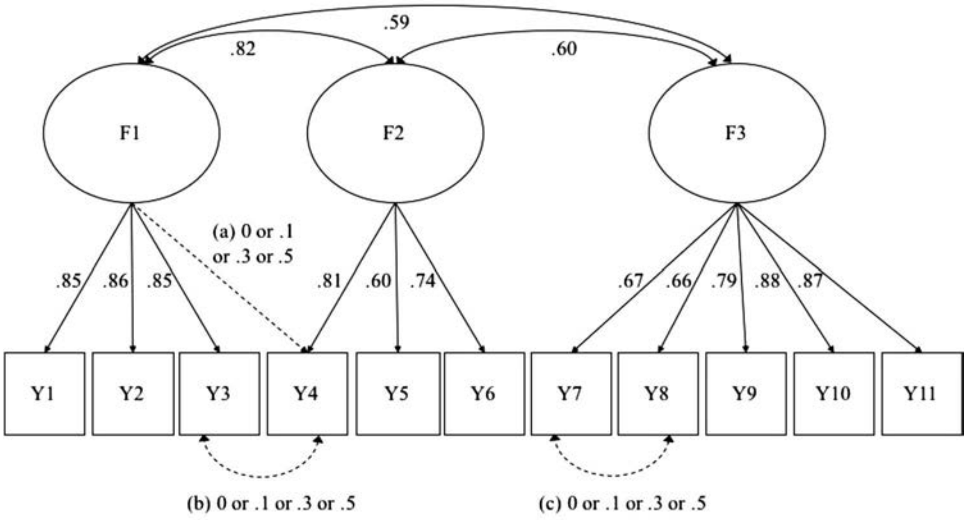

Figure 4. Simulation model from which random samples were generated. In this three-factor oblique confirmatory factor analytic, the solid lines represent parameters that were estimated for all models, whereas the dashed lines represent parameters that were manipulated, including (a) strength of cross-loading between factors, (b) strength of a between-factor correlated residual, and (c) strength of a within-factor correlated residual.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Method

We used the Monte Carlo simulation capabilities in M*plus* (Version 7.11; Muthén & Muthén, 1998–2015) to simulate sample data sets from a known population-level structure—a correlated three-factor model with a variety of model misspecifications, as described below—and then examined the performance of two confirmatory factor analytic (CFA) models commonly used in the psychopathology literature: a three-factor oblique CFA, representing the correlated factors model (see Figure 1), and a four-factor orthogonal bifactor CFA, representing the bifactor model (one general factor with three specific factors; Figure 2). Below, we describe how the simulated models were parameterized. All syntax can be found in the online supplementary materials.

Data Generation

Indicators. In psychopathology research, the structure and relative fit of CFA models are often investigated using large epidemiological samples. Thus, to increase the relevance of our simulations to psychopathology studies, we based population model parameters (i.e., factor correlations and loadings) on standardized solutions from the correlated three-factor model delineated by Lahey and colleagues (2012) using data from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC; for a full description of the sampling frame see Grant & Dawson, 2006). This true correlated factors model contained two just-identified factors with three indicators each (indicators denoted as Y1–Y6) and one factor with five indicators (Y7–Y11), corresponding to the distress, fear, and externalizing factors of Lahey et al.'s model (see Figure 4 for the correlated three-factor model used for simulations in the present study).

To investigate the performance of fit indices in models with categorical indicators, we parameterized 11 categorical indicators according to the proportion of endorsement for each in Wave 1 of NESARC (N = 43,093): Y1 (18.2%), Y2 (4.9%), Y3 (4.5%), Y4 (1.1%), Y5 (9.5%), Y6 (5.0%), Y7 (3.6%), Y8 (12.5%), Y9 (1.8%), Y10 (17.7%), and Y11 (1.3%). For investigation of fit index performance in models with skewed symptom count indicators, we parameterized continuous indicators with positively skewed response distributions (i.e., as typical of symptom counts in the general population) with skewness of 2.0 using the *MplusAutomation* package in R (Hallquist & Wiley, 2018). This level of skew is representative of real data distributions found in community-based mental health research (Curran, West, & Finch, 1996).

Estimators. Robust maximum likelihood (MLR) and mean-and-variance-corrected weighted least squares (WLSMV) estimators were used because they are the most common methods for handling discrete data and are robust to non-normality (Beauducel & Herzberg, 2006; Flora & Curran, 2004; Li, 2016; Rhemtulla, Brosseau-Liard, & Savalei, 2012; Savalei, 2014). For dichotomous diagnoses as indicators, all CFA models were fit to tetrachoric correlations using WLSMV, which is computationally less demanding than MLR when including correlated residuals between dichotomous indicators. For skewed continuous indicators, all CFA models were fit to Pearson correlations using MLR with a maximum of 1,000 iterations, because MLR is a continuous estimation method with statistical corrections to standard errors and chi-square statistics for non-normally distributed indicators and is used frequently in psychopathology research (Lahey et al., 2012, 2015; Olino et al., 2014; Snyder et al., 2017; Tackett, Daoud, De Bolle, & Burt, 2013).

Sample size. To investigate fit index bias as a function of sample size, we simulated data for nine samples of varying size (N = 500; 1,000; 2,000; 3,000; 4,000; 5,000; 10,000; 20,000; 40,000), representing a broad range of samples used in structural equation models in the psychopathology literature.

Population models and sample model comparisons. Our simulation study included the manipulation of four parameters (totaling 180 conditions): 2 estimators (WLSMV vs. MLR) and related indicator type (categorical vs. skewed continuous) × 9 sample sizes × 10 model misspecifications (one correctly specified, three levels of factor cross-loadings, three levels of correlated residuals *within*-factor, and three levels of correlated residuals *between*-factors) = 180 cells. We generated an empirical sampling distribution of 500 virtual random samples for each simulation condition.

For our population-level models, we simulated data from four correlated factors models, where one model was correctly specified and three were misspecified. We then fit two models to the data—and compared the fit of a bifactor versus a correlated factors model to examine fit index performance. The first model for simulating data was a three-factor correlated factors model with no misspecifications. We then investigated the impact of the three types of model misspecification in separate simulations (see Figure 4): (a) a cross-loading where one indicator loaded on two factors (i.e., Y4 on both F1 and F2), (b) a correlated residual between two indicators loading on different factors (i.e., between Y3 and Y4), and (c) a correlated residual between two indicators loading on the same factor (i.e., between Y7 and Y8). These three types of misspecification were each represented at four levels of strength, with the standardized loading/correlated residual value fixed to either .00 (i.e., the misspecification was not present), or to .10, .30, or .50—values considered weak, moderate, and strong, respectively (Cohen, 1988). Each dataset included only one misspecification manipulation, providing a conservative test of bias, as psychopathology data are likely to deviate from simple structure CFA models in multiple respects. The rationale for specifying the residual correlation between two indicators loading on the two just-identified factors (F1 and F2) was that this would be most likely as these factors are most highly correlated. For the cross-loading, the proportions of endorsement for Y4 correspond to panic disorder with or without agoraphobia, given previous evidence that this composite variable includes characteristics of both distress and fear transdiagnostic factors in the NESARC dataset (i.e., standardized factor loading of .29 on the distress factor and .45 on the fear factor; Greene & Eaton, 2016).

Model Fit

We conducted tests of model nesting and equivalence, which is often difficult to evaluate in practice and tends to be ignored as a result (Bentler & Satorra, 2010; Henley, Shook, & Peterson, 2006; Hershberger & Marcoulides, 2006; Maccallum, Wegener, Uchino, & Fabrigar, 1993; Raykov & Penev, 1999). Broadly defined, equivalent models differ in structure and substantive explanations of the data being described, but cannot be differentiated using measures of overall fit because they yield identical model-implied

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

covariance matrices, residuals, and goodness-of-fit indices, such as chi-square values and descriptive fit indices (Hershberger & Marcoulides, 2006). When defined in terms of nesting (Bentler & Satorra, 2010), models are *covariance matrix nested* when they have different degrees of freedom, but the implied covariance matrix under the more restricted model (e.g., correlated factor) can be perfectly reproduced under the more general model (e.g., bifactor). For simplicity, the term *model equivalence* will be retained throughout. To investigate model equivalence, we conducted separate tests for each type and level of misspecification (i.e., no misspecification, cross-loading, and correlated error) for the four population-level correlated factor models, with a sample size of $N = 1{,}000$. First, we obtained the covariance matrix implied by each data-generating correlated factor model and then fit each resulting matrix by both a bifactor model and a correlated factor model (with the relevant misspecification, such as a cross-loading, if present in the population-level model). In other words, we sought to characterize how well the bifactor model might accommodate a population-level correlated factor model (i.e., discrepancies due to approximation). Identification of potential model equivalence is important for providing a comprehensive account of the underlying reasons for fit indices' favoring of a bifactor model relative to a correlated factor model: (a) probifactor *bias* due to sampling error and capitalization on chance (i.e., discrepancies attributable to estimation), as opposed to (b) probifactor bias due to perfectly reproducing the population covariance matrix implied by one, or more, of our correlated three-factor models of interest.

Fit index criteria. We examined the performance of various fit indices in correctly identifying the data as emerging from a population-level correlated factors model versus incorrectly identifying the bifactor model as superior by fit. To do so, we investigated the following fit statistics, which are common in psychopathology research. First, to quantify fit of the data to each fitted model we used the root mean squared error of approximation (RMSEA; Steiger, 1990), for which good fit is indicated by values $< .06$. Second, to measure differences between sample and estimated variance and covariances we used the weighted root-mean-square residual (WRMR; Muthén & Muthén, 1998–2015) for dichotomous indicators, and the standardized root-mean-square residual (SRMR; Hu & Bentler, 1995) for continuous indicators. Good fit is indicated by WRMR <1.0 and SRMR $< .08$. Third, to assess improvement in fit relative to a saturated model, we used the comparative fit index (CFI; Bentler, 1990) and the Tucker–Lewis index (TLI). Values of CFI/TLI $> .95$ are common guidelines for good model fit (Hu & Bentler, 1999). Fourth, to compare the two models directly against each other, we used the Akaike information criterion (AIC; Akaike, 1987), Bayesian information criterion (BIC; Raftery, 1995), and the sample-size adjusted BIC (SABIC; Sclove, 1987), for which lower values are superior. These indices were not available for WLSMV, because they are not defined in least squares estimation.

Fitted model comparisons. Models were compared in several ways. First, to approximate the typical approach in the literature, we averaged each fit index across all 500 simulations of each of the 180 model parameterizations, and then we compared whether the bifactor or correlated factors model exhibited a superior mean value for each fit index. Situations in which means were equal for a given index were considered a tie, and thus as a failure of the fit index to correctly identify that the sample data were generated from a correlated factors model. Standard deviation (*SD*) units were also calculated for the mean values of each fit index. Second, to address the ubiquitous issue of all competing models fitting well in most psychopathology studies, we examined whether the size of differences (Δ) in mean TLI ($>.010$; Gignac, 2007) and AIC/BIC (<10; Raftery, 1995) values met established criteria when either model was found to fit best. Third, we examined the percentage of times that fit indices correctly identified the correlated model as superior, incorrectly favored the bifactor model, and the percentage of ties across all 500 simulations in each study condition. The threshold for strong model selection performance was $\geq 95\%$, because these results are intended to inform an applied perspective (e.g., if a researcher is comparing two models, one is correct and one is wrong, what is the probability they will choose the correct model?). Lastly, as a formal comparison of which model was closer to the true data generating model, the Vuong test for non-nested structural models (Vuong, 1989) was included as a test of both AIC and BIC differences between MLR models using the log-likelihood for model selection (Merkle, You, & Preacher, 2016). In each simulation setting, we examined the percentage of significant models ($p < .05$) as a test of whether the correlated factor or bifactor model fit better than the other according to AIC and BIC.

Results

Model Convergence

Across the 180,000 simulated data sets (180 cells × 500 random samples × 2 models fit to each), 3.92% failed to converge. Nearly all models that failed to converge were bifactor solutions (i.e., 7.84% of bifactor models vs. 0.008% of correlated factor models). Of the 45,000 correlated factors models estimated using WLSMV for dichotomous indicators, seven (0.02%) solutions did not converge, which all emerged from the smallest dataset with $N = 500$. Of the 45,000 estimated bifactor models using WLSMV, 5,988 (13.31%) solutions did not converge; this pattern was evident for samples of differing size, save for when sample size was equal to 40,000 in which case all bifactor solutions converged. For models estimated using MLR, all models that failed to converge were bifactor solutions. Of the 45,000 bifactor models generated for each sample size, a total of 1,069 (2.38%) solutions did not converge; 759 (1.69%) of these were from data sets with a sample size of 500, 252 (0.56%) from a sample size of 1,000, 44 (0.10%) from a sample size of 2,000, and 10 (0.02%) from a sample size of 3,000. Thus, as sample size increased, the proportion of models that successfully converged also increased. We also observed that nonconvergence might be related to response category or the type of estimator used, as nonconvergence was especially high for bifactor models estimated with WLSMV methods for dichotomous variables. When taken together, these results are consistent with previous observations that higher nonconvergence rates are associated with both small sample sizes and binary indicators (Flora & Curran, 2004).

Model Equivalence

For analyses involving the *correctly specified* correlated factor model, the bifactor model perfectly reproduced this implied

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

population-level covariance matrix. So, our data-generating correlated factor model could be perfectly reexpressed as a bifactor model when no cross-loadings or correlated errors were present (i.e., the more restricted correlated factor model is nested within the more general bifactor model). These two models also yielded identical fit index values for CFI, TLI, RMSEA, and SRMR, although the information criteria (AIC and BIC) did favor the more parsimonious correlated factor model (i.e., correlated factor model yielded lower values for AIC [−16] and BIC [−55.26]).

For analyses involving data-generation models with cross-loadings and correlated errors, the bifactor model was unable to perfectly reproduce these implied population-level correlation matrices (i.e., the bifactor model invariably produced some residuals). More specifically, when fit to each model-implied covariance matrix, the bifactor model provided a better fit for the cross-loading condition than the correlated error conditions, consistent with our results using mean values, percent correct, and the Vuong test. In all cases, as the size of the cross-loading and correlated error grew larger, the bifactor model showed poorer fit, especially when there was a within-factor correlated residual (e.g., RMSEA = .80 when this correlated residual was .5). The information criteria performed best across each condition (lower AIC/BIC values, ranging from −14.22 to −292.02), with TLI showing meaningful improvements in fit for the correlated factor model when between- and within-factor correlated residuals were moderate to large (improvements here were also observed for CFI and RMSEA values).

Fitted Model Comparisons

Models without misspecification. Table 1 presents the mean values of each model fit index for the fitted correlated factors (left half of each table) and bifactor models (right half) when the true correlated three-factor model was without misspecification (see Supplemental Materials Table 1 for *SD* values). The tables are color coded, such that dark gray shaded cells indicate a model showing superior values to the competing model, white cells indicate a model showing inferior values to the competing model, and light gray shaded cells indicate ties.

Across both MLR and WLSMV estimation methods, the SRMR and WRMR indices consistently favored the bifactor model. This is expected, despite these models' equivalent covariance matrices, as neither index penalizes for model complexity. For the remaining approximate fit indices—RMSEA, CFI, and TLI—most cases were ties, and, when not a tie, the bifactor model was deemed superior by fit more often than the correlated factors model. In contrast, the information criteria—BIC, SABIC, and to a lesser extent AIC—identified the correlated factor model as best fitting across all sample sizes. This trend is also expected as these indices include penalties for model complexity, with the BIC imposing the most severe penalty for less restricted models, which is particularly relevant to the bifactor model because it is less parsimonious than the correlated factors model.

Models with a cross-factor loading. Next, the models were compared based on data simulated from a correlated factors model that contained one indicator loading on two factors (see Figure 4). However, this cross-loading was not modeled in either the correlated factors or bifactor models that were fit to the data (see Figures 1 and 2). Table 2 presents the mean value of each global fit index for correlated factors and bifactor fitted models across all sample sizes and levels of misspecification (see Supplemental Materials Table 2 for *SD* values).

MLR. RMSEA, CFI, TLI, and SRMR all failed to correctly identify the correlated factors model as the true population-level model, in every single study cell, even when the cross-loading was small (0.1). Approximately two thirds of these comparisons incorrectly identified the bifactor model as superior, whereas the other third indicated a tie between models; ties were mostly only present when misspecification was small (cross-loading of 0.1), and no ties were present when the misspecification was large (cross-loading of 0.5). Similarly, AIC incorrectly favored the bifactor model in all but one comparison. However, BIC and SABIC performed somewhat better. BIC correctly identified the correlated factors model in nearly every comparison, although BIC incorrectly identified the bifactor model as superior when sample sizes were large ($N >$ 10,000) and the misspecification was moderate to large (cross-loading = 0.3 or 0.5). SABIC correctly supported the correlated factors model consistently when the cross-loading was small (0.1), inconsistently when the cross-loading was moderate (0.3), and incorrectly supported the bifactor model when the cross-loading was large (0.5) regardless of sample size.

WLSMV. Similar to the MLR results, all fit indices failed to correctly identify the correlated factors model as the true population-level model, except in a single cell. This was true for RMSEA, CFI, TLI, and WRMR. Across all cells and indices, approximately half favored the bifactor model and half produced a tie. In every case, WRMR favored the bifactor model, and all fit indices favored the bifactor model when sample sizes were small. As the size of the cross-loading increased, more indices that were tied came to favor the bifactor model. Indeed, at a cross-loading of 0.5, TLI came to favor the bifactor model in eight of nine comparisons.

Models with a between-factor correlated residual. We compared the fit of competing models when the true correlated factors model was specified to contain a correlated residual on *between*-factor indicators (see Figure 4). Table 3 presents the mean value of each global fit index across all sample sizes and levels of misspecification for correlated factors and bifactor fitted models (see Supplemental Materials Table 3 for *SD* values).

MLR. In most cases, RMSEA, CFI, and TLI fit indices failed to correctly identify the correlated factors model as the true population-level model. Many fit indices provided ties between models, with fewer providing support for the bifactor model, and even fewer correctly providing support for the correlated factors model. When the correlated residual was small (r = .1), indices tended to produce ties, with more support emerging for the bifactor model as the residuals became larger; when the residual was large (r = .5), all three indices incorrectly favored the bifactor model. SRMR incorrectly favored the bifactor model when the correlated residual was small, but correctly favored the correlated factor model only when the correlated residual was large. AIC incorrectly favored the bifactor model in nearly every comparison. BIC correctly identified the correlated factor model when the correlated residual was small, or when it was moderate and sample size was moderate; however, BIC incorrectly favored the bifactor model when the correlated residual was moderate at large sample sizes, or when the correlated residual was large. SABIC incorrectly sup-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 1
Mean Values for Each Model Fit Index in Which the True Correlated Three-Factor Model Contains No Misspecifications

	Fitted model using MLR for continuous indicators															
	Correlated factor								Bifactor							
Sample size	RMSEA	CFI	TLI	SRMR	AIC	BIC	SABIC	No. of models converged	RMSEA	CFI	TLI	SRMR	AIC	BIC	SABIC	No. of models converged
500	.015	.995	.995	.025	8851.72	9003.441	8889.18	500	.016	.996	.994	.019	8854.175	9039.618	8899.959	448
1,000	.008	.998	.999	.017	17751.55	17928.23	17813.89	500	.009	.998	.999	.013	17750.65	17966.59	17826.85	480
2,000	.005	.999	1.000	.012	35532.52	35734.15	35619.78	500	.005	.999	1.000	.010	35535.95	35782.39	35642.60	499
3,000	.004	1.000	1.000	.010	53298.68	53514.91	53400.53	500	.004	1.000	1.000	.008	53300.36	53564.64	53424.83	500
4,000	.004	1.000	1.000	.009	70956.16	71182.74	71068.35	500	.004	1.000	1.000	.007	70957.30	71234.23	71094.42	500
5,000	.003	1.000	1.000	.008	88842.42	89077.04	88962.65	500	.003	1.000	1.000	.006	88843.98	89130.73	88990.92	500
10,000	.002	1.000	1.000	.006	177680.50	177940.10	177825.60	500	.002	1.000	1.000	.004	177681.80	177999.10	177859.20	500
20,000	.001	1.000	1.000	.004	355295.30	355579.80	355465.40	500	.001	1.000	1.000	.003	355295.80	355643.60	355503.80	500
40,000	.001	1.000	1.000	.003	710765.10	711074.60	710960.20	500	.001	1.000	1.000	.002	710763.30	711141.60	711001.80	500

	Fitted model using WLSMV for dichotomous indicators															
	Correlated factor								Bifactor							
Sample size	RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged	RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged
500	.017	.984	.982	.766	—	—	—	497	.015	.988	.984	.640	—	—	—	300
1,000	.012	.991	.990	.758	—	—	—	500	.011	.994	.993	.626	—	—	—	341
2,000	.006	.997	.997	.711	—	—	—	500	.007	.997	.997	.604	—	—	—	423
3,000	.004	.999	.999	.678	—	—	—	500	.004	.999	.999	.576	—	—	—	442
4,000	.004	.999	.999	.684	—	—	—	500	.004	.999	.999	.583	—	—	—	467
5,000	.003	.999	1.000	.662	—	—	—	500	.003	.999	1.000	.563	—	—	—	474
10,000	.002	1.000	1.000	.664	—	—	—	500	.002	1.000	1.000	.565	—	—	—	498
20,000	.001	1.000	1.000	.665	—	—	—	500	.001	1.000	1.000	.567	—	—	—	497
40,000	.001	1.000	1.000	.656	—	—	—	500	.001	1.000	1.000	.558	—	—	—	500

Note. Dark grey shading reflects model superiority by fit. Light grey shading reflects instances in which the competing models are equivalent. Unshaded cells reflect model inferiority by fit. Number of free parameters for robust maximum likelihood (MLR) models: correlated factor (36) and bifactor (44). Number of free parameters for weighted least squares with adjusted means and variances (WLSMV) models: correlated factor (25) and bifactor (33). Degrees of freedom for correlated factor models (41) versus Bifactor (33). RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean residual; WRMR = weighted root-mean-square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion; SABIC = sample-size adjusted BIC.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 2
Mean Values for Each Model Fit Index in Which the Strength of a Factor Cross-Loading is Manipulated in the True Correlated Three-Factor Model

		Fitted model using MLR for continuous variables															
		Correlated factor								Bifactor							
Strength of cross-loading	Sample size	RMSEA	CFI	TLI	SRMR	AIC	BIC	SABIC	No. of models converged	RMSEA	CFI	TLI	SRMR	AIC	BIC	SABIC	No. of models converged
.1	500	.015	.995	.995	.025	8891.84	9043.57	8929.30	500	.015	.996	.995	.019	8890.06	9075.51	8935.85	432
	1,000	.009	.998	.998	.017	17791.56	17968.24	17853.90	500	.009	.998	.999	.014	17791.00	18006.94	17867.19	485
	2,000	.006	.999	.999	.013	35666.88	35868.51	35754.14	500	.005	.999	1.000	.010	35665.22	35911.66	35771.87	498
	3,000	.004	.999	1.000	.010	53485.74	53701.96	53587.58	500	.004	1.000	1.000	.008	53485.22	53749.50	53609.70	500
	4,000	.004	1.000	1.000	.009	71338.04	71564.63	71450.23	500	.004	1.000	1.000	.007	71336.72	71613.66	71473.85	500
	5,000	.004	1.000	1.000	.008	89152.64	89387.26	89272.86	500	.003	1.000	1.000	.006	89150.94	89437.70	89297.88	500
	10,000	.004	1.000	1.000	.006	178371.10	178630.60	178516.20	500	.003	1.000	1.000	.005	178365.90	178683.20	178543.30	500
	20,000	.004	1.000	1.000	.005	356937.50	357222.10	357107.70	500	.003	1.000	1.000	.004	356926.30	357274.00	357134.20	500
	40,000	.004	1.000	1.000	.004	714014.80	714324.30	714209.90	500	.002	1.000	1.000	.003	713992.40	714370.60	714230.80	500
.3	500	.018	.994	.993	.026	8922.83	9074.56	8960.29	500	.017	.995	.994	.020	8913.14	9098.58	8958.92	422
	1,000	.012	.997	.997	.019	17890.93	18067.60	17953.27	500	.011	.998	.997	.015	17879.51	18095.45	17955.71	468
	2,000	.009	.998	.998	.014	35817.24	36018.87	35904.49	500	.008	.999	.999	.011	35809.79	36056.23	35916.44	493
	3,000	.009	.999	.998	.012	53746.95	53963.18	53848.79	500	.007	.999	.999	.010	53737.52	54001.80	53861.99	499
	4,000	.009	.999	.998	.011	71684.89	71911.47	71797.08	500	.006	.999	.999	.009	71670.87	71947.81	71807.99	500
	5,000	.009	.999	.998	.010	89597.63	89832.25	89717.85	500	.006	.999	.999	.008	89579.57	89866.32	89726.51	500
	10,000	.009	.999	.998	.009	179200.90	179460.50	179346.10	500	.006	.999	.999	.007	179163.70	179481.00	179341.20	500
	20,000	.010	.999	.998	.008	358295.70	358580.20	358465.80	500	.007	.999	.999	.007	358219.40	358567.10	358427.30	500
	40,000	.010	.999	.998	.007	716792.70	717102.10	716987.70	500	.007	.999	.999	.006	716639.00	717017.30	716877.50	500
.5	500	.021	.993	.991	.027	8842.07	8993.80	8879.53	500	.018	.995	.993	.022	8817.78	9003.23	8863.57	382
	1,000	.015	.996	.995	.021	17777.78	17954.46	17840.12	500	.012	.998	.997	.017	17756.63	17972.57	17832.83	450
	2,000	.014	.997	.996	.017	35583.87	35785.50	35671.12	500	.011	.998	.997	.014	35560.12	35806.56	35666.77	494
	3,000	.014	.997	.996	.015	53369.23	53585.46	53471.07	500	.010	.999	.998	.013	53344.00	53608.28	53468.48	500
	4,000	.014	.997	.996	.014	71210.88	71437.46	71323.07	500	.010	.999	.998	.012	71176.84	71453.77	71313.96	500
	5,000	.014	.997	.996	.014	89049.68	89284.30	89169.91	500	.010	.999	.998	.012	89007.56	89294.32	89154.50	500
	10,000	.014	.997	.996	.012	178165.60	178425.20	178310.80	500	.010	.999	.998	.011	178082.60	178399.80	178260.00	500
	20,000	.014	.997	.996	.012	356459.90	356744.40	356630.00	500	.010	.999	.998	.011	356290.60	356638.40	356498.60	500
	40,000	.014	.997	.996	.011	712991.30	713300.80	713186.30	500	.010	.999	.998	.010	712655.80	713034.10	712894.20	500

		Fitted model using WLSMV for dichotomous variables															
		Correlated factor								Bifactor							
Strength of cross-loading	Sample size	RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged	RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged
.1	500	.016	.986	.985	.758	—	—	—	500	.014	.992	.999	.621	—	—	—	292
	1,000	.010	.993	.993	.729	—	—	—	500	.010	.995	.994	.615	—	—	—	356
	2,000	.005	.998	.999	.674	—	—	—	500	.005	.999	.999	.568	—	—	—	401
	3,000	.003	.999	1.000	.661	—	—	—	500	.004	.999	1.000	.563	—	—	—	440
	4,000	.003	.999	1.000	.662	—	—	—	500	.003	.999	1.000	.563	—	—	—	449
	5,000	.003	.999	1.000	.666	—	—	—	500	.003	.999	1.000	.565	—	—	—	467
	10,000	.002	1.000	1.000	.659	—	—	—	500	.002	1.000	1.000	.558	—	—	—	487
	20,000	.002	1.000	1.000	.669	—	—	—	500	.002	1.000	1.000	.565	—	—	—	500
	40,000	.001	1.000	1.000	.678	—	—	—	500	.001	1.000	1.000	.569	—	—	—	500
.3	500	.018	.985	.982	.771	—	—	—	500	.015	.990	.987	.628	—	—	—	297
	1,000	.010	.994	.993	.722	—	—	—	500	.009	.996	.995	.595	—	—	—	353
	2,000	.005	.998	.999	.674	—	—	—	500	.005	.998	.999	.569	—	—	—	430

(*table continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 2 (*continued*)

Strength of cross-loading	Sample size	Correlated factor								Bifactor							
		RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged	RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged
	3,000	.004	.999	1.000	.662	—	—	—	500	.004	.999	1.000	.556	—	—	—	452
	4,000	.003	.999	1.000	.667	—	—	—	500	.003	.999	1.000	.561	—	—	—	463
	5,000	.003	.999	1.000	.664	—	—	—	500	.003	1.000	1.000	.559	—	—	—	470
	10,000	.003	1.000	1.000	.679	—	—	—	500	.002	1.000	1.000	.567	—	—	—	486
	20,000	.003	1.000	1.000	.711	—	—	—	500	.002	1.000	1.000	.586	—	—	—	495
	40,000	.003	1.000	1.000	.763	—	—	—	500	.002	1.000	1.000	.610	—	—	—	500
.5	500	.017	.988	.986	.753	—	—	—	500	.014	.993	.991	.604	—	—	—	313
	1,000	.009	.995	.995	.704	—	—	—	500	.008	.997	.996	.583	—	—	—	378
	2,000	.005	.999	.999	.669	—	—	—	500	.005	.999	.999	.560	—	—	—	414
	3,000	.004	.999	.999	.672	—	—	—	500	.004	.999	1.000	.558	—	—	—	434
	4,000	.004	.999	.999	.675	—	—	—	500	.004	.999	1.000	.564	—	—	—	446
	5,000	.004	.999	.999	.685	—	—	—	500	.004	.999	1.000	.564	—	—	—	446
	10,000	.004	.999	.999	.712	—	—	—	500	.003	1.000	1.000	.581	—	—	—	471
	20,000	.004	1.000	.999	.777	—	—	—	500	.003	1.000	1.000	.618	—	—	—	491
	40,000	.004	1.000	.999	.895	—	—	—	500	.004	1.000	1.000	.692	—	—	—	500

Note. Dark grey shading reflects model superiority by fit. Light grey shading reflects instances in which the competing models are equivalent. Unshaded cells reflect model inferiority by fit. Number of free parameters for robust maximum likelihood (MLR) models: correlated factor (36) and bifactor (44). Number of free parameters for weighted least squares with adjusted means and variances (WLSMV) models: correlated factor (25) and bifactor (33). Degrees of freedom for correlated factor models (41) versus Bifactor (33). RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean residual; WRMR = weighted root-mean-square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion; SABIC = sample-size adjusted BIC.

ported the bifactor model when the correlated residual was moderate or large.

WLSMV. In most cases, all fit indices failed to correctly identify the correlated factors model as the true population-level model. Many fit indices provided ties between models, with fewer providing support for the bifactor model, and even fewer correctly providing support for the correlated factors model. WRMR favored the bifactor model in every cell; in small samples all indices incorrectly favored the bifactor model. RMSEA showed the most accurate performance, although it only correctly identified the correlated factors model in about one third of comparisons.

Models with a within-factor correlated residual. We evaluated the fit of correlated factors and bifactor models when the true correlated factor model was specified to include a correlated residual on *within*-factor indicators (see Figure 4). Table 4 presents the mean value of each global fit index across all sample sizes and levels of misspecification for both models (see Supplemental Materials Table 4 for *SD* values).

MLR. In stark contrast to our previous sets of results, RMSEA and TLI were more likely to correctly identify the correlated factor model across nearly all levels of within-factor correlated residuals and sample sizes. The CFI largely produced ties at small to moderate levels of model misspecification (r = .1 and .3), but correctly identified the correlated factors model when the model showed a large misspecification (r = .5). SRMR consistently favored the bifactor model incorrectly, as did AIC in most cases. BIC and SABIC correctly identified the correlated factor model, except for when sample sizes became large in the large misspecification condition.

WLSMV. All fit indices consistently failed to identify the correlated factors model as superior. Approximately half of the cells produced a tie and half incorrectly favored the bifactor model. At larger correlated residual values (r = .5), all indices incorrectly favored the bifactor model, with WRMR always supporting the bifactor model at any level of misspecification. Further, WRMR began to deteriorate as sample size increased in the moderate to large misspecification conditions, such that neither model provided an acceptable absolute fit to the data (i.e., WRMR $>$ 1.0).

Differences in fit index and information criteria values. Using Gignac's (2007) practical difference criterion of ΔTLI $\geq$ .010, we observed that improvement in TLI values for both the correlated factor and bifactor models never exceed this benchmark when using WLSMV, although the bifactor model's TLI values did approach this criterion in all misspecification conditions when sample size was small (N = 500 and 1,000; ΔTLI range: .001 to .009). When using MLR, a different pattern emerged such that the correlated factor model TLI value consistently exceeded the bifactor's (ΔTLI range: .016 to .024) when within-factor correlated residuals were .5 across each of our nine sample size conditions, indicating minimal impact of sample size on TLI in this context (Marsh et al., 2005). Consistent with WLSMV results, improvements in the bifactor's TLI values remained below ΔTLI $\geq$ .010 when estimated using MLR.

With regard to differences for the information criteria ($<$10; Raftery, 1995), our MLR results are consistent with the pattern of mean value results in that the bifactor model only met this criterion for ΔBIC when sample size was large ($N \geq$ 1,000) and between-factor correlated residuals or cross-loadings were specified at moderate to large levels. Notably, when *between-factor correlated*

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3
Mean Values for Each Model Fit Index in Which the Strength of a Between-Factor Correlated Residual is Manipulated in the True Correlated Three-Factor Model

		Fitted model using MLR for continuous variables															
		Correlated factor								Bifactor							
Strength of correlated residual	Sample size	RMSEA	CFI	TLI	SRMR	AIC	BIC	SABIC	No. of models converged	RMSEA	CFI	TLI	SRMR	AIC	BIC	SABIC	No. of models converged
.1	500	.017	.994	.994	.025	8861.55	9013.27	8899.01	500	.017	.995	.993	.019	8858.98	9044.42	8904.76	432
	1,000	.011	.997	.997	.018	17722.04	17898.72	17784.38	500	.012	.998	.997	.014	17720.37	17936.31	17796.57	483
	2,000	.008	.999	.998	.013	35488.21	35689.84	35575.46	500	.009	.999	.998	.010	35485.65	35732.09	35592.30	497
	3,000	.008	.999	.999	.011	53222.24	53438.47	53324.08	500	.008	.999	.998	.009	53221.26	53485.54	53345.74	500
	4,000	.007	.999	.999	.009	70996.65	71223.24	71108.85	500	.007	.999	.999	.008	70994.97	71271.90	71132.09	500
	5,000	.007	.999	.999	.008	88747.24	88981.86	88867.46	500	.007	.999	.999	.007	88745.00	89031.75	88891.94	500
	10,000	.007	.999	.999	.006	177529.50	177789.10	177674.70	500	.007	.999	.999	.005	177523.30	177840.50	177700.70	500
	20,000	.008	.999	.999	.005	355169.00	355453.60	355339.20	500	.008	.999	.999	.004	355154.40	355502.10	355362.30	500
	40,000	.008	.999	.999	.004	709302.30	709611.70	709497.30	500	.008	.999	.999	.004	709270.40	709648.60	709508.80	500
.3	500	.029	.988	.984	.026	8822.29	8974.02	8859.75	500	.029	.990	.983	.022	8806.08	8991.52	8851.87	423
	1,000	.025	.991	.988	.020	17702.15	17878.83	17764.49	500	.026	.992	.987	.017	17703.66	17919.60	17779.86	477
	2,000	.025	.992	.989	.015	35409.65	35611.29	35496.91	500	.025	.993	.989	.014	35396.14	35642.58	35502.79	497
	3,000	.024	.992	.989	.014	53105.31	53321.54	53207.16	500	.024	.993	.989	.013	53082.42	53346.70	53206.90	500
	4,000	.024	.992	.989	.013	70828.67	71055.26	70940.86	500	.024	.993	.989	.012	70796.32	71073.26	70933.45	500
	5,000	.024	.992	.989	.012	88539.84	88774.46	88660.06	500	.024	.993	.989	.012	88499.35	88786.10	88646.29	500
	10,000	.024	.992	.989	.011	177184.80	177444.40	177330.00	500	.024	.993	.989	.011	177100.20	177417.50	177277.70	500
	20,000	.024	.992	.989	.010	354519.30	354803.80	354689.40	500	.024	.993	.989	.011	354348.30	354696.00	354556.20	500
	40,000	.024	.992	.989	.009	708831.80	709141.20	709026.80	500	.024	.994	.989	.010	708488.40	708866.70	708726.80	500
.5	500	.047	.972	.963	.029	8824.12	8975.84	8861.58	500	.045	.979	.965	.028	8809.70	8995.14	8855.48	376
	1,000	.044	.975	.966	.023	17668.67	17845.34	17731.01	500	.042	.981	.968	.024	17634.58	17850.52	17710.78	452
	2,000	.043	.975	.967	.020	35414.27	35615.90	35501.53	500	.041	.981	.969	.022	35343.88	35590.32	35450.53	483
	3,000	.042	.976	.968	.018	53090.98	53307.21	53192.83	500	.040	.982	.970	.021	52989.01	53253.29	53113.48	491
	4,000	.042	.976	.968	.018	70791.52	71018.10	70903.71	500	.040	.982	.970	.021	70667.74	70944.68	70804.87	498
	5,000	.041	.976	.968	.017	88466.81	88701.43	88587.03	500	.040	.982	.970	.021	88310.61	88597.36	88457.55	498
	10,000	.041	.976	.968	.016	176782.60	177042.10	176927.70	500	.040	.982	.970	.020	176474.50	176791.70	176651.90	500
	20,000	.041	.976	.968	.016	353539.30	353823.80	353709.40	500	.040	.982	.970	.020	352915.70	353263.40	353123.60	500
	40,000	.041	.976	.968	.016	707295.60	707605.10	707490.70	500	.040	.982	.970	.020	706055.50	706433.70	706293.90	500

		Fitted model using WLSMV for dichotomous variables															
		Correlated factor								Bifactor							
Strength of correlated residual	Sample size	RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged	RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged
.1	500	.018	.983	.979	.781	—	—	—	500	.015	.988	.984	.645	—	—	—	297
	1,000	.012	.991	.990	.757	—	—	—	500	.011	.993	.991	.633	—	—	—	355
	2,000	.007	.997	.997	.713	—	—	—	500	.006	.997	.997	.603	—	—	—	416
	3,000	.005	.998	.998	.706	—	—	—	500	.005	.998	.998	.600	—	—	—	442
	4,000	.003	.999	1.000	.669	—	—	—	500	.003	.999	1.000	.572	—	—	—	456
	5,000	.003	.999	1.000	.666	—	—	—	500	.003	.999	1.000	.567	—	—	—	470
	10,000	.002	1.000	1.000	.662	—	—	—	500	.002	1.000	1.000	.560	—	—	—	491
	20,000	.001	1.000	1.000	.664	—	—	—	500	.002	1.000	1.000	.569	—	—	—	498
	40,000	.001	1.000	1.000	.684	—	—	—	500	.001	1.000	1.000	.588	—	—	—	500
.3	500	.017	.985	.982	.765	—	—	—	499	.014	.990	.987	.628	—	—	—	278
	1,000	.011	.992	.990	.756	—	—	—	500	.011	.994	.992	.630	—	—	—	342
	2,000	.007	.997	.997	.719	—	—	—	500	.008	.997	.996	.616	—	—	—	412

(*table continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3 (*continued*)

Strength of correlated residual	Sample size	Correlated factor RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged	Bifactor RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged
	3,000	.005	.998	.998	.703	—	—	—	500	.006	.998	.998	.601	—	—	—	425
	4,000	.004	.999	.999	.693	—	—	—	500	.005	.999	.999	.596	—	—	—	434
	5,000	.004	.999	.999	.692	—	—	—	500	.004	.999	.999	.596	—	—	—	457
	10,000	.003	.999	.999	.714	—	—	—	500	.004	.999	.999	.615	—	—	—	479
	20,000	.004	1.000	.999	.763	—	—	—	500	.004	1.000	.999	.664	—	—	—	496
	40,000	.004	1.000	.999	.853	—	—	—	500	.004	1.000	.999	.755	—	—	—	500
.5	500	.018	.983	.980	.782	—	—	—	499	.016	.988	.984	.644	—	—	—	267
	1,000	.012	.992	.991	.757	—	—	—	500	.011	.994	.992	.635	—	—	—	332
	2,000	.009	.996	.995	.745	—	—	—	500	.009	.996	.995	.634	—	—	—	392
	3,000	.007	.997	.997	.743	—	—	—	500	.008	.997	.996	.640	—	—	—	415
	4,000	.007	.998	.997	.746	—	—	—	500	.007	.998	.997	.646	—	—	—	429
	5,000	.006	.998	.998	.748	—	—	—	500	.007	.998	.997	.650	—	—	—	440
	10,000	.007	.998	.998	.822	—	—	—	500	.007	.998	.997	.722	—	—	—	465
	20,000	.007	.998	.998	.956	—	—	—	500	.008	.999	.998	.847	—	—	—	487
	40,000	.007	.999	.998	1.168	—	—	—	500	.008	.999	.998	1.051	—	—	—	500

Note. Dark grey shading reflects model superiority by fit. Light grey shading reflects instances in which the competing models are equivalent. Unshaded cells reflect model inferiority by fit. Number of free parameters for robust maximum likelihood (MLR) models: correlated factor (36) and bifactor (44). Number of free parameters for weighted least squares with adjusted means and variances (WLSMV) models: correlated factor (25) and bifactor (33). Degrees of freedom for correlated factor models (41) versus Bifactor (33). RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean residual; WRMR = weighted root-mean-square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion; SABIC = sample-size adjusted BIC.

residuals were .5, the bifactor outperformed the correlated factors model across seven sample sizes ($N \geq 2{,}000$). These improvements in BIC values also showed a marked decrease as sample size became larger (ΔBIC = -25.58 when $N = 2{,}000$ vs. ΔBIC = -1171.40 when $N = 40{,}000$). In contrast, we observed only three of 72 instances in which the correlated factor model was favored by BIC and did not met the criterion of less than 10, meaning BIC performed well. The degree of these differences in BIC values was also less extreme compared with when the bifactor was favored (ΔBIC range: -10.02 to -69.50). Finally, ΔAIC results also mirrored our mean value results as lower AIC values were generally associated with the bifactor model across nearly all types and levels of misspecification. These values also tended to consistently decrease by more than 10 when sample size was large ($\geq$20,000) and/or misspecification levels were moderate to high (ΔAIC range: -11.20 to -1240.10). There was no single instance in which the correlated factor model was favored by AIC and met the criterion of ΔAIC $<$10 (range: -0.52 to -9.69).

Percent accuracy. Using a threshold of $\geq$95% for strong performance across all 500 simulations in each misspecification condition, we examined the percentage of times that fit indices were able to correctly identify the correlated model as superior, incorrectly favored the bifactor model, and the percentage of ties (see Supplemental Materials Tables 5 through 12 for detailed results). For models estimated using WLSMV, none of our simulated results met the specified benchmark. For MLR, we observed a substantial decrease in the number of times fit indices were able to identify the correct model relative to the mean value results. Specifically, BIC and SABIC performed well across all study conditions until the underlying model contained misspecifications of .3, whereas AIC never met our accuracy threshold. Regarding MLR's fit indices (RMSEA, CFI, TLI, and SRMR), no index met the threshold when the underlying model contained a cross-loading or a between-factor correlated residual. However, RMSEA & TLI did perform well when a within-factor correlated residual was present, but only in samples $\geq$10,000 when the misspecification was small (.1) and when misspecifications were moderate to large (.3 and .5).

Vuong test. When the Vuong test for AIC and BIC was applied to our simulated continuous data, we found results that were consistent with the observed pattern in both our mean value and percent accuracy results (see Supplemental Materials Tables 13 through 16 for detailed results). The Vuong test for AIC was never able to identify the fitted correlated factors model as being closer to the true data generating model. For BIC, the Vuong test performed well when the underlying model contained small misspecifications (.1), However, test performance for BIC steadily decreased across all study conditions once model misspecifications reached .3.

Discussion

Several studies on the structure of psychopathology have applied both correlated factors and bifactor models. When these two models' fit indices have been directly compared, they have consistently favored a bifactor representation of observed comorbidity patterns (i.e., in 95% of studies). However, research from other fields suggests that traditional fit statistics are biased in favor of the bifactor model (e.g., Bonifay & Cai, 2017; Gignac, 2016;

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 4
Mean Values for Each Model Fit Index in Which the Strength of a Within-Factor Correlated Residual is Manipulated in the True Correlated Three-Factor Model

		Fitted model using MLR for continuous variables															
		Correlated factor								Bifactor							
Strength of correlated residual	Sample size	RMSEA	CFI	TLI	SRMR	AIC	BIC	SABIC	No. of models converged	RMSEA	CFI	TLI	SRMR	AIC	BIC	SABIC	No. of models converged
.1	500	.016	.995	.994	.025	8857.28	9009.00	8894.74	500	.018	.995	.993	.020	8852.42	9037.86	8898.21	448
	1,000	.013	.997	.996	.018	17750.13	17926.81	17812.47	500	.014	.997	.995	.015	17744.16	17960.10	17820.35	488
	2,000	.011	.998	.997	.014	35499.07	35700.70	35586.32	500	.012	.998	.997	.011	35500.02	35746.46	35606.67	499
	3,000	.011	.998	.997	.012	53213.54	53429.77	53315.39	500	.012	.998	.997	.010	53215.25	53479.53	53339.72	500
	4,000	.011	.998	.998	.010	70961.15	71187.73	71073.34	500	.012	.998	.997	.009	70963.08	71240.01	71100.20	500
	5,000	.011	.998	.998	.010	88712.63	88947.25	88832.85	500	.012	.998	.997	.008	88714.30	89001.06	88861.24	500
	10,000	.011	.998	.998	.008	177582.40	177842.00	177727.60	500	.012	.998	.997	.007	177583.90	177901.10	177761.30	500
	20,000	.011	.998	.998	.007	355159.40	355443.90	355329.50	500	.012	.998	.997	.006	355160.70	355508.40	355368.60	500
	40,000	.011	.998	.998	.006	710148.90	710458.40	710344.00	500	.012	.998	.997	.006	710149.70	710527.90	710388.10	500
.3	500	.036	.982	.976	.029	8817.20	8968.92	8854.66	500	.040	.983	.971	.025	8813.09	8998.53	8858.87	434
	1,000	.034	.984	.979	.024	17699.72	17876.40	17762.06	500	.038	.984	.974	.021	17700.83	17916.77	17777.03	477
	2,000	.033	.985	.980	.021	35438.35	35639.98	35525.61	500	.037	.985	.974	.019	35440.85	35687.29	35547.50	496
	3,000	.033	.985	.980	.020	53137.92	53354.15	53239.76	500	.037	.985	.975	.018	53136.50	53400.78	53260.97	500
	4,000	.033	.985	.980	.019	70877.86	71104.45	70990.06	500	.037	.985	.975	.018	70874.73	71151.67	71011.86	500
	5,000	.033	.985	.980	.019	88618.00	88852.62	88738.22	500	.037	.985	.975	.017	88614.33	88901.08	88761.27	500
	10,000	.032	.985	.980	.018	177136.30	177395.90	177281.50	500	.036	.985	.975	.017	177129.30	177446.50	177306.70	500
	20,000	.032	.985	.980	.017	354625.30	354909.90	354795.50	500	.036	.985	.975	.017	354607.20	354954.90	354815.10	500
	40,000	.032	.985	.980	.017	708923.50	709233.00	709118.60	500	.036	.985	.975	.016	708887.40	709265.70	709125.90	500
.5	500	.064	.951	.934	.037	8806.39	8958.12	8843.85	500	.074	.946	.910	.033	8805.63	8991.07	8851.41	444
	1,000	.061	.953	.936	.033	17666.37	17843.05	17728.71	500	.069	.950	.917	.030	17660.62	17876.56	17736.81	488
	2,000	.059	.954	.938	.031	35308.73	35510.36	35395.99	500	.067	.952	.921	.028	35295.66	35542.10	35402.31	500
	3,000	.058	.954	.938	.030	53072.47	53288.70	53174.32	500	.066	.953	.921	.027	53054.81	53319.09	53179.29	500
	4,000	.058	.954	.938	.030	70760.55	70987.14	70872.75	500	.066	.953	.922	.027	70737.75	71014.69	70874.88	500
	5,000	.058	.954	.938	.029	88436.33	88670.95	88556.56	500	.065	.953	.922	.027	88407.17	88693.93	88554.11	500
	10,000	.058	.954	.939	.029	176906.70	177166.30	177051.90	500	.065	.953	.922	.027	176848.80	177166.10	177026.30	500
	20,000	.058	.954	.938	.028	353690.80	353975.30	353860.90	500	.065	.953	.922	.026	353569.20	353916.90	353777.10	500
	40,000	.057	.954	.939	.028	707689.60	707999.10	707884.70	500	.065	.953	.922	.026	707455.40	707833.70	707693.80	500

		Fitted model using WLSMV for dichotomous variables															
		Correlated factor								Bifactor							
Strength of correlated residual	Sample size	RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged	RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged
.1	500	.017	.983	.981	.775	—	—	—	499	.014	.989	.986	.629	—	—	—	330
	1,000	.011	.992	.992	.743	—	—	—	500	.010	.995	.993	.621	—	—	—	364
	2,000	.006	.997	.998	.703	—	—	—	500	.007	.998	.997	.598	—	—	—	399
	3,000	.005	.998	.999	.692	—	—	—	500	.005	.999	.999	.591	—	—	—	444
	4,000	.004	.999	.999	.683	—	—	—	500	.004	.999	.999	.586	—	—	—	462
	5,000	.004	.999	.999	.684	—	—	—	500	.004	.999	.999	.590	—	—	—	474
	10,000	.003	1.000	1.000	.700	—	—	—	500	.003	1.000	.999	.598	—	—	—	496
	20,000	.003	1.000	1.000	.730	—	—	—	500	.003	1.000	.999	.627	—	—	—	500
	40,000	.003	1.000	1.000	.796	—	—	—	500	.003	1.000	.999	.689	—	—	—	500
.3	500	.019	.982	.977	.797	—	—	—	500	.016	.989	.984	.649	—	—	—	315
	1,000	.014	.991	.988	.775	—	—	—	500	.012	.993	.990	.645	—	—	—	361
	2,000	.011	.995	.993	.780	—	—	—	500	.011	.996	.993	.661	—	—	—	419

(*table continues*)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 4 (*continued*)

Strength of correlated residual	Sample size	Correlated factor								Bifactor							
		RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged	RMSEA	CFI	TLI	WRMR	AIC	BIC	SABIC	No. of models converged
	3,000	.010	.996	.994	.803	—	—	—	500	.010	.996	.994	.678	—	—	—	454
	4,000	.010	.996	.995	.831	—	—	—	500	.010	.997	.994	.703	—	—	—	452
	5,000	.010	.996	.995	.854	—	—	—	500	.010	.997	.995	.719	—	—	—	472
	10,000	.010	.997	.995	.991	—	—	—	500	.010	.997	.995	.835	—	—	—	490
	20,000	.011	.997	.995	1.227	—	—	—	500	.011	.997	.995	1.038	—	—	—	500
	40,000	.011	.997	.995	1.603	—	—	—	500	.011	.997	.995	1.352	—	—	—	500
.5	500	.023	.977	.970	.842	—	—	—	499	.018	.986	.979	.666	—	—	—	325
	1,000	.020	.984	.979	.863	—	—	—	500	.016	.990	.985	.687	—	—	—	343
	2,000	.019	.988	.984	.932	—	—	—	500	.017	.992	.987	.756	—	—	—	404
	3,000	.019	.989	.985	1.015	—	—	—	500	.017	.993	.988	.809	—	—	—	441
	4,000	.019	.989	.985	1.094	—	—	—	500	.017	.993	.988	.862	—	—	—	462
	5,000	.019	.989	.985	1.164	—	—	—	500	.017	.993	.989	.902	—	—	—	460
	10,000	.019	.989	.986	1.491	—	—	—	500	.017	.993	.989	1.145	—	—	—	497
	20,000	.019	.989	.986	2.003	—	—	—	500	.017	.993	.989	1.514	—	—	—	500
	40,000	.019	.989	.985	2.773	—	—	—	500	.017	.993	.988	2.080	—	—	—	500

Note. Dark grey shading reflects model superiority by fit. Light grey shading reflects instances in which the competing models are equivalent. Unshaded cells reflect model inferiority by fit. Number of free parameters for robust maximum likelihood (MLR) models: correlated factor (36) and bifactor (44). Number of free parameters for weighted least squares with adjusted means and variances (WLSMV) models: correlated factor (25) and bifactor (33). Degrees of freedom for correlated factor models (41) versus Bifactor (33). RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; SRMR = standardized root mean residual; WRMR = weighted root-mean-square residual; AIC = Akaike information criterion; BIC = Bayesian information criterion; SABIC = sample-size adjusted BIC.

Morgan et al., 2015; Murray & Johnson, 2013). The aim of the current study was to extend this line of research to common scenarios of mental disorder structural modeling by conducting a test of potential bias that is specific to the bifactor model of psychopathology. To do this, we simulated data from known population-level correlated factors models, with different types and degrees of model misspecifications, and ascertained to what extent fit indices identified a correlated factors model as data-generating model. Second, we conducted complementary tests of model equivalence by evaluating whether the bifactor model could perfectly reproduce the population covariance matrices implied by our each the four correlated factor models of interest (i.e., population-level simple structure, a between-factor cross-loading, a between-factor correlated residual, or a within-factor correlated residual).

General Fallibility of Fit Indices

Overall findings from sample model comparisons. Our study revealed different types of probifactor model fit index bias across a wide range of modeling scenarios. When misspecifications were present, we observed a frequent failure of all fit indices to identify the correlated factors model as the underlying population-level model. This was the case across estimators and sample sizes. In only one scenario—the use of the MLR estimator with a within-factor correlated residual—did three fit statistics (RMSEA, TLI, and BIC) provide consistent support for the correct correlated factors model from which the data were generated. However, even in that scenario, most other fit indices performed inconsistently, with some correctly identifying the correlated factors model when model misspecifications were small (AIC) *or* large (CFI), and when sample sizes were small *or* large, depending on the index. Furthermore, performance decreased for all fit statistics in each study condition when assessed according to both the percentage of correctly identified models and the Vuong test. This general pattern of fit indices' insensitivity to the population-level model was evident despite this study's conservative tests of model misspecifications (i.e., including only one misspecification at a time).

Overall, larger samples sometimes improved performance, but generally led to worse performance, especially for BIC and other information criteria. Further, BIC and SABIC demonstrated increasing probifactor bias as a function of increasing levels of model misspecification, a trend that was evident across all types of model fit comparisons. Previous research has shown that the bifactor model can better account for unmodeled complexity because of its less parsimonious structure relative to the correlated factors model (Murray & Johnson, 2013). Unfortunately, even for fit indices that have stronger penalties for model complexity and that favor parsimony (e.g., BIC and, to a lesser extent, AIC; Burnham & Anderson, 2004), these penalties proved inadequate to correctly identify a correlated factors model in simulated data with misspecifications likely to be present in psychopathology modeling scenarios (Greene & Eaton, 2016; Rodriguez-Seijas et al., 2015). Thus, although BIC, SABIC, and AIC penalize for model complexity in the form of greater numbers of freely estimated parameters, they do not penalize the bifactor model for additional forms of complexity that go beyond the number of parameters (i.e., functional form, defined as the way a model's equations specify

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

and combine parameters and variables; for an extended discussion see Bonifay & Cai, 2017). This shortcoming results in the possibility that more complex models may be deemed as "best fitting" by virtue of their ability to accommodate minor misspecifications that are not meaningful (i.e., high fitting propensity; Bonifay & Cai, 2017; Bonifay, Lane, & Reise, 2017; Reise et al., 2016). Thus, a balance is needed between goodness-of-fit and structural complexity. One way to achieve this balance is through the use of minimum description length (MDL)–based approaches (e.g., see Markon & Jonas, 2016), which take into account both the number of freely estimated parameters and the model's functional form to help researchers arrive at relatively simple models that provide adequate, albeit less than perfect, fit to the data (Bonifay & Cai, 2017).

There was a notable trend for some fit indices to be better at correctly selecting the correlated factors model when the population model contained a *within*-factor correlated residual (i.e., fit index performance improved as levels of model misspecification increased). In other words, the best way we found to identify the correct general model was to parameterize increasingly *in*correct models (e.g., TLI values for the correlated factor model only meaningfully exceeded the bifactor's when within-factor correlated residuals were .5). Conversely, this pattern also indicates that the bifactor model is especially good at accommodating misspecifications that span the specific factors (i.e., between-factor correlated residuals and cross-loadings), which can be captured by the general factor as common variance, although it is less effective at accounting for misspecifications that fall within specific factors. The issue of model misspecification is especially relevant to structural researchers in that the size of the *within*- and *between*-factor correlated residuals that we modeled appear quite reasonable and likely common. That is, correlations among residuals of the sizes *r*s = .1, .3, or .5 are not uncommon in structural investigations—particularly in studies that use multiple scales from a single measure to capture different constructs, which introduce shared method variance and increase the likelihood that corresponding indicators show correlated residuals. Correlated residuals may also emerge when disorders have similar symptoms. Although most studies do not report residual correlation matrices because adequate fit is obtained without modeling these correlations, Rodriguez-Seijas and colleagues (2015) found it necessary to include two correlated residuals in a correlated-factor model between (a) major depressive episode and dysthymia ($r = .7$) and (b) alcohol use disorder and drug use disorder ($r = .8$), likely because of similar indicator content. The size of these correlated residuals exceeded even the largest correlated residual in the present study, suggesting the size of model misfit may be greater in real-world scenarios.

The observed pattern of nonconvergence for bifactor models is not unique to the present study. Previous simulation studies have also reported a negative association between nonconvergence of bifactor models and sample size (Maydeu-Olivares & Coffman, 2006; Morgan et al., 2015; Murray & Johnson, 2013), especially in the context of binary indicators (Flora & Curran, 2004), although nonconvergence problems may also be attributed to the bifactor's orthogonal parameterization. That is, we did not allow the bifactor's specific factors to be correlated (as the model is commonly parameterized in psychopathology research), which might reduce convergence problems attributable to misspecification, low factor loadings, or small sample sizes. Nonetheless, such correlated specific factors not only violate the classic representation of an orthogonal bifactor model (Spearman, 1904), but also further complicate inferences drawn from the bifactor model by changing the meaning of *p* and suggesting the presence of additional common variance beyond *p* that is contributing to the specific factor's interrelatedness (Eid, Geiser, Koch, & Heene, 2017; Reise, 2012). Therefore, when considering whether to allow for oblique bifactor solutions, it is necessary to carefully weigh the advantages (e.g., higher convergence rates and better approximations of bifactor simple structure; Jennrich & Bentler, 2012) and disadvantages (e.g., oblique solutions are potentially less stable than orthogonal solutions; Lorenzo-Seva & Ferrando, 2018) before adopting an oblique modeling strategy.

Model nesting and equivalence. Discussion and identification of equivalent models is infrequent, despite the methodological significance of this issue in fitted model comparisons (Bentler & Satorra, 2010; Henley et al., 2006; Hershberger & Marcoulides, 2006; Maccallum et al., 1993; Raykov & Penev, 1999). Equivalent models arrive at the same model-implied covariance matrix and yield equivalent fit index values, but are not equivalent in structure (Hershberger & Marcoulides, 2006). Inferences based on data-model fit are severely restricted for equivalent models, because one model cannot be supported without all equivalent models being supported. Limitations are also placed on inferences about the causal relations implied by a hypothesized model as the size and direction of a model's structural relations, and its external correlates, depend on which equivalent model is selected.

We conducted tests of model equivalence by evaluating discrepancies between the population covariance matrices implied by each of the four correlated factor models and the bifactor model-implied covariance matrix when it was fit to a population covariance matrix. These supplementary analyses indicated that fit indices' favoring of the bifactor model was not always characterized by bias, because in one instance the bifactor model perfectly accommodated the population-level covariance matrix implied by the correlated factor model. Specifically, when the data-generating model was a *simple structure* correlated three-factor model it was perfectly reexpressed as a four-factor bifactor model. Thus, although these two models represent two distinct hypothetical casual structures with different substantive interpretations, they were also equivalent in terms of data-model fit. These findings underscore the idea that CFI, TLI, RMSEA, and SRMR fit-based comparisons between our three-factor correlated factor and four-factor bifactor models are not useful exercises in the unlikely situation that a population-level model contains no cross-loadings or correlated errors. In these cases, our simulation study with data generated from a pure correlated factor model implies that fit indices are not *biased* toward the bifactor model—they are favoring the bifactor model, albeit not unfairly. The information criteria, however, consistently selected the more parsimonious structure in our tests of model equivalence. As such, AIC and BIC appear to be more appropriate than CFI, TLI, RMSEA, and SRMR for situations where two, more and less restricted, models are bound to arrive at the same covariance matrix. These conclusions are reflected in the mean value and percent correct results in which RMSEA, CFI, and TLI mostly produced ties between the two models when the population-level correlated factor model was perfectly specified. In contrast, the information criteria helped identify improvements

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

in parsimony as they tended to select the simpler model under no misspecifications—an expected result given the penalty these indices impose on less restricted models.

In all other cases where the data-generating model contained cross-loadings and correlated errors, the bifactor model did *not* perfectly fit the population-level covariance matrix implied by the correlated factor model. Therefore, for simulation study conditions involving *misspecified* correlated factor models of interest, every measure of fit—CFI, TLI, RMSEA, SRMR, AIC, BIC, and SABIC—did show some susceptibility to probifactor bias. These instances of probifactor bias for fit index mean values and percent correct appear to be attributable to our bifactor model's additional fourth dimension erroneously accommodating misspecifications by (re)packaging these unmodeled complexities as common variance, even though they are not. This is not a desirable feature. Notably, several common fit indices tended to be biased in favor of the bifactor model when the population-level correlated factor model contained between-factor cross-loadings or trivial between- or within-factor correlated residuals ($r = .1$), a result consistent with prior work (Murray & Johnson, 2013; Yu, 2002). This was the case for RMSEA, CFI, TLI, WRMR, and SRMR, which are all measures that provide information on the lack of overall model fit to the data, not the degree to which a model is useful (Revelle & Wilt, 2013).

Overall, we take these findings to indicate that the bifactor model should not fit any better than the correlated factor model, other than its ability to better account for misspecifications, which is a statistical feature, rather than a substantive argument for utilizing a bifactor model. Regardless, in applied scenarios, fit indices are not capable of identifying the true data-generating mechanisms because these mechanisms are always unknown. Thus, the extent to which one of two equivalent, or nigh-equivalent, models can provide corroborating evidence for a theory depends on whether competing hypotheses posed by the equivalent alternative can be ruled out on theoretically substantive grounds (Hershberger & Marcoulides, 2006), such as whether the core assumptions underlying a structural representation are aligned with substantive considerations for classifying psychopathology.

Implications

It is rare to find a poor fitting model in most psychopathology studies, yet the use of fit indices and information criteria is likely to continue for the foreseeable future. This dilemma highlights the difficulty in achieving improvements in fit that are sufficiently large to warrant the estimation of additional parameters, as discussed in previous simulation studies (Gignac, 2016; Murray & Johnson, 2013). Even so, the literature remains unclear about how large of a difference constitutes "enough improvement," speaking to the complications with generating a golden rule that is applicable to every study and/or dataset. In the present study, better TLI values ($\geq .010$) for the correlated factor model only occurred in the presence of moderate to large *within-factor* correlated residuals. When judged by the criterion of $\Delta\text{AIC} < 10$, lower AIC values were associated with the bifactor model, across nearly all samples sizes and types/levels of misspecification. However, the correlated factors model was able to consistently outperform the bifactor by $\Delta\text{BIC} < 10$, except for when sample size was large and/or large misspecifications spanned across the specific factors, similar to our mean value results. These results add to previous discussions about these difference criteria for TLI and AIC/BIC values not being substantial enough for these statistics to overcome probifactor bias (Murray & Johnson, 2013). The observed trend for TLI may be interpreted as evidence that sufficiently large differences in fit are not to be expected (Gignac, 2016), especially in applied research where candidate models tend to all fit well, again highlighting the central role of subjective judgment in structural modeling practices.

The inherent bias of fit statistics in favor of the bifactor model is attributable to a combination of its model characteristics, which, as a consequence, increase its ability to accommodate unmodeled complexity in the data and/or its propensity to overfit by capitalizing on chance fluctuations in the data that arise from sampling error (i.e., irrelevant departures from the model; Bonifay & Cai, 2017; Bonifay et al., 2017; Reise et al., 2016). These two issues are most distinguishable in scenarios where our data generating model contained a cross-loading or correlated residual. There, we observed increased probifactor bias across all fit indices, including deteriorated performances of BIC and SABIC, as a function of increasing levels of misspecification (.5) and sample size. These results suggest bias in fit is attributable to model misfit rather than random noise, following similar lines of simulation research in which simple structure CFA models are misspecified (i.e., applied to data containing unmodeled cross-loadings and/or correlated residuals; Mansolf & Reise, 2017; Murray & Johnson, 2013). As a consequence, the mistaken inference of bifactor superiority seems to be driven by the general dimension's erroneous accommodation of misspecifications through capturing theoretically unexplained variance and repackaging it as common variance, even though it is not.

It is worth noting that in applied research, all CFA models are misspecified approximations of the data they summarize (Browne & Cudeck, 1993; Cudeck & Henly, 1991; MacCallum & Austin, 2000). That is, in contrast to the population models created in Monte Carlo simulation studies, real-world population model characteristics are always unknown (i.e., "there are no true models to discover"; Cudeck & Henly, 2003). This dilemma means that misspecifications—discrepancies between the tool doing the estimating and reality—are to be expected, at least if we assume that the population model is not a simple structure model. Therefore, when evaluating goodness of fit indices for a series of models with increasing degrees of fitting propensity, researchers should not overinterpret models that are more highly parameterized compared to simpler models. Measures of fit are more likely to favor structures with more parameters, especially when sample sizes are large (Cudeck & Henly, 1991). In this way, a correlated factor model could also be easily overinterpreted compared to a more restrictive unidimensional model in a circumstance where unmodeled complexities give rise to an unreliable factor based on repeated item content within indicators (i.e., when associations between some indicators are better explained by correlated residuals methods effects; Brown, 2015), instead of a substantively useful latent construct (e.g., externalizing). Importantly, our simulation scenarios were designed to represent misspecifications that should have no bearing on conclusions about the validity of a bifactor model's general factor of psychopathology. As such, our results should be understood as a reflection of a broader statistical issue: the fallibility of fit indices in judging the validity of structural represen-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

tations (Browne & Cudeck, 1993; Maydeu-Olivares, 2017; Preacher & Merkle, 2012). Utility considerations are most important for model selection decisions.

Another implication of our results extends to fit indices' frequent selection of the bifactor over the higher-order factor model. The higher-order and correlated factor models are equivalent when three lower-order factors are present (in which case both models may be, but are not always, nested within a bifactor model; Mansolf & Reise, 2017). Hence, our probifactor fit index bias results can be viewed as yet another instance in which the bifactor is erroneously selected as superior when it is fit to simulated data generated by a misspecified version of a higher-order model (Gignac, 2016; Maydeu-Olivares & Coffman, 2006; Molenaar, 2016; Morgan et al., 2015). As described by Mansolf and Reise (2017), this outcome is related to the bifactor's specific rank constraints that it makes on the data. Thus, comparisons of the unique rank constraints implied by these common measurement models, plus the degree to which they are violated, provide valuable insights into what distinguishes these models and how they may come to yield different, or even equivalent, fits to the data. For example, within each specific factor in the higher-order model, the ratio of specific factor loadings (for observed variables) to general factor loadings (indirect effects) are proportional (Yung, Thissen, & McLeod, 1999). Therefore, the bifactor and high-order factor models will also display equivalent fit to the data when the bifactor's ratio of general to specific factor loadings are also proportional. However, this proportionality condition for the bifactor's general to specific factor variance is rarely achieved in real-life modeling scenarios, leading to frequent observations of probifactor fit index bias due to these common violations of rank and proportionality constraints (Gignac, 2016; Mansolf & Reise, 2017; Molenaar, 2016). Model constraints provide a more technical alternative to understanding CFA models, with distinct implications for conceptualizing structure, and evidence is accruing for their contribution to our understanding of when and why the bifactor model will fit better relative to more parsimonious models (for a detailed exposition of these issues see Mansolf & Reise, 2017).

Interpretability. In order for a model to be a useful representation of diagnostic comorbidity,[3] it must also be falsifiable. To quote Sir Karl Popper, "falsifiability is the criterion of demarcation" between science and nonscience (Popper, 1963). Therefore, if a bifactor theory can easily accommodate all heretofore published states of affairs *based on fit*, then there is no observable difference between that theory's verisimilitude or falsity (for a discussion of problems associated with overreliance on fit indices in testing quantitative theories, framed in terms of the philosophy of science and the history of psychology see Roberts & Pashler, 2000). By this logic, theories that posit a general liability of psychopathology do not score well on the criterion of falsifiability given the high probability of such models outperforming more restrictive models on measures of fit in data across various samples, measures, and methods (e.g., Arias, Ponce, Martínez-Molina, Arias, & Núñez, 2016; Carragher et al., 2016; Caspi et al., 2014; Castellanos-Ryan et al., 2016; Laceulle et al., 2015; Lahey et al., 2015; Martel et al., 2017; Olino et al., 2014). This is not to say that the correlated factors model is clearly falsifiable as it also tends to fit most psychopathology data sets well; however, it is troubling if the bifactor model's features allow it to outperform simpler models, despite concerns about its relative lack of parsimony and questions about the interpretability of its latent factors (Bonifay et al., 2017; Widiger & Oltmanns, 2017). As such, the most pressing question currently is how to proceed with evaluating perhaps unfalsifiable explanations for the superiority of any competing structural representation[4] (i.e., explanations for a correlation matrix's positive manifold are difficulty to falsify; Van Der Maas et al., 2006).

When the criterion of substantive interpretability is used for adjudication of factor models, the bifactor model's general factor has various interpretations (Lahey, Krueger, et al., 2017), which are not generally as clear as those of the correlated factor model. For instance, one study demonstrated that the bifactor model's general factor of psychopathology correlated very strongly with the correlated factors models' internalizing and distress factors—more strongly than with the bifactor model's purported distress factor correlated with the correlated factors model's distress factor (Kim & Eaton, 2015). This raises the possibility that the bifactor's distress factor is being mislabeled because of a misinterpretation of the residual variance captured by this specific factor after controlling for the general factor (i.e., a nuisance factor that does not represent a meaningful construct; DeMars, 2013). Other studies have found the general factor to be closely linked to thought disorder symptoms such that increasing levels of general psychopathology may correspond to increased risk for experiencing disordered thought processes (Caspi et al., 2014). Hence, although the general factor is robustly related to external correlates and future outcomes (Lahey et al., 2012; Lahey, Krueger, et al., 2017), the utility of the each factor within the bifactor model may vary, just as the definition of this general factor is liable to vacillate from study to study.

Another reason the bifactor model's specific factors are often more difficult to interpret than the correlated factor internalizing-externalizing dimensions is attributable to substantially attenuated loadings (Caspi et al., 2014; Gomez, Stavropoulos, Vance, & Griffiths, 2018; Laceulle et al., 2015; Olino et al., 2014), which result in poor factor identification and irregular factor loadings (e.g., the distress factor tends to be a unipolar depression factor, due to low loadings on GAD; Greene & Eaton, 2017; Kim & Eaton, 2015; Lahey et al., 2012), and sometimes opposite effects (e.g., childhood externalizing problems being protective against future pure internalizing problems; Caspi et al., 2014). Other

[3] We view the purpose of these structural models as purely representational and aimed at informing current conceptualizations of how psychopathology is classified, assessed, treated, and researched as a result of the known limitations of our extant classification rubrics. In contrast, the National Institute of Mental Health's Research Domain Criteria (RDoC) are explicitly directed towards the purposes of mapping/identifying etiological components and processes. Because this discussion is beyond the scope of this article, we refer the reader to the work of Clark, Cuthbert, Lewis-Fernández, Narrow, and Reed (2017) for a recent exposition on the different utilities associated with classification models versus RDoC.

[4] For recent work that suggests such explanations can be falsified with developmental data, we refer readers to investigations by Kievit et al. (2017) and Hofman et al. (2018) on mutualism of cognitive abilities in latent change models. These studies show that G can be separated out from mutualism because only mutualism should feature paths from time1–ability1 to time2–ability2 (*coupling effects*; the higher intercept on ability1 at time1, the higher value of ability2 at time2). Such findings would not be expected under G.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

unexpected solutions are also common, such as when strongly indicated factors in the correlated factors model show an appreciable decrease in magnitude when modeled as specific factors in a bifactor model (e.g., distress and internalizing disorders in studies with adults, children, and adolescents; Gomez et al., 2018; Lahey et al., 2012), or vanish altogether when indicators simultaneously shift and strongly load on the general factor instead (e.g., diagnostic variables with previously strong loadings on the thought disorder factor in a correlated factor model only loaded on *p* in the bifactor; Caspi et al., 2014). This pattern of results contradicts one of the major reasons for considering a bifactor approach—modeling construct-relevant multidimensionality (Morin, Arens, & Marsh, 2016)—when indicators are *expected* to reflect both *p* and specific factor variance. Such inconsistencies have led to investigations aiming to clarify the conditions under which the bifactor model is justified by researchers' study objectives, as well as alternative approaches to defining these models a priori (for a close examination of anomalous results and suggestions for addressing application issues with bifactor models see Eid et al., 2017; Eid, Krumm, Koch, & Schulze, 2018; Heinrich, Zagorscak, Eid, & Knaevelsrud, 2018).

In the bifactor context, an internalizing dimension reflects the shared pathology among anxiety and mood disorders after partialing out what these conditions share with all other diagnoses, which is further complicated when some studies allow the bifactor model specific factors to correlate (Carragher et al., 2016; Caspi et al., 2014) and others do not (Gomez et al., 2018; Lahey et al., 2012, 2015), such that the meaning and external correlates of these dimensions tend to differ notably across studies. These findings demonstrate that specific factors from bifactor solutions are not isomorphic with their counterparts from correlated-factor models. Thus, there is a need for a clear understanding of the specific factors' properties in matters of substantive interpretation and, more broadly, the use of alternative statistics for evaluations of model quality and latent factors' reliabilities (e.g., explained common variance [ECV], construct replicability [*H*], and *omega* as seen in Gomez et al., 2018; Rodriguez, Reise, & Haviland, 2016a, 2016b).

Finally, we interpret the results of this study as critical of the use of fit indices to demonstrate the bifactor model of psychopathology offers a superior representation of psychopathology relative to other competing models—not that the bifactor model of psychopathology itself is without worth. Indeed, we believe the bifactor model likely has notable utility and believe other (non-fit-based) methods of examining and adjudicating models will support the bifactor model to some extent (Gomez et al., 2018; Lahey, Krueger, et al., 2017; Oltmanns, Smith, Oltmanns, & Widiger, 2018). Thus, although the bifactor approach to decomposing variance into common and unique sources might be difficult to interpret from some perspectives (e.g., clinically meaningful and interrelated constructs), it may fit well with other perspectives (e.g., psychiatric genetics, psychometric scale development, etc.) and show utility for various purposes (e.g., specific factors statistically control for general factor variance). When aims of modeling call for such forms of variance decomposition, the bifactor model is a helpful tool. The correlated factors model is also consistent with the general factor that presumably underpins factor intercorrelations at the level of broad spectra (Kotov et al., 2017). If the goal is to model a general factor *and* assess questions of specificity, researchers may find Goldberg's method helpful, which treats hierarchy as construct (Eaton, in press; Forbes et al., 2017; Goldberg, 2006; Kim & Eaton, 2017).

Moving forward. There are a range of underused approaches for evaluating the fit of competing models with varying levels of complexity. First, cross-validation methods, which can be employed for estimations of overall error, have been shown to be more useful for model selection than several fit indices (e.g., for a study based on both real and simulated data see McFarland, 2016). Cross-validation is especially helpful for addressing the problem of overfitting in covariance structures due to capitalization on chance variance/sample characteristics (Cudeck & Browne, 1983; McFarland, 2016); and, using a single sample, the quantity of overall error can also be approximated via the expected value of the cross-validation index (ECVI; Browne & Cudeck, 1989; Browne & Cudeck, 1993). Such practices improve precision for distinguishing among discrepancies due to errors of approximation (reflective of the difference between population vs. model-implied covariance matrices when a model is fit to the population covariance matrix) versus discrepancies due to errors of estimation (reflective of the difference between the sample vs. model-implied covariance matrices). In this context, the former would be indicative of problems with sampling error variability and/or the bifactor's capitalization on chance, whereas the latter would point to the bifactor's built-in capacity to better accommodate misspecifications relative to the correlated factors model. Second, the Vuong test for non-nested models compares models' fits to select which is likely closer to the underlying data generating model and therefore best (Merkle et al., 2016). The Vuong test is specific to information criteria and has shown to be useful for model selection in recent applications (Anderson et al., 2018). Third, explorations of data using exploratory methods are valuable because they can help researchers ascertain the degree to which confirmatory models may misfit the data (i.e., rather than through post hoc assessments of fit and modification indices; Reise, 2012), and can be further enhanced through the use of cross-validation samples. For example, factor structure stability can be evaluated by comparing results across various rotation criteria (Schmitt & Sass, 2011), and/or the extent to which factor structures shift across both exploratory and confirmatory methods (Goldberg & Velicer, 2006). Lastly, exploratory structural equation modeling is a powerful method for performing confirmatory tests of a priori latent structures, providing an attractive alternative to overly restrictive CFA frameworks (i.e., as a result of the emphasis on simple structure in CFA; Marsh, Morin, Parker, & Kaur, 2014).

Adoption of alternative methods for adjudicating between models also hinges on the need for additional information beyond whether a model's fit indices exceed a specified threshold (e.g., CFI $\geq$ .95). In this way, fit indices function as dichotomous decision-making heuristics akin to significance tests using $p < .05$, which researchers use as rules for whether they should continue pursuing a specific model. However, fit indices do not supply information directly relevant to each models' substantive interpretability nor the subsequent development of the competing theories accommodated by these models. As such, future studies may benefit from the incorporation of additional model statistics that can be used as tests of researchers' conceptualizations of the theoretical constructs to be modeled. After all, structural relations among latent constructs are what underlie decisions to fit a hy-

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

pothesized model to the data. One such statistic is the coefficient *H*, a reliability estimate of latent constructs in SEM that increases with the strength of each observed indicators' factor loading and, as a consequence, their relative contributions to defining these constructs (Hancock & Mueller, 2001). With regard to *p*, previous research has shown that it is reasonable to expect higher *H* values compared to the specific factors (Gomez et al., 2018), which tend to be less stable across samples. This highlights the importance of indexing measurement quality (i.e., relations between observed variables and factors in measurement models) when gauging the inferred latent constructs themselves. Measurement quality is especially important in light of a paradoxical relationship described by Hancock and Mueller (2011), wherein poor measurement quality is associated with better fit index performance for structural models (because of space limitations we refer the reader to these authors' article for details on how an additional modeling step can allow for evaluation of structural models independent of their measurement models). Taken together, such demonstrations of structural model quality can provide valuable information for hypothesis development and testing, while also broadening the range of possibilities for addressing the deleterious effects of model misspecification on fit index performance.

Lastly, after judging the acceptability of model fit and parameter estimates' strength and interpretability, models can then be subjected to more "risky" tests in the spirit of Meehl's (1978) Popperian assertion that "[a] theory is corroborated to the extent that we have subjected it to such risky tests; the more dangerous tests it has survived, the better corroborated it is." These tests are best conducted in the later stages of construct validation when substantive theory and the accumulation of evidence support a priori predictions about latent structures (Brown, 2015). For instance, much can be learned from continued examinations of competing models' associations with external criteria, multitrait multimethod analyses of both *p* and specific factors' construct validity, multiple-group analyses of the generalizability of these latent constructs to distinct demographic groups (e.g., Greene & Eaton, 2017; Tackett, Lahey, et al., 2013), and the development of comorbidity over time (e.g., Murray, Eisner, & Ribeaud, 2016).

Limitations

Similar to all Monte Carlo simulations, our findings may not generalize beyond the chosen set of conditions studied here. This is because the 180,000 models we fit to simulated data sets cover only a small part of the whole model parameter space. For instance, we did not manipulate model parameters beyond misspecifications (e.g., we did not manipulate number of indicators per factor or numbers of factors). We also did not test models where the strength of interfactor correlations varied within the correlated factors model nor the impact of multiple misspecifications being present simultaneously. Rather, we aimed to provide an illustration of fit index bias in situations that likely occur in many data sets, and hope that future studies will extend our manipulations in various directions. Prior simulation work has shown that the interpretability of common fit indices, such as RMSEA and CFI, is hindered by decreases in factor loadings and corresponding increases in unique variances, regardless of whether models are simple or complex (Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011). This is an important avenue for further investigation as ongoing efforts to assess differing aspects of these competing models are warranted to the extent that they are apropos of patterns of findings in applied research (e.g., lower loadings for the bifactor's specific factors relative to correlated factors). Another consideration is that, although we included a large number of sample size conditions, we did not investigate sample sizes with fewer than 500 participants, which may have produced differing results given the particularities of each fit statistic's relationship to sample size (Marsh et al., 2005). We made this decision because studies that undertook bifactor modeling have generally had sample sizes at least that large. Lastly, our simulated data were drawn from one population-level model, the correlated-factor model, and future research should test the extent to which competing models are able to fit data from alternative population-level structures, such as a true population-level bifactor structure. We did not examine true bifactor structure because we expected superior, unbiased performance of the bifactor model under such a scenario to be a foregone conclusion (Gignac, 2016; McFarland, 2016; Morgan et al., 2015).

Conclusions

This study was designed to inform future research on the *p*-factor directly. Our interpretation of the results provides cautionary evidence against overinterpretation of fit statistics for model selection decisions in general. Our study suggests that the numerous structural studies supporting the bifactor model over the correlated factors model based on fit should be interpreted as inconclusive, given our demonstration that all fit indices showed probifactor model bias. Thus, future psychopathology modeling scenarios should not rely solely on which model provides the best fit to the data. More important are considerations of data characteristics and clear statements about substantive comparison criteria (Cudeck & Henly, 1991), including the degree to which the theoretical assumptions underlying bifactor versus correlated-factor representations converge with theories for classifying psychopathology (i.e., misinterpretation can be avoided by taking the purpose of the model into account; Bonifay et al., 2017; DeMars, 2013; Maydeu-Olivares, 2017; Morgan et al., 2015; Murray & Johnson, 2013).

Along these lines, it must be recognized that *biases in relative fit indices do not invalidate favored models or imply that alternative models are more valid*. Converging lines of evidence support the utility of a general factor of psychopathology, as this dimension can be postulated even without reference to superior fit. Although alternative methods and statistics for model acceptability are available (e.g., Vuong test, cross-validation, coefficient *H*), the choice between correlated factors (or higher-order) and bifactor models can be made on practical grounds where the study aims justify the modeling approach. The former is most effective at identifying clusters of variables that are positively interrelated, whereas the latter is most effective at decomposing indicators' variance into that which is shared versus unique.

References

Achenbach, T. M. (1966). The classification of children's psychiatric symptoms: A factor-analytic study. *Psychological Monographs: General and Applied, 80*, 1–37. http://dx.doi.org/10.1037/h0093906

Achenbach, T. M., Conners, C. K., Quay, H. C., Verhulst, F. C., & Howell, C. T. (1989). Replication of empirically derived syndromes as a basis for

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

taxonomy of child/adolescent psychopathology. *Journal of Abnormal Child Psychology, 17,* 299–323. http://dx.doi.org/10.1007/BF00917401

Achenbach, T. M., Ivanova, M. Y., & Rescorla, L. A. (2017). Empirically based assessment and taxonomy of psychopathology for ages 11/2–90+ years: Developmental, multi-informant, and multicultural findings. *Comprehensive Psychiatry, 79,* 4–18. http://dx.doi.org/10.1016/j.comppsych.2017.03.006

Achenbach, T. M., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin, 131,* 361–382. http://dx.doi.org/10.1037/0033-2909.131.3.361

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317–332. http://dx.doi.org/10.1007/BF02294359

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Anderson, A. E., Marder, S., Reise, S. P., Savitz, A., Salvadore, G., Fu, D. J., . . . Bilder, R. M. (2018). Bifactor modeling of the positive and negative syndrome scale: Generalized psychosis spans schizoaffective, bipolar, and schizophrenia diagnoses. *Schizophrenia Bulletin, 44,* 1204–1216. http://dx.doi.org/10.1093/schbul/sbx163

Andrews, G., Goldberg, D. P., Krueger, R. F., Carpenter, W. T., Hyman, S. E., Sachdev, P., & Pine, D. S. (2009). Exploring the feasibility of a meta-structure for DSM-V and ICD-11: Could it improve utility and validity? *Psychological Medicine, 39,* 1993–2000. http://dx.doi.org/10.1017/S0033291709990250

Arias, V. B., Ponce, F. P., Martínez-Molina, A., Arias, B., & Núñez, D. (2016). General and specific attention-deficit/hyperactivity disorder factors of children 4 to 6 years of age: An exploratory structural equation modeling approach to assessing symptom multidimensionality. *Journal of Abnormal Psychology, 125,* 125–137. http://dx.doi.org/10.1037/abn0000115

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13,* 186–203. http://dx.doi.org/10.1207/s15328007sem1302_2

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246. http://dx.doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M., & Satorra, A. (2010). Testing model nesting and equivalence. *Psychological Methods, 15,* 111–123. http://dx.doi.org/10.1037/a0019625

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

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science, 5,* 184–186. http://dx.doi.org/10.1177/2167702616657069

Brown, T. A. (2015). *Confirmatory factor analysis for applied research.* New York, NY: Guilford Press Publications.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24,* 445–455. http://dx.doi.org/10.1207/s15327906mbr2404_4

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage Focus Editions, 154,* 136–136.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33,* 261–304. http://dx.doi.org/10.1177/0049124104268644

Carragher, N., Teesson, M., Sunderland, M., Newton, N. C., Krueger, R. F., Conrod, P. J., . . . Slade, T. (2016). The structure of adolescent psychopathology: A symptom-level analysis. *Psychological Medicine, 46,* 981–994. http://dx.doi.org/10.1017/S0033291715002470

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., . . . Moffitt, T. E. (2014). The p factor one general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science, 2,* 119–137. http://dx.doi.org/10.1177/2167702613497473

Castellanos-Ryan, N., Brière, F. N., O'Leary-Barrett, M., Banaschewski, T., Bokde, A., Bromberg, U., . . . the IMAGEN Consortium. (2016). The structure of psychopathology in adolescence and its common personality and cognitive correlates. *Journal of Abnormal Psychology, 125,* 1039–1052. http://dx.doi.org/10.1037/abn0000193

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41,* 189–225. http://dx.doi.org/10.1207/s15327906mbr4102_5

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Vol. 2). Hillsdale, NJ: Erlbaum.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 147–167. http://dx.doi.org/10.1207/s15327906mbr1802_2

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

Cudeck, R., & Henly, S. J. (2003). A realistic perspective on pattern representation in growth data: Comment on Bauer and Curran (2003). *Psychological Methods, 8,* 378–383. http://dx.doi.org/10.1037/1082-989X.8.3.378

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16–29. http://dx.doi.org/10.1037/1082-989X.1.1.16

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing, 13,* 354–378. http://dx.doi.org/10.1080/15305058.2013.799067

Eaton, N. R. (in press). The broad importance of integration: Psychopathology research and hierarchy as construct. *European Journal of Personality*.

Eaton, N. R., Keyes, K. M., Krueger, R. F., Balsis, S., Skodol, A. E., Markon, K. E., . . . Hasin, D. S. (2012). An invariant dimensional liability model of gender differences in mental disorder prevalence: Evidence from a national sample. *Journal of Abnormal Psychology, 121,* 282–288. http://dx.doi.org/10.1037/a0024780

Eaton, N. R., Krueger, R. F., Markon, K. E., Keyes, K. M., Skodol, A. E., Wall, M., . . . Grant, B. F. (2013). The structure and predictive validity of the internalizing disorders. *Journal of Abnormal Psychology, 122,* 86–92. http://dx.doi.org/10.1037/a0029598

Eaton, N. R., Krueger, R. F., & Oltmanns, T. F. (2011). Aging and the structure and long-term stability of the internalizing spectrum of personality and psychopathology. *Psychology and Aging, 26,* 987–993. http://dx.doi.org/10.1037/a0024406

Eaton, N. R., Rodriguez-Seijas, C., Carragher, N., & Krueger, R. F. (2015). Transdiagnostic factors of psychopathology and substance use disorders: A review. *Social Psychiatry and Psychiatric Epidemiology, 50,* 171–182. http://dx.doi.org/10.1007/s00127-014-1001-2

Eaton, N. R., South, S. C., & Krueger, R. F. (2010). The meaning of comorbidity among common mental disorders. In T. Millon, R. F. Krueger, & E. Simonsen (Eds.), *Contemporary directions in psychopathology: Scientific foundations of the DSM-V and ICD-11* (pp. 223–241). New York, NY: Guilford Press.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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

Eid, M., Krumm, S., Koch, T., & Schulze, J. (2018). Bifactor models for predicting criteria by general and specific factors: Problems of nonidentifiability and alternative solutions. *Journal of Intelligence, 6,* 42. http://dx.doi.org/10.3390/jintelligence6030042

Farmer, R. F., Seeley, J. R., Kosty, D. B., Olino, T. M., & Lewinsohn, P. M. (2013). Hierarchical organization of axis I psychiatric disorder comorbidity through age 30. *Comprehensive Psychiatry, 54,* 523–532. http://dx.doi.org/10.1016/j.comppsych.2012.12.007

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466–491. http://dx.doi.org/10.1037/1082-989X.9.4.466

Forbes, M. K., Kotov, R., Ruggero, C. J., Watson, D., Zimmerman, M., & Krueger, R. F. (2017). Delineating the joint hierarchical structure of clinical and personality disorders in an outpatient psychiatric sample. *Comprehensive Psychiatry, 79,* 19–30. http://dx.doi.org/10.1016/j.comppsych.2017.04.006

Forbush, K. T., & Watson, D. (2013). The structure of common and uncommon mental disorders. *Psychological Medicine, 43,* 97–108. http://dx.doi.org/10.1017/S0033291712001092

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time . . . Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment, 28,* 1354–1367. http://dx.doi.org/10.1037/pas0000275

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

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. http://dx.doi.org/10.1016/j.intell.2016.01.006

Goldberg, L. R. (2006). Doing it all bass-ackwards: The development of hierarchical factor structures from the top down. *Journal of Research in Personality, 40,* 347–358. http://dx.doi.org/10.1016/j.jrp.2006.01.001

Goldberg, L. R., & Velicer, W. F. (2006). Principles of exploratory factor analysis. *Differentiating normal and abnormal personality, 2,* 209–337.

Gomez, R., Stavropoulos, V., Vance, A., & Griffiths, M. D. (2018). Re-evaluation of the Latent structure of common childhood disorders: Is there a general psychopathology factor (p-factor)? *International Journal of Mental Health and Addiction*. Advance online publication. http://dx.doi.org/10.1007/s11469-018-0017-3

Grant, B. F., & Dawson, D. A. (2006). Introduction to the National Epidemiologic Survey on Alcohol and Related Conditions. *Alcohol Research & Health, 29,* 74–78.

Greene, A. L., & Eaton, N. R. (2016). Panic disorder and agoraphobia: A direct comparison of their multivariate comorbidity patterns. *Journal of Affective Disorders, 190,* 75–83. http://dx.doi.org/10.1016/j.jad.2015.09.060

Greene, A. L., & Eaton, N. R. (2017). The temporal stability of the bifactor model of comorbidity: An examination of moderated continuity pathways. *Comprehensive Psychiatry, 72,* 74–82. http://dx.doi.org/10.1016/j.comppsych.2016.09.010

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling, 25,* 621–638. http://dx.doi.org/10.1080/10705511.2017.1402334

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Srbom (Eds.), *Structural equation modeling: Present and future–Festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International.

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

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,* 319–336. http://dx.doi.org/10.1037/a0024917

Heinrich, M., Zagorscak, P., Eid, M., & Knaevelsrud, C. (2018). Giving G a meaning: An application of the bifactor-(s-1) approach to realize a more symptom-oriented modeling of the Beck Depression Inventory-II. *Assessment*. Advance online publication. http://dx.doi.org/10.1177/1073191118803738

Henley, A. B., Shook, C. L., & Peterson, M. (2006). The presence of equivalent models in strategic management research using structural equation modeling: Assessing and addressing the problem. *Organizational Research Methods, 9,* 516–535. http://dx.doi.org/10.1177/1094428106290195

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). Charlotte, NC: Information Age Publishing.

Hofman, A., Kievit, R., Stevenson, C., Molenaar, D., Visser, I., & van der Maas, H. (2018, February 28). *The dynamics of the development of mathematics skills: A comparison of theories of developing intelligence*. http://dx.doi.org/10.31219/osf.io/xa2ft

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2,* 41–54. http://dx.doi.org/10.1007/BF02287965

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Thousand Oaks, CA: Sage.

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

Jennrich, R. I., & Bentler, P. M. (2012). Exploratory bi-factor analysis: The oblique case. *Psychometrika, 77,* 442–454. http://dx.doi.org/10.1007/s11336-012-9269-1

Kievit, R. A., Lindenberger, U., Goodyer, I. M., Jones, P. B., Fonagy, P., Bullmore, E. T., . . . the Neuroscience in Psychiatry Network. (2017). Mutualistic coupling between vocabulary and reasoning supports cognitive development during late adolescence and early adulthood. *Psychological Science, 28,* 1419–1431. http://dx.doi.org/10.1177/0956797617710785

Kim, H., & Eaton, N. R. (2015). The hierarchical structure of common mental disorders: Connecting multiple levels of comorbidity, bifactor models, and predictive validity. *Journal of Abnormal Psychology, 124,* 1064–1078. http://dx.doi.org/10.1037/abn0000113

Kim, H., & Eaton, N. R. (2017). A hierarchical integration of person-centered comorbidity models: Structure, stability, and transition over time. *Clinical Psychological Science, 5,* 595–612. http://dx.doi.org/10.1177/2167702617704018

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., . . . Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology, 126,* 454–477. http://dx.doi.org/10.1037/abn0000258

Kotov, R., Ruggero, C. J., Krueger, R. F., Watson, D., Yuan, Q., & Zimmerman, M. (2011). New dimensions in the quantitative classification of mental illness. *Archives of General Psychiatry, 68,* 1003–1011. http://dx.doi.org/10.1001/archgenpsychiatry.2011.107

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Kramer, M. D., Krueger, R. F., & Hicks, B. M. (2008). The role of internalizing and externalizing liability factors in accounting for gender differences in the prevalence of common psychopathological syndromes. *Psychological Medicine, 38,* 51–61. http://dx.doi.org/10.1017/S0033291707001572

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

Krueger, R. F., & Eaton, N. R. (2015). Transdiagnostic factors of mental disorders. *World Psychiatry, 14,* 27–29. http://dx.doi.org/10.1002/wps.20175

Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology, 2,* 111–133. http://dx.doi.org/10.1146/annurev.clinpsy.2.022305.095213

Krueger, R. F., Tackett, J. L., & MacDonald, A. (2016). Toward validation of a structural approach to conceptualizing psychopathology: A special section of the Journal of Abnormal Psychology. *Journal of Abnormal Psychology, 125,* 1023–1026. http://dx.doi.org/10.1037/abn0000223

Laceulle, O. M., Vollebergh, W. A., & Ormel, J. (2015). The structure of psychopathology in adolescence replication of a general psychopathology factor in the TRAILS Study. *Clinical Psychological Science, 3,* 850–860. http://dx.doi.org/10.1177/2167702614560750

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,* 971–977. http://dx.doi.org/10.1037/a0028355

Lahey, B. B., Krueger, R. F., Rathouz, P. J., Waldman, I. D., & Zald, D. H. (2017). A hierarchical causal taxonomy of psychopathology across the life span. *Psychological Bulletin, 143,* 142–186. http://dx.doi.org/10.1037/bul0000069

Lahey, B. B., Rathouz, P. J., Keenan, K., Stepp, S. D., Loeber, R., & Hipwell, A. E. (2015). Criterion validity of the general factor of psychopathology in a prospective study of girls. *Journal of Child Psychology and Psychiatry, 56,* 415–422. http://dx.doi.org/10.1111/jcpp.12300

Lahey, B. B., Rathouz, P. J., Van Hulle, C., Urbano, R. C., Krueger, R. F., Applegate, B., . . . Waldman, I. D. (2008). Testing structural models of *DSM–IV* symptoms of common forms of child and adolescent psychopathology. *Journal of Abnormal Child Psychology, 36,* 187–206. http://dx.doi.org/10.1007/s10802-007-9169-5

Lahey, B. B., Zald, D. H., Perkins, S. F., Villalta-Gil, V., Werts, K. B., Van Hulle, C. A., . . . Poore, H. E. (2017). Measuring the hierarchical general factor model of psychopathology in young adults. *International Journal of Methods in Psychiatric Research, 27,* e1593. http://dx.doi.org/10.1002/mpr.1593

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48,* 936–949. http://dx.doi.org/10.3758/s13428-015-0619-7

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

Lorenzo-Seva, U., & Ferrando, P. J. (2018). A general approach for fitting pure exploratory bifactor models. *Multivariate Behavioral Research*. Advance online publication. http://dx.doi.org/10.1080/00273171.2018.1484339

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51,* 201–226. http://dx.doi.org/10.1146/annurev.psych.51.1.201

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114,* 185–199. http://dx.doi.org/10.1037/0033-2909.114.1.185

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

Markon, K. E., & Jonas, K. G. (2016). Structure as cause and representation: Implications of descriptivist inference for structural modeling across multiple levels of analysis. *Journal of Abnormal Psychology, 125,* 1146–1157. http://dx.doi.org/10.1037/abn0000206

Markon, K. E., Krueger, R. F., & Watson, D. (2005). Delineating the structure of normal and abnormal personality: An integrative hierarchical approach. *Journal of Personality and Social Psychology, 88,* 139–157. http://dx.doi.org/10.1037/0022-3514.88.1.139

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 225–340). Mahwah, NJ: Erlbaum.

Marsh, H. W., Morin, A. J., Parker, P. D., & 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,* 85–110. http://dx.doi.org/10.1146/annurev-clinpsy-032813-153700

Martel, M. M., Pan, P. M., Hoffmann, M. S., Gadelha, A., do Rosário, M. C., Mari, J. J., . . . Salum, G. A. (2017). A general psychopathology factor (P factor) in children: Structural model analysis and external validation through familial risk and child global executive function. *Journal of Abnormal Psychology, 126,* 137–148. http://dx.doi.org/10.1037/abn0000205

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

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods, 11,* 344–362. http://dx.doi.org/10.1037/1082-989X.11.4.344

McFarland, D. J. (2016). Modeling general and specific abilities: Evaluation of bifactor models for the WJ-III. *Assessment, 23,* 698–706. http://dx.doi.org/10.1177/1073191115595070

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834. http://dx.doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (2001). Comorbidity and taxometrics. *Clinical Psychology: Science and Practice, 8,* 507–519. http://dx.doi.org/10.1093/clipsy.8.4.507

Meehl, P. E. (2004). What's in a taxon? *Journal of Abnormal Psychology, 113,* 39–43. http://dx.doi.org/10.1037/0021-843X.113.1.39

Merkle, E. C., You, D., & Preacher, K. J. (2016). Testing nonnested structural equation models. *Psychological Methods, 21,* 151–163. http://dx.doi.org/10.1037/met0000038

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

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,* 2–20. http://dx.doi.org/10.3390/jintelligence3010002

Morin, A. J., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling, 23,* 116–139. http://dx.doi.org/10.1080/10705511.2014.961800

Murray, A. L., Eisner, M., & Ribeaud, D. (2016). The development of the general factor of psychopathology 'p factor' through childhood and adolescence. *Journal of Abnormal Child Psychology, 44,* 1573–1586. http://dx.doi.org/10.1007/s10802-016-0132-1

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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,* 407–422. http://dx.doi.org/10.1016/j.intell.2013.06.004

Muthén, B., & Muthén, L. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Olbert, C. M., Gala, G. J., & Tupler, L. A. (2014). Quantifying heterogeneity attributable to polythetic diagnostic criteria: Theoretical framework and empirical application. *Journal of Abnormal Psychology, 123,* 452–462. http://dx.doi.org/10.1037/a0036068

Olino, T. M., Dougherty, L. R., Bufferd, S. J., Carlson, G. A., & Klein, D. N. (2014). Testing models of psychopathology in preschool-aged children using a structured interview-based assessment. *Journal of Abnormal Child Psychology, 42,* 1201–1211. http://dx.doi.org/10.1007/s10802-014-9865-x

Olino, T. M., McMakin, D. L., & Forbes, E. E. (2018). Toward an empirical multidimensional structure of anhedonia, reward sensitivity, and positive emotionality: An exploratory factor analytic study. *Assessment, 25,* 679–690. http://dx.doi.org/10.1177/1073191116680291

Oltmanns, J. R., Smith, G. T., Oltmanns, T. F., & Widiger, T. A. (2018). General factors of psychopathology, personality, and personality disorder: Across domain comparisons. *Clinical Psychological Science, 6,* 581–589. http://dx.doi.org/10.1177/2167702617750150

Patalay, P., Fonagy, P., Deighton, J., Belsky, J., Vostanis, P., & Wolpert, M. (2015). A general psychopathology factor in early adolescence. *The British Journal of Psychiatry, 207,* 15–22. http://dx.doi.org/10.1192/bjp.bp.114.149591

Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London, UK: Routledge & Kegan Paul.

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

Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods, 17,* 1–14. http://dx.doi.org/10.1037/a0026804

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111–163. http://dx.doi.org/10.2307/271063

Raykov, T., & Penev, S. (1999). On structural equation model equivalence. *Multivariate Behavioral Research, 34,* 199–244. http://dx.doi.org/10.1207/S15327906Mb340204

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47,* 667–696. http://dx.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,* 818–838. http://dx.doi.org/10.1080/00273171.2016.1243461

Revelle, W., & Wilt, J. (2013). The general factor of personality: A general critique. *Journal of Research in Personality, 47,* 493–504. http://dx.doi.org/10.1016/j.jrp.2013.04.012

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17,* 354–373. http://dx.doi.org/10.1037/a0029315

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

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98,* 223–237. http://dx.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,* 137–150. http://dx.doi.org/10.1037/met0000045

Rodriguez-Seijas, C., Eaton, N. R., & Krueger, R. F. (2015). How transdiagnostic factors of personality and psychopathology can inform clinical assessment and intervention. *Journal of Personality Assessment, 97,* 425–435. http://dx.doi.org/10.1080/00223891.2015.1055752

Rodriguez-Seijas, C., Eaton, N. R., Stohl, M., Mauro, P. M., & Hasin, D. S. (2017). Mental disorder comorbidity and treatment utilization. *Comprehensive Psychiatry, 79,* 89–97. http://dx.doi.org/10.1016/j.comppsych.2017.02.003

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling, 21,* 149–160. http://dx.doi.org/10.1080/10705511.2013.824793

Schmitt, T. A., & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement, 71,* 95–113. http://dx.doi.org/10.1177/0013164410387348

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52,* 333–343. http://dx.doi.org/10.1007/BF02294360

Seeley, J. R., Kosty, D. B., Farmer, R. F., & Lewinsohn, P. M. (2011). The modeling of internalizing disorders on the basis of patterns of lifetime comorbidity: Associations with psychosocial functioning and psychiatric disorders among first-degree relatives. *Journal of Abnormal Psychology, 120,* 308–321. http://dx.doi.org/10.1037/a0022621

Silva, R., Scheines, R., Glymour, C., & Spirtes, P. (2006). Learning the Structure of Linear Latent Variable Models. *Journal of Machine Learning Research, 7,* 191–246.

Simms, L. J., Grös, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety, 25,* E34–E46. http://dx.doi.org/10.1002/da.20432

Slade, T., & Watson, D. (2006). The structure of common *DSM–IV* and ICD-10 mental disorders in the Australian general population. *Psychological Medicine, 36,* 1593–1600. http://dx.doi.org/10.1017/S0033291706008452

Snyder, H. R., Young, J. F., & Hankin, B. L. (2017). Strong homotypic continuity in common psychopathology-, internalizing-, and externalizing-specific factors over time in adolescents. *Clinical Psychological Science, 5,* 98–110.

Spearman, C. (1904). General Intelligence, objectively determined and measured. *The American Journal of Psychology, 15,* 201–292. http://dx.doi.org/10.2307/1412107

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25,* 173–180. http://dx.doi.org/10.1207/s15327906mbr2502_4

Tackett, J. L., Daoud, S. L. S. B., De Bolle, M., & Burt, S. A. (2013). Is relational aggression part of the externalizing spectrum? A bifactor model of youth antisocial behavior. *Aggressive Behavior, 39,* 149–159. http://dx.doi.org/10.1002/ab.21466

Tackett, J. L., Lahey, B. B., van Hulle, C., Waldman, I., Krueger, R. F., & Rathouz, P. J. (2013). Common genetic influences on negative emotionality and a general psychopathology factor in childhood and adolescence. *Journal of Abnormal Psychology, 122,* 1142–1153. http://dx.doi.org/10.1037/a0034151

Trull, T. J., & Durrett, C. A. (2005). Categorical and dimensional models of personality disorder. *Annual Review of Clinical Psychology, 1,* 355–380. http://dx.doi.org/10.1146/annurev.clinpsy.1.102803.144009

van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & van der Maas, H. L. (2017). What is the p-factor of psychopathology? Some risks of general factor modeling. *Theory & Psychology, 27,* 759–773. http://dx.doi.org/10.1177/0959354317737185

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

van der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113,* 842–861. http://dx.doi.org/10.1037/0033-295X.113.4.842

Vollebergh, W. A., Iedema, J., Bijl, R. V., de Graaf, R., Smit, F., & Ormel, J. (2001). The structure and stability of common mental disorders: The NEMESIS study. *Archives of General Psychiatry, 58,* 597–603. http://dx.doi.org/10.1001/archpsyc.58.6.597

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica, 57,* 307–333. http://dx.doi.org/10.2307/1912557

Waldman, I. D., Poore, H. E., van Hulle, C., Rathouz, P. J., & Lahey, B. B. (2016). External validity of a hierarchical dimensional model of child and adolescent psychopathology: Tests using confirmatory factor analyses and multivariate behavior genetic analyses. *Journal of Abnormal Psychology, 125,* 1053–1066. http://dx.doi.org/10.1037/abn0000183

Waszczuk, M. A., Zimmerman, M., Ruggero, C., Li, K., MacNamara, A., Weinberg, A., . . . Kotov, R. (2017). What do clinicians treat: Diagnoses or symptoms? The incremental validity of a symptom-based, dimensional characterization of emotional disorders in predicting medication prescription patterns. *Comprehensive Psychiatry, 79,* 80–88. http://dx.doi.org/10.1016/j.comppsych.2017.04.004

Watson, D. (2009). Differentiating the mood and anxiety disorders: A quadripartite model. *Annual Review of Clinical Psychology, 5,* 221–247. http://dx.doi.org/10.1146/annurev.clinpsy.032408.153510

Widiger, T. A., & Oltmanns, J. R. (2017). The general factor of psychopathology and personality. *Clinical Psychological Science, 5,* 182–183.

Wright, A. G., Krueger, R. F., Hobbs, M. J., Markon, K. E., Eaton, N. R., & Slade, T. (2013). The structure of psychopathology: Toward an expanded quantitative empirical model. *Journal of Abnormal Psychology, 122,* 281–294. http://dx.doi.org/10.1037/a0030133

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

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Los Angeles: University of California Los Angeles.

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,* 113–128. http://dx.doi.org/10.1007/BF02294531

Received February 8, 2018
Revision received March 10, 2019
Accepted March 12, 2019 ■