

This article is in press at
Psychological Methods.

Reassessing the fitting propensity of factor models

Wes Bonifay^{1,2}, Li Cai³, Carl F. Falk⁴, & Kristopher J. Preacher⁵

¹ University of Missouri

² Missouri Prevention Science Institute

³ University of California, Los Angeles

⁴ McGill University

⁵ Vanderbilt University

Corresponding author: Wes Bonifay bonifayw@missouri.edu

Supplemental materials: <https://osf.io/b8apk/>

Funding statement: Bonifay and Cai are funded by the U.S. Department of Education, Institute of Education Sciences, through Grant R305D210032.

Abstract

Model complexity is a critical consideration when evaluating a statistical model. To quantify complexity, one can examine fitting propensity (FP), or the ability of the model to fit well to diverse patterns of data. The scant foundational research on FP has focused primarily on proof-of-concept rather than practical application. To address this oversight, the present work joins a recently published study in examining the FP of models that are commonly applied in factor analysis. We begin with a historical account of statistical model evaluation, which refutes the notion that complexity can be fully understood by counting the number of free parameters in the model. We then present three sets of analytic examples to better understand the FP of exploratory and confirmatory factor analysis models that are widely used in applied research. We characterize our findings relative to previously disseminated claims about factor model FP. Finally, we provide some recommendations for future research on FP in latent variable modeling.

Keywords: factor analysis, structural equation modeling, fitting propensity, model evaluation

Reassessing the fitting propensity of factor models

Statistical models are routinely used in scientific research to represent and test theories about associations among the variables of interest. Models are not simply diagrams or statistical computing code or forecasting devices – they are also mechanisms whereby scientists reason about the world (Morgan, 2012; Muthukrishna & Heinrich, 2019).¹ To ensure that our models lead to sound inferences and reasoning, it is crucial that they are subjected to rigorous evaluation. Perhaps the most obvious model evaluation technique is to inspect goodness-of-fit: the degree to which the model fits the observed data. However, among several shortcomings of goodness-of-fit testing (see Roberts & Pashler, 2000; 2002), a particular limitation is that the selected model may have a tendency to fit well to many possible data patterns, and thus its good fit to the presently observed data is more likely a red herring than a confirmation of true insight. Preacher (2006) referred to this tendency as *fitting propensity* (FP) and demonstrated that FP is tied not only to the number of parameters in the model, but also to its functional form (i.e., the configuration of the parameters in the model). By ignoring FP in model evaluation, we run the risk of reducing the falsifiability of the theories and hypotheses that our models represent.

To the best of our knowledge, since Preacher (2006) coined the phrase, the methodological research on FP in latent variable modeling has comprised just a handful of studies: Bonifay and Cai (2017) and Ergin (2020) examined FP in the context of categorical data modeling, and Falk and Muthukrishna (2023) developed and demonstrated software for FP analysis of continuous data models. These previous researchers specifically set up their FP simulations to illustrate the presence of *configural complexity* by isolating it from *parametric*

¹ Crombie (1988) listed hypothetical modeling as one of the six styles of scientific thinking, along with mathematical postulation, experiment, taxonomy, statistical analysis, and historical derivation.

complexity, i.e., they varied the model structures while holding the number of parameters constant. This work provided a proof-of-concept for FP by demonstrating that structural equation and item response theory models may be subject to inherent complexity (Romeijn, 2017) that inflates fit statistics, reduces generalizability, and misdirects inference, and cannot be quantified by counting parameters (Myung et al., 2004).

Building on these initial efforts at establishing FP concepts, methods, and software, recent scholarship has demonstrated its practical importance in applied modeling research. As one example, the psychopathology literature has seen widespread adoption of the so-called *p*-factor theory of general psychopathology (Caspi et al., 2014). This theory is typically modeled using a bifactor configuration, with one general factor representing the shared variance among all psychopathology indicators in the observed data, and multiple specific factors representing additional shared variance within subsets of the indicators. Hundreds of *p*-factor studies have appeared in the decade since its “discovery” (Watts et al., 2024), mostly supported by evidence of the bifactor model’s good fit to psychopathology data. However, the bifactor model is known to have high FP (Bonifay & Cai, 2017; Murray & Johnson, 2013; Reise et al., 2023): It can bend to accommodate a wide range of highly dissimilar data patterns, including implausible patterns (Reise et al., 2016), so that its good fit is an artifact of the statistical model rather than a strong test of the *p*-factor theory (Bonifay et al., 2016; Bornovalova et al., 2020; Kan et al., 2024; Greene et al., 2019; 2022; 2023; Rosenström et al., 2019; 2021; van Bork et al., 2017; Waldman et al., 2023; Watts et al., 2019; 2020; 2023; 2024). As a direct result of this high FP, recurrent goodness-of-fit across studies has been broadly misinterpreted as evidence of successful replication of the *p*-factor theory (Bonifay et al., 2024), despite markedly inconsistent factor loadings (Watts et al., 2019) and specific factors with low reliability and zero or negative

loadings (Forbes et al., 2021; Constantinou & Fonagy, 2019), among other methodological and theoretical problems (Watts et al., 2024). In sum, the early conceptual work on FP led to the practical realization that the *p*-factor appears to be one of the “cherished models” that Preacher (2006) said “may have to be abandoned or replaced if their past successes can be ascribed more to FP than to any insight they lend into the process that actually generated the data” (p. 254).

Recognizing this practical value, Bader and Moshagen (2022) contributed to the nascent literature by investigating FP in the context of factor analysis. This was a sensible research aim; fit-based model selection is a widespread practice in factor analysis and other latent variable modeling, but as Dunn (2000) advised, “Only when different models have equal propensities to fit data can the extent to which they actually do fit the data be used to adjudicate between them” (p. 174). That is, the FP of competing models must be established, and found to be equal, prior to consulting their goodness-of-fit to the observed data. Bader and Moshagen’s work in this area involved fitting several widely used factor structures (e.g., unidimensional, correlated-factors, bifactor, and higher-order models) to three million data variance-covariance matrices that had been randomly and uniformly sampled from a given space of continuous data. The authors compared these models in terms of familiar structural equation modeling (SEM) fit statistics (e.g., RMSEA, TLI) and consequently made several definitive assertions about the FP of factor models and utility of particular fit statistics.

Overall, they concluded that “the fitting propensity of the models assessed was mostly driven by the different number of free parameters, whereas there was no evidence for additional differences in the functional form” (p. 12).² More specifically, and contrary to previous research (e.g., Murray & Johnson, 2013), Bader and Moshagen claimed “no evidence for a more flexible

² Page numbers are from the advance online version of Bader and Moshagen (2022).

functional form of bifactor models” and that RMSEA and TLI allow for “a fair comparison of goodness-of-fit between the bifactor model and typical competitors” (p. 13). We detail some additional concerns with this work below, in our Discussion section.

Bader and Moshagen’s (2022) emphasis on commonly used models and familiar model comparison procedures represents a needed addition to the scant FP literature. Preacher (2006) and Bonifay and Cai (2017), in particular, exposed differences in FP between models that had the same number of parameters, but would not represent competing theories in a model selection task typical of substantive psychological research. However, Bader and Moshagen’s (2022) assertions are undermined by several important issues [most notably, their claims about the number of parameters (as quoted above) were based on model comparisons that did not control for the number of parameters, as detailed in our Discussion section]. These issues warrant further analysis on the topic of FP in factor analysis modeling, and FP investigations more broadly.

In the remainder of this paper, we first consider a historical account regarding aspects of model complexity beyond just the number of parameters. We then present a series of simulation studies that are designed to illustrate the FPs of models that are commonly compared in applications of exploratory and confirmatory factor analysis. Following the presentation of our results, we contrast our findings with those of Bader and Moshagen (2022). Finally, we provide some recommendations for future research regarding FP with latent variable models.

Model complexity: Beyond the number of parameters

Statistical model complexity is often thought to correspond strictly to the number of parameters in the model, as evident in both scientific writing (e.g., “a complex model always has more parameters than a simple model;” Samani et al., 2019; p. 90) and mathematical formulations (e.g., the penalty for complexity in the AIC (Akaike, 1973) relative fit index: $2k$,

where k is the number of model parameters). We find at least three problems with this framing, especially when the number of parameters is used to form inferences about model FP.

First, that more free parameters will lead to better fit is a foregone conclusion. To use an illustrative example, for m data pairs, a polynomial regression model with $m - 1$ parameters will fit perfectly (i.e., $R^2 = 1.0$), and the model's goodness-of-fit will decrease as increasingly more parameters are removed or constrained. Early contributors to mathematical statistics knew this fact well. Neyman (1939) stated: "If a good fit may be attained only by introducing a great number of parameters, it usually means that the method of introducing those parameters is not very successful, and therefore it does not seem worthwhile to discuss in greater detail" (p. 53). This is precisely why many fit indices in SEM (e.g., RMSEA) adjust for parsimony by directly controlling for the degrees of freedom/number of parameters.

Second, we echo many methodologists in upholding generalizability as a primary goal of statistical modeling in psychological research (Myung, Pitt, & Kim, 2004; Yarkoni, 2022). Yet, the number of parameters alone is not an indicator of a model's generalizability, and in fact, models with more free parameters (i.e., greater parametric complexity) are *less* likely to generalize to future data. See, for example, Hitchcock and Sober (2004): "For statisticians, applied mathematicians, and other scientists who frequently construct models to make sense of data, it is a well-confirmed fact that more complex models often do poorly when predicting new data" (p. 11). Much earlier, Gini (1926) wrote, "[It] is well known that a series of terms is described by a curve of a given type all the more faithfully the greater the number of its parameters. But this does not mean that one curve with a greater number of parameters extrapolates better values beyond the limits of observation, or that using it in inserting the missing terms of a series we obtain results more approximate to reality" (p. 716). Many

commonly used SEM fit statistics, such as those used in routine model selection tasks, prioritize accommodation of the currently observed data over the accurate prediction of unseen but potential data, and in doing so, run the risk of selecting a model that will fail to generalize because it is overly sensitive to sample-dependent idiosyncrasies.

The third, and most important, point is that the number of parameters is only one source of model complexity. The model's functional form, i.e., how the parameters are configured in the model equation, is an additional source of complexity (Myung & Pitt, 1997; Pitt & Myung, 2002; Preacher, 2003; 2006). This issue also has a deep history in statistical modeling. For instance, Ferguson (1954) noted the importance of controlling for the number of parameters: "Thurstone regards the complexity of a test as the number of parameters which it involves ... Two tests may, however, have the same complexity in the Thurstone sense and yet may appear to differ markedly in complexity, or its opposite parsimony, in some intuitive sense" (p. 283). Yet, even by 1954, this idea was not new. Moore (1908) wrote:

"The impossibility of rigidly defining what is simple and what is complex has not escaped statisticians. It is quite possible that from a particular point of view the equation of one type of curve might be more simple than that of another type, and yet be more complex when viewed in another light. It might, for example, have fewer constants, and, from this point of view, be more simple than another equation with a greater number of parameters. But the evaluation of the constants in the first case might entail an extremely complex operation, while, in the latter case, no difficulty would be encountered" (p. 18).

Accordingly, a half-century of statistical research has focused on elucidating these distinct components of model complexity. Relative fit statistics such as the Akaike information criterion (Akaike, 1974) and the Bayesian information criterion (Schwarz, 1978) penalize for

parametric complexity, i.e., by discounting good fit if it comes at the expense of an overly large number of parameters. Information-theoretic statistics such as the stochastic information criterion (SIC; Rissanen, 1986), Fisher information approximation (Rissanen, 1996), normalized maximum likelihood (Rissanen, 2001), information complexity (Bozdogan, 2000), and others penalize for configural complexity, i.e., by discrediting good fit if it can be attributed to an overly flexible functional form. The SIC, for instance, characterizes the complexity that comes from the geometrical properties of the model local to the parameter, as measured by the Fisher information matrix, above and beyond the number of parameters. One can loosely think of the SIC as accounting for complexity by weighing each parameter according to its importance / informativeness / curvature (all similar concepts under Fisher information) when the model is fit to data.

To further complicate the issue, the number of parameters and their configuration may not be the only sources driving model complexity. Preacher (2006) grounded his reasoning in the principle of minimum description length (MDL; Grünwald, 2000). A complete description of any model-data fitting scenario would entail additional specifications that could potentially affect any statement of complexity (e.g., estimator choice, factor extraction method, Bayesian hyperparameter settings, and so on; see MacCallum & Tucker, 1991). In sum, characterizing model complexity is itself a complex subject that cannot be reduced to reliance on traditional measures or methods.

In sum, “we must never mistake the number of parameters in a model for its actual complexity” (Romeijn, 2017, p. 808). Given this admonition, we believe that further evidence of configural complexity and model FP is worthy of dissemination. To that end, we turn now to a set of studies inspired by Bader and Moshagen’s (2022) rationale that the methodological

literature would benefit from a better understanding of the FP of factor models that are commonly compared in psychological research. By specifying models with the same number of parameters, we were able to more clearly expose the complexity inherent in each model's configuration.

Methods

Models. We considered three sets of 16-variable factor models that are routinely applied in the substantive psychological literature, and that contain the same number of freely estimated parameters within each set. These criteria allowed us to disentangle the parametric and configural sources of complexity. To reiterate, the models described below were chosen expressly because a) they reflect real-world modeling specifications and b) they arrange the same number of parameters in different configurations; the results should not be taken as an endorsement or critique of any particular model, construct, study, or research team.

Set 1. Our first set comprised three bifactor models that differed in how the specific (or group) factors were configured, i.e., with three, four, or five specific factors (as illustrated in Figure 1). All three models included 48 freely estimated parameters. Our first model set is therefore intended to illustrate that bifactor models with the same number of parameters, but different configurations of specific factors will differ in their propensity to fit well to data.

Set 1 also corresponds to a common real-world model selection scenario. In particular, applications of exploratory bifactor (Schmid & Leiman, 1957; Jennrich & Bentler, 2011) and exploratory structural equation modeling (Morin et al., 2016) models often involve selection between bifactor structures with different numbers of specific factors. For example, in a study measuring childhood intelligence, Dombrowski and colleagues (2015) compared bifactor models with two, three, four, and five specific factors. Other bifactor vs. bifactor comparisons can be

found in the applied measurement of depression (Norton et al., 2013), intelligence (Lecerf, & Canivez, 2022), hypomania (Stanton et al., 2019), narcissism (Engyel et al., 2020), emotional/behavioral disorder (Lambert et al., 2018), attention-deficit/hyperactivity disorder (ADHD; Arias et al., 2016), “general psychopathology” (Gomez et al., 2019; Moore et al., 2020), well-being (Fadda et al., 2017), cognitive assessment (Dombrowski, 2014), suicidal beliefs (Bryan & Harris, 2019), sport psychology (Myers et al., 2014), fatigue (Yost et al., 2017), work performance (Giordano et al., 2020), and interpersonal relationships (Chen et al., 2022), to name a few. Note that the design matrices in these applications, as in any exploratory bifactor model, did not include fixed zero loadings; the varying specifications within each study therefore differed in degrees of freedom. To control for the number of parameters, our Set 1 models do include fixed zero parameters, thereby allowing us to investigate the configural complexity inherent in bifactor model comparisons involving different numbers of specific factors.

Set 2. Our second model set comprised two highly familiar measurement models: a (confirmatory) bifactor model and a two-factor exploratory factor analysis (EFA) model, both with 47 freely estimated parameters (as shown in Figure 2). This particular bifactor model includes a single indicator (y_1) that loads only on the general factor. Eid and colleagues (2017) also discussed this model, which they labeled as a “bifactor $S(I-1)$ model” (where the notation refers to a single indicator that does not load on any specific factor). They reasoned that “one measured variable serves as a marker or ‘gold standard’ measure for G [the general factor] and therefore does not load onto a specific factor,” noting also that “in many applications of the conventional bifactor approach one loading is missing on at least one specific factor” (p. 552). Zhang et al. (2021) also considered a similar structure, which they called an “augmented bifactor model.”

This type of specification has appeared in applied measurement of depression (Tibubos et al., 2021), quality of life (Gibbons et al., 2007), math literacy (Cai et al., 2011), perfectionism (Seong et al., 2021), general and specific intelligence (Feraco & Cona, 2022), crystallized intelligence (Watrin et al., 2022), working memory (Gaye, 2022), the “dark triad” personality (Wehner et al., 2021), job satisfaction (Zhang et al., 2021), job crafting (Schachler et al., 2019), shock anxiety (Ritzka et al., 2020), and misophonia (Remmert et al., 2021), and, presumably, other constructs. Regardless of the meaning or import of an indicator that loads only on the general factor in a bifactor model, our reason for selecting this model is simply its well-established use in substantive research.

That said, the two models in our second set are rarely, if ever, considered in model selection studies. There are at least three explanations for avoiding direct comparison of these models. First, the bifactor model is a confirmatory factor analysis (CFA) model that is intended to stipulate and test some a priori theory about the variables of interest, whereas the EFA model is mostly atheoretic in nature. Second, CFA and EFA, if used in the same study, are usually applied at different stages and thus their fit would not be directly compared. Third, as its name implies, EFA represents the least theoretically constrained solution that ought to fit any observed data better than a corresponding CFA model. Although it is possible to compare the goodness-of-fit of EFA and CFA models, such comparisons add very little value to substantive scientific inference. Thus, we specified the models in Set 2 to better address the differences between EFA and CFA model complexity, despite the fact that these model classes would not be contrasted in typical fit-based model selection scenarios.

Set 3. Our third set included two models with 42 free parameters: a correlated-factors model with five factors and a bifactor *S*-1 model (as presented in Figure 3). The correlated-

factors model is widely used and thus requires no further elaboration. However, we expect that some readers may not be familiar with the bifactor *S-1* model (which is distinct from the bifactor *S(I-1)* model in Set 2). This model was proposed by Eid and colleagues (2017), who reasoned that a “reference domain” can be formed by removing one specific factor from a bifactor model (thus the moniker “bifactor *S-1*”) and thereby improving the model interpretability. As depicted in Figure 3B, indicators y_1 to y_7 form the reference domain in our bifactor *S-1* model. In addition, the two specific factors in our model are correlated, which is uncommon in traditional, or “strict,” bifactor models (Reise et al., 2013). Eid et al. justify this allowance for correlated specific factors by emphasizing the importance of the reference domain, “These correlations indicate partial relationships between domains after accounting for variance that all domains share with the reference domain” (p. 550).

The purported improvement in interpretability has led to a proliferation of bifactor *S-1* models measuring dozens of broad and narrow psychological constructs, including depression (Heinrich et al., 2020), anxiety (Hoffman et al., 2021), mood disorder (Peterson et al., 2022), externalizing disorders (Thöne et al., 2021; Hand & Lonigan, 2022), psychopathy (Olderbak et al., 2021), “general psychopathology” (Heinrich et al., 2021; Wendt et al., 2022), ADHD and oppositional defiant disorders (Burns et al., 2020; Junghänel et al., 2020), intelligence (Eid et al., 2018), memory (Kofler et al., 2020), social skills (Panayiotou et al., 2022), self-efficacy (Backfisch et al., 2021), academic behaviors (Kittelman et al., 2021), perfectionism (Gäde et al., 2017), emotional intelligence (Simonet et al., 2021), forgiveness (Forster et al., 2020), activism and radicalism (Pavlović et al., 2022), boldness (Patrick et al., 2019), and circadian rhythms (Panjeh et al., 2022), among others.

Particularly relevant to the current analysis is Eid’s (2020) recommendation that “it is

advisable to apply both the multidimensional model with correlated first-order factors and the bifactor $S-1$ model” (p. 897). Many of the studies listed above followed this advice by fitting a correlated-factors model and a bifactor $S-1$ model to the same data for the purposes of fit-based model selection. Thus, the two models in our third set are not only used in substantive research – they are often directly compared within a single study.

Analysis. We used the *ockhamSEM* package (Falk & Muthukrishna, 2023) in R (R Core Team, 2012) to evaluate the fitting propensity of the seven models described above. To approximate the complete data space, we used the onion method (Lewandowski et al., 2009) to randomly and uniformly sample 10,000 correlation matrices from the ellipsope that encompasses all possible correlation matrices (Joe, 2006; Waller, 2017). We intentionally did not restrict the correlations to be positive; this option would have constrained our inferences to the plausible data subspace, but our concern was model performance in the complete data space. We elaborate on this choice in the Discussion section below.

We fit all seven models using the *lavaan* package (Rosseel, 2012) in R and a simulated sample size of $N = 1,000$. A known complication in FP analysis is that parameter estimation is plagued by non-convergence when data are randomly sampled from the complete space. Preacher (2006) noted that improper Monte Carlo solutions can be used to assess FP and increased the maximum number of maximum likelihood iterations to 2,000 (cf. the default maximum of 150 iterations in the *lavaan* package [Rosseel, 2012] in R) to allow the estimator extra time to reach convergence. Consequently, not all models converged, but fit indices were always produced and contextualized as “the best fit obtainable after 2,000 iterations” (p. 238). Bonifay and Cai (2017) followed this reasoning, but increased the number of iterations tenfold to garner further confidence in any non-converged results. In the present analysis, we modified the

default optimization settings in the *lavaan* package to include two relevant arguments:

`control=list(iter.max=1500)` increases the optimization iterations from the default of

150 to 1,500 (which can, of course, be set at even higher values) and

`optim.force.converged=TRUE` “pretend[s] the model has converged, no matter what”

(Rosseel, 2023; p. 56). In plain language, these adjustments gave the estimator ten times as long to find a solution, and then accepted the final guess as good enough to use for an investigation of FP.

To quantify and report FP, we retained the minimum of the maximum likelihood fit function (denoted f_{\min} in *lavaan*) from fitting each of the seven models to each of the 10,000 data sets. Lower f_{\min} values indicated relatively better model fit. To communicate our findings, we relied primarily on data visualization. For each model set, we created empirical cumulative distribution functions (ECDFs) to illustrate the proportions of all data sets (the y -axis) that yielded f_{\min} values at or below any given threshold of interest (the x -axis). In each figure, the leftmost ECDF curve indicates the model with highest fitting propensity.

On their own, ECDFs oversimplify FP analysis by obscuring any common and/or unique regions of the complete data space that are occupied by different models: Two models may fit well to the same proportions of the overall data, but each proportion may comprise meaningfully distinct data patterns. Thus, to better characterize FP, researchers should supplement their ECDFs with area-proportional Euler plots that depict regions of the data space that are “occupied” by each model (Pitt, Kim, & Myung, 2003; Myung, Pitt, & Kim, 2004; see the right panel of Figures 4-6 of the current work). For example, Bonifay and Cai (2017) found that two diagnostic classification models had overlapping ECDFs, but upon inspection of the corresponding Euler plot (which they called an “amoeba plot”), it became clear that these two

models occupied partially distinct regions of the complete data space. Of course, the same reasoning applies even when ECDFs do not overlap: Even when there is a more obvious difference between cumulative proportions, an Euler plot will still provide unique information about intersections, disjunctions, and relative complements among the regions occupied by each model, and will indicate the proportion of the space that is not fit by any of the models under consideration.

In this study, we used the *eulerr* package (Larsson, 2021) in R to create area-proportional Euler plots (Wilkinson, 2011) that provide additional insight into the role of functional form in factor model complexity. For a given fit index, such as f_{\min} , these plots display the overlapping and non-overlapping areas of the complete data space that are associated with each model. Larger ellipses indicate higher FP, each overlapping area indicates the particular subspace that can be fitted by multiple models, and each non-overlapping area indicates the particular subspace that can be fitted by a single model. In each Euler plot below, the depicted numerical values denote the proportion of all 10,000 correlation matrices for which our arbitrarily specified threshold of $f_{\min} < 4.0$ was achieved. A lower or higher threshold would affect the absolute size of the ellipses (making them smaller or larger, respectively), but would affect the ranking of their relative sizes only when the corresponding ECDFs intersect due to varying steepness, which did not occur in our results. In other words, our general findings about relative FP are not dependent on the choice of threshold.

The fitting propensity specifications, factor model estimation details, f_{\min} values, and plotting code are freely available at <https://osf.io/b8apk/>.

Results

Set 1 (Comparing bifactor models). Figure 4 presents the results from fitting the three

bifactor models in Set 1 to 10,000 16×16 correlation matrices that were randomly and uniformly sampled from the complete continuous data space. As reflected in the ECDFs in the plot on the left, Model 1A (a bifactor model with three specific factors) had a higher FP than Model 1B (a bifactor model with four specific factors), which itself had higher FP than Model 1C (a bifactor model with five specific factors). In sum, bifactor FP is inversely related to the number of specific factors.

The right side of Figure 4 presents the corresponding Euler plot. Model 1A met the arbitrary threshold of $f_{\min} < 4$ with 41.7% of all matrices and 9.9% were unique to Model 1A; Model 1B met the threshold with 35.0% of all matrices, 3.9% of which were unique to this model; and Model 1C met the threshold for only 29.6% of all matrices, with just 1.7% being unique to this model. In addition, we see that 21.2% of all matrices yielded $f_{\min} < 4$ with all three models, and that 49.7% of all matrices did not achieve this threshold with any of the models in Set 1.

Set 2 (Comparing CFA and EFA models). Figure 5 depicts the ECDFs and Euler plots of Models 2A (a “confirmatory” bifactor model) and 2B (an exploratory factor analysis model). There is a readily apparent distance between the ECDFs in the left panel of Figure 5: The EFA model produced $f_{\min} < 4$ in 43.5% of all matrices, while the bifactor model produced $f_{\min} < 4$ in 31.4% of all matrices. In other words, relative to the bifactor model, the EFA model reached our arbitrary fit threshold in over 1,200 more data sets.

This result can be further elaborated by inspecting the Euler plot on the right side of Figure 5. Aside from showing that the EFA region (the dashed outline) is larger than the bifactor region, this plot also indicates that both models achieved $f_{\min} < 4$ in a quarter of the data sets, while the EFA uniquely fit an additional 18.9% of the matrices. Perhaps more surprising is that

680 matrices (or 6.8% of the complete data space) yielded $f_{\min} < 4$ when fit with a “confirmatory” bifactor model, but not with an exploratory model.

Set 3 (Comparing correlated-factors and bifactor S-1 models). Figure 6 presents the findings from fitting all correlation matrices with the models in Set 3. The ECDFs on the left side of this figure show that the Models 3A (a correlated-factors model) and 3B (A bifactor S-1 model) were visually near indistinguishable in terms of their *cumulative* distributions. However, the Euler plot on the right illustrates that, although the regions occupied by each model are roughly equivalent in size, they are not completely overlapping. While 15.7% of all matrices returned $f_{\min} < 4$ with both models, 7.7% achieved $f_{\min} < 4$ only with the correlated-factors model (the solid outline), a separate 9.8% achieved $f_{\min} < 4$ only with the bifactor S-1 model.

Discussion

Statistical model evaluation should prioritize generalizability to future/unseen data over good fit to the observed data (Hitchcock & Sober, 2004; Pitt & Myung, 2002). Unfortunately, generalizability can be easily obscured by model complexity, which itself can be elusive when all models under consideration contain the same number of parameters. Our results correct some crucial misunderstandings and shed some light on previously unrecognized issues in the FP of factor analysis models. Below, we contextualize the results of our FP analyses, contrast our findings with those of Bader and Moshagen (2022), elaborate on the relationships among FP and goodness-of-fit, and consider some next steps in FP research.

Summary of Findings

Following previous research on model complexity (e.g., Myung, Pitt, & Kim, 2005; Preacher, 2006), we designed each model set to include different configurations of the same number of freely estimated parameters, thereby isolating the contribution of functional form to

FP. Across all sets, the regions of the complete data space that were occupied by each model varied in terms of size, location, or both size and location (Figures 4, 5, and 6). Thus, variability in FP cannot be attributed solely to the number of parameters.

Set 1. Our first set of models comprised three bifactor models, all of which contained $q = 48$ free parameters, but different numbers of specific factors. The results from fitting these models to 10,000 correlation matrices presented a novel finding about bifactor model FP. Whereas previous research established that the bifactor model is better than other structures at fitting data (e.g., Bonifay & Cai, 2017; Markon, 2019; Reise et al., 2016; Watts et al., 2020), the current study builds upon such findings by demonstrating that FP also varies across different bifactor specifications.

More specifically, our results suggest that bifactor models with fewer specific factors are predisposed to fit well, relative to bifactor models with more specific factors. An explanation for this pattern can be found by recognizing that the bifactor model is a particular form of hierarchical factor model (Cai & Houts, 2021). In such a model, the parameters associated with the secondary tier of factors account for any residual dependence that remains after the influence of the primary factor(s) has been removed. Consequently, a bifactor model with fewer specific factors (assuming all items load on at least one factor) will tend to cover a larger proportion of the residual covariances.

To make this issue more concrete, recall that the total number of unique covariances in a given matrix is equal to $k(k - 1)/2$, where k is the number of variables. Our simulation data included 16 items, resulting in a total of $16(16 - 1)/2 = 120$ residual covariance terms to be accounted for by the different models in Set 1. As shown in Figure 1, Model 1A included three specific factors of six, five, and five items, respectively, thereby covering $[6(6 - 1)/2] +$

$[5(5 - 1)/2] \times 2 = 15 + 20 = 35$ of the 120 covariances, or 29.17%. Model 1B included four specific factors of four items each, thereby covering $[4(4 - 1)/2] \times 4 = 24$ residual covariances, or 20.0% of the total number. Model 1C included five specific factors; four items loaded on the first specific factor, and three items loaded on each of the four remaining specific factors. This model therefore accounted for $[4(4 - 1)/2] + [3(3 - 1)/2] \times 4 = 6 + 12 = 18$ residual covariances, or 15.0% of the total. Had we fit a bifactor model with eight specific factors, then each would cover a single residual covariance, and thus the full model would account for $8/120$, or just 6.67% of all residual covariances. In sum, this pattern demonstrates that, all else (including the number of parameters) being equal, bifactor structures with fewer specific factors will cover a larger proportion of the total residual covariances. Such models overfit data by covering more residual covariances, thereby explaining away the residual associations that we can only assume are uniformly distributed throughout the data space.

The findings from Set 1 have implications for both methodology and practice. The methodological implication is that FP is a complicated issue that cannot be swept aside with conclusive declarations about whether certain models are (or are not) excessively flexible. The practical implication concerns applications of the exploratory bifactor model. In such research, goodness-of-fit is routinely used to select among bifactor models with different numbers of specific factors (e.g., Dombrowski et al., 2015). However, our results indicate that models with fewer specific factors may exhibit better fit simply because they are better at capturing residual covariances in any given data set and not because they reflect some meaningful, interpretable pattern in the data. Researchers who wish to apply a series of exploratory bifactor models must acknowledge that the best fit does not necessarily correspond to the most generalizable structure (see also Greene et al., 2019; Montoya & Edwards, 2021).

Set 2. Our second set of models included a “confirmatory” bifactor model and an EFA model, both containing $q = 47$ parameters. Despite having the same number of parameters, these models differ in their intended uses: The bifactor model is a theory-driven model that is intended to test the user’s a priori expectations about the relationships among the items, whereas the EFA model is a mainly data-driven model that is intended to uncover the optimal loading pattern, regardless of any user expectations other than the number of factors.

Although the EFA model is overall better at fitting data, the bifactor model is undeniably flexible: The Euler plot on the right side of Figure 5 shows that 6.8% of all matrices (i.e., 680 data sets) were better fit by the bifactor model than by the EFA model [which replicates the same finding by Bonifay and Cai (2017)]. That is, a purportedly theory-driven bifactor model is so accommodating that it may, on occasion, outperform a data-driven exploratory model. This prospect is particularly concerning because excess flexibility can make a model “so weak that there is no way to find evidence either for or against it” (Wexler, 1978, p. 346). Whether the bifactor model has reached that degree of flexibility remains to be seen; but for the time being, results from our second analysis suggest that bifactor modeling could very well be considered more of an exploratory technique than a theory-testing one.

Set 3. Our third analysis compared the FP of a correlated-factors model and a bifactor $S-1$ model, both with $q = 42$ free parameters. The take-home message from this analysis is that ECDFs obscure an important nuance of model complexity: Competing models that appear to fit data equally well may fit distinct regions of the data space. In other words, superimposed ECDFs do not necessarily mean that the models are indistinguishable. Supplementary information, such as that provided by Euler plots, is needed if one wants to better understand the unique complexities of competing models.

For Set 3, the equivalence in overall FP allows for a fair comparison of the goodness-of-fit of Model 3A (five correlated factors) and Model 3B (bifactor $S-1$) to an observed data set (Dunn, 2000). That is, because these models have the same a priori likelihood of fitting well to any given data set from the complete space, an applied researcher can properly compare their ability to fit well to one particular data set. The Euler plot in Figure 6 adds context by approximating how likely it is to obtain each possible outcome. For example, a researcher might apply both models to a single data set and observe four outcomes: Neither model fits well (which, according to Figure 6, occurred in ~66.9% of all possible data sets); both models fit well (~15.7%); Model 3A fits well, but Model 3B does not (~9.8%); or Model 3B fits well, but Model 3A does not (~7.7%). The last two outcomes are the most unlikely, so observing either will give the researcher greater confidence that the better-fitting model is capturing something distinctive and meaningful (Roberts & Pashler, 2000). If the observed data set falls in the overlapping region of the Euler plot, then the choice between Models 3A and 3B becomes more of a traditional model selection task: Given that both models fit well and have equivalent FPs, which one is better in terms of interpretability, validity, parameter estimates, and so on? See Preacher and Yaremych (2023) for a review of model selection criteria in SEM.

It is also worth noting that the average fit can misrepresent the FP. This observation complicates interpretation in at least two ways. Most importantly, the average reduces and obscures the informative differences between the shapes of the ECDFs; it is an oversimplification that ignores meaningful variation in FP. Accordingly, previous FP researchers drew their inferences from the proportion of fit statistics at or below a certain threshold (e.g., $\text{RMSR} \leq .08$ in Preacher [2006]; $Y2/N \leq .05$ in Bonifay and Cai [2017]). However, when ECDFs are non-uniform *and* overlapping, then a simple proportion is also insufficient. This is

because rankings of model FP will be contingent on the particular threshold values that are selected for calculating that proportion. For this exact reason, Preacher wrote, “Even in the absence of a benchmark criterion value for good fit, nonoverlapping CDFs are sufficient to illustrate that two or more models differ in terms of FP” (p. 238).

Regarding Bader and Moshagen (2022)

As mentioned earlier, Bader and Moshagen (2022; hereafter B&M) recently examined FP through a practical lens, focusing on several factor structures that are commonly used in applied model comparisons. Their primary motivation was a critique of the extant literature on FP: The previously “contrasted instances did not necessarily represent models that are frequently compared during theory-guided model selections” (p. 5). To address this gap, B&M conducted two simulation studies. The first compared 25 EFA and CFA models with varying numbers of indicators and factors, and the second compared 14 CFA models with varying numbers of indicators in configurations that routinely appear in the applied literature: unidimensional, correlated-factors, bifactor, and higher-order structures. In both studies, they fit each model to 3,000,000 variance-covariance matrices that were sampled at random from a subspace of the complete data, using the default *lavaan* settings for estimation convergence and optimization. Here, we briefly summarize their claims, discuss the implications of accepting their claims, contrast their findings with our own, and examine their methodology for sources of these discrepant outcomes.

Claims. Based on their simulation study results, B&M made several claims about the FP of factor models. In general, they argued that model complexity can be fully quantified by counting parameters. They expressed this point in statements such as, “[The] difference [in FP] was entirely explained by the difference in the number of free parameters” (p. 13); “[FP] was

mostly driven by the different number of free parameters, whereas there was no evidence for additional differences in the functional form” (p. 12); and “[T]he number of free parameters represents a valid indicator of the degree of fitting propensity of commonly used types of CFA and EFA models” (p. 12).

Beyond these general conclusions about the number of parameters, B&M also reported specific findings about factor model complexity. They claimed that bifactor model FP should be attributed only to its number of parameters [e.g., “[T]here was no evidence for a more flexible functional form of bifactor models, indicating that when accounting for the number of free parameters ..., a fair comparison of goodness-of-fit between the bifactor model and typical competitors ... can be made.” (p. 13)]. Second, B&M claimed that EFA models do not have more flexible functional forms compared to CFA models [e.g., “We did not find evidence for a more flexible functional form of EFA as compared to CFA models beyond their higher number of free parameters” (p. 7)].

Implications. As of writing, B&M’s assertion that “considering the concept of fitting propensity ... can readily be achieved by relying on traditional fit indices accounting for the number of free parameters” (p. 13) has already been used to defend model selection in a number of published studies. For example, citing “current recommendations (Bader & Moshagen, 2022) ... $CFI \geq .97$, $NNFI \geq .97$, $RMSEA \leq .05$, and $SRMR \leq .05$,” Schroeders et al. (2024) state, “Our evaluation of model fit was not based on model fit indices alone but also took into account model complexity” (p. 7). Similar citations appear in Gnambis and Schroeders (2022), Mascia et al. (2023), and Garcia-Fernández et al. (2024), the latter noting specifically that fit indices like RMSEA “fully account for the greater flexibility of the bifactor model and thereby provide an adequate control of fitting propensity (Bader & Moshagen, 2022)” (p. 4). Based on such

statements, it appears that B&M have given researchers a false sense of security in the use of traditional fit statistics for model selection, encouraging them to make decisions with no concern for the inherent complexity that influences model fit. Further, B&M's claim of "no evidence" of bifactor flexibility will likely continue to be cited as justification for potential misuse of the bifactor model in the future, leading to many of the same problems encountered in the p -factor literature discussed earlier. Our overall concern is that B&M's claims, if unchecked, may result in further proliferation of weakly generalizable factor models, slowing progress in psychology and allied disciplines.

Contrasting findings. In our analysis, we explored the same research question as B&M – how does FP affect commonly used factor analysis models? Our results contradict several of B&M's conclusions. First, in the historical review from our introduction, we demonstrated that the number of parameters cannot provide a complete picture of model complexity. All three of our simulation studies underscore this point: Without controlling for the number of parameters, there is no way to isolate the degree to which FP is driven by the functional form. B&M did not find evidence for or against functional form complexity because their research design disallowed the production of such evidence.

Second, our analysis of competing bifactor structures (simulation Set 1) contextualized B&M's declaration of "no evidence for a more flexible functional form of bifactor models" by showing that bifactor model flexibility varies even within competing bifactor models. In fact, our Model 1B (a bifactor model with four specific factors) is the exact same structure that B&M used as the basis for their conclusions about bifactor FP. If they had modified their bifactor structure, either by adding/removing specific factors or modifying the number of items per specific factor,

then they likely would have detected non-ignorable differences in FP within a single family of practical and frequently applied models.

Third, our analysis of EFA and CFA complexity (simulation Set 2) yielded results that are contrary to B&M's claim of "no differences in the functional form of these [bifactor and exploratory two-factor] models" (p. 12). Rather, when controlling for the flexibility afforded by the number of parameters, the distinct functional forms of these models do, in fact, result in different FPs. As anticipated, and in opposition to B&M's claim about EFA and CFA complexity, the EFA form is, overall, inherently better at fitting data.

Methodology. The general and specific claims and contrasting findings described above can be attributed to several methodological issues in B&M's research design. In short, B&M's discussion (p. 11 ff.) sometimes extrapolated beyond the inferences that could be drawn from their specific simulation design and data analysis. We have identified four problems with their design, analysis, and presentation of results.

First, B&M followed previous FP research (Falk & Muthukrishna, 2023; Preacher, 2006) by restricting the complete data space to the plausible data space by only considering positively correlated data matrices. Or, as they put it, "we limited the data space to variance-covariance matrices in the positive manifold" (p. 6). Yet, neither their Results nor their Discussion mention this limitation or characterize their findings as conditional on the plausible data space. This omission is particularly glaring in their subsection on the bifactor model, given that a bifactor model does have a tendency to fit well due to its ability to fit data from outside the plausible data space. B&M cited Reise and colleagues (2016), who directly addressed this very issue: "There are many reasons why a bifactor model may provide a better fit. One of the worrisome reasons is that it can better accommodate implausible, possibly invalid, response patterns" (p. 834). A study

aimed at making conclusive statements about models' FPs should be designed to allow those models to express that flexibility beyond the plausible space.

A second methodological issue concerns B&M's decision to draw inferences about FP only when the maximum likelihood estimator converged under typical/default settings (which follows Falk and Muthukrishna (2023), but is not in accordance with earlier work on FP). Specifically, they fitted all models to 1 million matrices in each of three test length conditions and then analyzed only those results for which all models converged. They do not report this final number for Study 1, but for Study 2, this resulted in discarding ~330,000 cases (33%) from the 9-item condition, ~620,000 (62%) from the 16-item condition, and 780,000 (78%) from the 30-item condition. This decision has the same effect as relying on the plausible space: It limits the definition of FP and requires that the results should be qualified as dependent on typical convergence criteria.

To be clear, whether or not to restrict the data space and/or accept default convergence criteria are matters personal preference. A researcher who makes such choices, however, should communicate their findings using language that does not extend beyond what the research design allows. For example, in the case of B&M, "the assessed models indeed systematically differed in their flexibility to fit diverse data patterns" (p. 13) should become "models ... differed in their flexibility to fit *plausible* data patterns *using typical optimization methods and convergence criteria*". As is the case in many scientific studies, communicating all results in detail tends to make the findings much narrower, but more accurate. Alternatively, the estimation optimization settings we used (i.e., `optim.force.converged = TRUE` and increasing `iter.max` to some large number) would have enabled B&M to calculate fit statistics for FP analysis without disposing of over 1.7 million results in total. Beyond such detailed reporting, the emerging FP

literature would also benefit from a better understanding of these research preferences and their impact on model evaluation.

Third, in Set 3 of our simulations, we illustrated that overreliance on ECDFs alone can mask important differences in FP. Although ECDFs and averages in fit statistics may be more concise to report, researchers ought to be cognizant that models with apparent high FP may not have uniformly higher FP when examined using Euler plots, for example. There may be subsets of the data space for which a model in lower average FP model actually fits better. However, B&M based their inferences about FP on the ECDFs of familiar SEM fit statistics: “[T]here were only negligible to small differences in the distribution of the RMSEA...there were only negligible to small differences in the distribution of the TLI” (p. 7), and subsequently, “The RMSEA and, to a slightly lesser extent, the TLI accounted for differences in fitting propensity” (p. 10). Several of B&M’s conclusions were based in part on observed similarity between the cumulative distributions of RMSEA and TLI. They hypothesized that “the RMSEA and TLI should exhibit comparable average values and distributions across competing models” (p. 5) and, after finding that the RMSEA ECDFs were equivalent for all models of the same dimensionality, concluded that “accounting for the number of free parameters of the models was largely sufficient to address differences in fitting propensity” (p. 7).

Relatedly, B&M also based their conclusions on the average fit statistics across all models that converged when fit to plausible data. For instance, in B&M’s Figure 2, the ECDFs of RMSEA are non-uniform for all test lengths and they are clearly overlapping in the 9- and 16-item conditions. Through a visual inspection of the 9-item RMSEA plot in this figure, one can see that a threshold of $RMSEA \leq .30$ would rank the FPs as bifactor > correlated factors > single-factor; but a threshold of $RMSEA \leq .35$ would rank them as correlated factors > bifactor

> single-factor; and a threshold of $RMSEA \leq .40$ would rank them as correlated factors > single-factor > bifactor. Although these RMSEA values are far larger than the established benchmarks for good fit, the main problem for FP analysis is their ambiguous rankings. The same issue applies to B&M's results for TLI and SRMR, as evidenced by the non-uniform and overlapping ECDFs in their Figure S2.

FP, fit, and future research

To reiterate, we speculate that any of our concerns about B&M's methodology can be attributed to a dearth of consistent prior FP analyses to serve as guidelines. Thus, for researchers who may wish to incorporate FP in their substantive research or conduct their own FP analyses, we offer some final considerations – on applied modeling, goodness-of-fit testing, and traditional SEM fit statistics – and general recommendations for future FP studies.

Fitting propensity and applied modeling. The “fit contests” (Reise et al., 2016) that constitute most applied model selection procedures are impacted by the FPs of the models under consideration. So how should applied researchers address FP? One option is for them to run their own FP analyses (e.g., by using the *ockhamSEM* package (Falk & Muthukrishna, 2023) in R, as described above), thereby characterizing and contextualizing the goodness-of-fit of their candidate models. If one model has relatively low FP, then its good fit is unlikely in general and thus more meaningful if achieved in the observed data. Conversely, if another model has relatively high FP, then its good fit is likely in general and thus meaningless if achieved in the observed data. In the latter case, we refer again to Preacher (2006): If a model is established as having high FP, then it may have to be abandoned or replaced.

If applied researchers opt not to run their own FP analyses, they should at the very least allude to the available research on FP and model complexity (as cited throughout our

Introduction), to temper their inferential claims when they are based solely on goodness-of-fit, to discuss inherent complexity as a limitation of their findings, and so forth. Transparency in scientific communication is crucial to enhancing trust in the reported evidence (Elliott, 2022), and being transparent about applied model selection means disclosing the potentially adverse effects of FP on the decision-making process.

Fitting propensity and goodness-of-fit testing. In their seminal paper on model evaluation, Roberts and Pashler (2000) questioned, “How persuasive is a good fit?” They searched for answers to this question in philosophy of science and the history of psychology and found no support for the use of good fit to corroborate a theory. The same authors later wrote that “When used to evaluate complex psychological theories, goodness-of-fit tests have been too easy to pass” (Roberts & Pashler, 2002; p. 605). To make fit tests harder to pass, they argued that researchers should value goodness-of-fit only if the model’s flexibility has been curtailed via constraints. Or, to put it more bluntly, “if a theory [model] does not constrain possible outcomes, the fit is meaningless” (Roberts & Pashler, 2000; p. 359).

FP quantifies the degree to which a model is (not) constrained. A model with relatively high FP is less constrained, so may be pliable enough to fit well to a wide range of dissimilar data patterns. Such a model will be equipped to fit the useless noise that has permeated the observed data along with the useful signal that will generalize beyond the observed data. But a model with relatively low FP is more constrained, so may be rigid enough that it fits well to only a limited range of similar data patterns. In such a case, good fit is more meaningful, because it reflects that the model captures a larger signal-to-noise ratio than that of competing models. Accordingly, we believe that FP should become a routine consideration in statistical model evaluation so that researchers can gauge the persuasiveness of their good fit.

Fitting propensity and SEM fit statistics. Additional problems with future FP research may arise from the use of traditional goodness-of-fit statistics. First, the established benchmarks for goodness-of-fit (e.g., CFI > .95) are irrelevant to FP analysis. B&M stated, “concerning the CFI, RMSEA, and TLI, values often considered acceptable occurred only for a minuscule proportion of variance-covariance matrices across all models and conditions” (p. 13). According to the MDL principle, this is exactly what one should anticipate when considering the complete (or plausible) data space. In his introduction to MDL, Grünwald (2005) uses a simple binary data example to demonstrate modeling as information compression. Out of all possible 10,000-bit sequences (i.e., the complete data space), the proportion of sequences with four times as many 0s as 1s (i.e., the model) is “vanishingly small” (p. 5); specifically, the proportion of the complete data space that is occupied by this model is less than 2^{-2787} . Grünwald also provides this guidance: “Having a short description length for the data is equivalent to identifying the data as belonging to a tiny, very *special* subset out of all a priori possible data sequences” (p. 8). This is an unavoidable aspect of information compression methods.

Similarly, one should not expect to achieve typical SEM benchmarks of good fit except with a “tiny, very special subset” of data matrices. But such an outcome has no bearing on relative FP. Preacher (2006) addressed this issue explicitly when discussing his comparison of two models with the same number of parameters, but different functional forms: “Neither model fit particularly well overall by standard criteria, but the large difference in the proportion of data patterns fit by the two models at any value for RMSR is telling” (p. 241). Whether two (or more) models (with the same number of parameters arranged in different configurations) fit well according to familiar conventions is beside the point: What matters is whether the functional form of one model makes it predisposed to fit data relatively better than the other model(s).

A related point is that some traditional SEM fit statistics may be conceptually inappropriate for FP investigations. Incremental fit indices, such as CFI and TLI, compare the fitted model to a null model, which is often a model implying full independence of the variables. The full independence model can indeed represent one kind of randomly generated data with no underlying relationships among the variables. Therefore, it would be unreasonable to assume that the fitted model can substantially outperform the independence model when the data are random. Further research is needed to understand whether the known idiosyncrasies of the available SEM fit indices (see, e.g., Herzog et al., 2007; Pavlov et al., 2021; Rose et al., 2017) influence their use as criteria in FP studies.

Accordingly, we recommend two paths forward. Research on FP should sidestep these traditional fit statistics and instead focus on relative comparison of the minimum of a discrepancy function, free from the various adjustments inherent in the computation of SEM fit measures, so that competing models with the same number of parameters, but different functional forms, can be compared fairly. At the same time, researchers who are interested in better understanding SEM fit statistics could explore their performance in the context of the complete data space, perhaps uncovering heretofore unknown properties or (dis)advantages.

Recommendations for future FP studies. We emphasize once more that the effect of functional form on FP cannot be fully understood without controlling for other explanations of good fit. The number of parameters is the most obvious driver of good fit, but other aspects of the research design and analysis may also impact the propensity of a model to fit well. In fact, many (or perhaps all) of the analytic steps that the researcher chooses along the “garden of forking paths” (Gelman & Loken, 2013) will affect the description length of the model (Fanelli, 2019) and the corresponding FP. The extent to which other “researcher degrees of freedom” (see

e.g., Wicherts et al., 2016) may affect FP is yet to be examined.

More generally, future research must be careful and thorough when considering the FP of factor analysis models. The conceptual underpinnings of FP have a relatively long history – the MDL principle dates back to Rissanen (1978) and the idea of modeling as information compression is even older (e.g., Kolmogorov, 1965; Shannon, 1948) – but the methodology of FP is fairly recent. Preacher (2006) coined the term “fitting propensity” and established that the intractable integration of normalized maximum likelihood (Rissanen, 2001) and other formulations of the MDL principle can be circumvented by fitting models to many data sets that have been randomly and uniformly sampled from the complete data space. Preacher’s approach has appeared in only a handful of publications since (e.g., Bonifay & Cai, 2017; Ergin, 2020; Falk & Muthukrishna, 2023), and none of these studies established methodological best practices for FP analysis. Such recommendations, to be acquired through further methodological refinement and exemplary applications of FP analysis, are needed to enhance statistical model evaluation in the behavioral sciences.

Acknowledgments

We are grateful to Sonja Winter, Ashley Watts, Hanamori Skoblow, Ashley Greene, and Daniele Fanelli for reviewing a draft of this paper and providing critical feedback. The contributions of Bonifay and Cai were supported by the U.S. Department of Education, Institute of Education Sciences, U.S. Department of Education, through Grant R305D210032. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The computation for this work was performed on the high-performance computing infrastructure provided by Research Support Solutions and in part by the National Science Foundation under grant number CNS-1429294 at the University of Missouri,

Columbia MO. DOI: <https://doi.org/10.32469/10355/69802>.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.
- 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*(1), 125-137.
- Backfisch, I., Scherer, R., Siddiq, F., Lachner, A., & Scheiter, K. (2021). Teachers' technology use for teaching: Comparing two explanatory mechanisms. *Teaching and Teacher Education*, *104*, 103390.
- Bader, M., & Moshagen, M. (2022). Assessing the fitting propensity of factor models. *Psychological Methods*. Advance online publication.
- Bonifay, W. (2024, August 26). Reassessing the fitting propensity of factor models. Retrieved from osf.io/b8apk
- Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate Behavioral Research*, *52*(4), 465-484.
- Bonifay, W., Winter, S. D., Skoblow, H. F., & Watts, A. L. (2024). Good fit is weak evidence of replication: Increasing rigor through prior predictive similarity checking. *Assessment*, *10731911241234118*.
- Bornovalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis in psychopathology research: Appreciating benefits and limitations. *Biological Psychiatry*, *88*(1), 18-27.
- Bozdogan, H. (2000). Akaike's information criterion and recent developments in information

- complexity. *Journal of Mathematical Psychology*, 44(1), 62-91.
- Bryan, C. J., & Harris, J. A. (2019). The structure of suicidal beliefs: A bifactor analysis of the Suicide Cognitions Scale. *Cognitive Therapy and Research*, 43(2), 335-344.
- Burns, G. L., Geiser, C., Servera, M., Becker, S. P., & Beauchaine, T. P. (2020). Application of the bifactor S-1 model to multisource ratings of ADHD/ODD symptoms: An appropriate bifactor model for symptom ratings. *Journal of Abnormal Child Psychology*, 48(7), 881-894.
- Cai, L., & Houts, C. R. (2021). Longitudinal analysis of patient-reported outcomes in clinical trials: Applications of multilevel and multidimensional item response theory. *Psychometrika*, 86(3), 754-777.
- Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, 16(3), 221-248.
- Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, 2(2), 119-137.
- Chan, Y. F., Leung, D. Y., Fong, D. Y., Leung, C. M., Lee, A. M. (2010). Psychometric evaluation of the Hospital Anxiety and Depression Scale in a large community sample of adolescents in Hong Kong. *Quality of Life Research*, 19(6), 865-873
- Chen, S., Liao, F., Murphy, D., & Joseph, S. (2022). Measurement invariance of the English, Chinese, and Spanish versions of the Barrett-Lennard Relationship Inventory. *Measurement and Evaluation in Counseling and Development*, 55(1), 30-47.
- Constantinou, M., & Fonagy, P. (2019). *Evaluating bifactor models of psychopathology using model-based reliability indices*. PsyArxiv. <https://doi.org/10.31234/osf.io/6tf7j>

- Crombie, A. C. (1988). Designed in the mind: Western visions of science, nature and humankind. *History of Science*, 26(1), 1-12.
- Dombrowski, S. C. (2014). Exploratory bifactor analysis of the WJ-III cognitive in adulthood via the Schmid-Leiman procedure. *Journal of Psychoeducational Assessment*, 32(4), 330-341.
- Dombrowski, S. C., Canivez, G. L., Watkins, M. W., & Beaujean, A. A. (2015). Exploratory bifactor analysis of the Wechsler Intelligence Scale for Children—Fifth Edition with the 16 primary and secondary subtests. *Intelligence*, 53, 194-201.
- Dunn, J. C. (2000). Model complexity: The fit to random data reconsidered. *Psychological Research*, 63(2), 174-182.
- Eid, M. (2020). Multi-faceted constructs in abnormal psychology: Implications of the bifactor S-1 model for individual clinical assessment. *Journal of Abnormal Child Psychology*, 48(7), 895-900.
- Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, 22(3), 541.
- 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(3), 42.
- Elliott, K. C. (2022). A taxonomy of transparency in science. *Canadian Journal of Philosophy*, 52(3), 342–355.
- Engyel, M., Urbán, R., Bandi, S., & Nagy, L. (2020). Dimensionality of narcissism: A Bifactorial model of the Narcissistic Personality Inventory using single-stimulus response formats. *Current Psychology*, 1-15.
- Ergin, E. A. (2020). *Fitting Propensities of Item Response Theory Models* (Doctoral dissertation,

Fordham University).

Fadda, D., Scalas, L. F., Meleddu, M., & Morin, A. J. (2017). A bifactor-ESEM representation of the Questionnaire for Eudaimonic Wellbeing. *Personality and Individual Differences, 116*, 216-222.

Falk, C. F., & Muthukrishna, M. (2023). Parsimony in model selection: Tools for assessing fit propensity. *Psychological Methods, 28*(1), 123-136.

Fanelli, D. (2019). A theory and methodology to quantify knowledge. *Royal Society Open Science, 6*(4), Article 181055. <https://doi.org/10.1098/rsos.181055>

Fanelli, D. (2022, January 7). The "tau" of science - How to measure, study, and integrate quantitative and qualitative knowledge. <https://doi.org/10.31222/osf.io/67sak>

Feraco, T., & Cona, G. (2022). Differentiation of general and specific abilities in intelligence. A bifactor study of age and gender differentiation in 8-to 19-year-olds. *Intelligence, 94*, 101669.

Ferguson, G.A. (1954). The concept of parsimony in factor analysis. *Psychometrika, 19*, 281-290.

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

Forster, D. E., Billingsley, J., Russell, V. M., McCauley, T. G., Smith, A., Burnette, J. L., Ohtsubo, Y., Schug, J., Lieberman, D., & McCullough, M. E. (2020). Forgiveness takes

- place on an attitudinal continuum from hostility to friendliness: Toward a closer union of forgiveness theory and measurement. *Journal of Personality and Social Psychology*, 119(4), 861-880.
- Gäde, J. C., Schermelleh-Engel, K., & Klein, A. G. (2017). Disentangling the common variance of perfectionistic strivings and perfectionistic concerns: A bifactor model of perfectionism. *Frontiers in psychology*, 8, 160.
- Gaye, F. (2022). Working Memory and Math in Children with and without ADHD. [Master's thesis, Florida State University].
- Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 348, 1-17.
- Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., ... & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, 31(1), 4-19.
- Gini, C. (1926). The contributions of Italy to modern statistical methods. *Journal of the Royal Statistical Society*, 89, 703-724.
- Giordano, C., Ones, D. S., Waller, N. G., & Stanek, K. C. (2020). Exploratory bifactor measurement models in vocational behavior research. *Journal of Vocational Behavior*, 120, 103430.
- Gnambs, T., & Schroeders, U. (2024). Reliability and factorial validity of the Core Self-Evaluations Scale: A meta-analytic investigation of wording effects. *European Journal of Psychological Assessment*. Advance online publication. <https://doi.org/10.1027/1015->

5759/a000783

- Gomez, R., Stavropoulos, V., Vance, A., & Griffiths, M. D. (2019). 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*, *17*(2), 258-278.
- Greene, A. L., Eaton, N. R., Forbes, M. K., Fried, E. I., Watts, A. L., Kotov, R., & Krueger, R. F. (2022). Model fit is a fallible indicator of model quality in quantitative psychopathology research: A reply to Bader and Moshagen. *Journal of Psychopathology and Clinical Science*, *131*(6), 696-703.
- Greene, A. L., Watts, A. L., Forbes, M. K., Kotov, R., Krueger, R. F., & Eaton, N. R. (2023). Misbegotten methodologies and forgotten lessons from Tom Swift's electric factor analysis machine: A demonstration with competing structural models of psychopathology. *Psychological Methods*, *28*(6), 1374-1403.
- Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*(1), 133-152.
- Grünwald P. (2005). Minimum description length tutorial. In P. D. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications* (pp. 23-80). Cambridge, MA: MIT Press.
- Hand, E. D., & Lonigan, C. J. (2022). Examining the relations between preschooler's externalizing behaviors and academic performance using an S-1 bifactor model. *Research on Child and Adolescent Psychopathology*, *50*(5), 577-589.
- Heinrich, M., Geiser, C., Zagorscak, P., Burns, G. L., Bohn, J., Becker, S. P., ... & Knaevelsrud, C. (2021). On the meaning of the "P Factor" in symmetrical bifactor models of psychopathology: Recommendations for future research from the Bifactor-(S- 1)

- perspective. *Assessment*, 30(3), 487-507 <https://doi.org/10.1177/10731911211060298>
- Heinrich, M., Zagorscak, P., Eid, M., & Knaevelsrud, C. (2020). 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*, 27(7), 1429-1447.
- Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 361-390.
- Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science*, 55, 1-34.
- Hoffmann, M. S., Brunoni, A. R., Stringaris, A., Viana, M. C., Lotufo, P. A., Benseñor, I. M., & Salum, G. A. (2021). Common and specific aspects of anxiety and depression and the metabolic syndrome. *Journal of Psychiatric Research*, 137, 117-125.
- Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika*, 6, 537-549.
- Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10), 2177-2189.
- Junghänel, M., Rodenacker, K., Dose, C., & Döpfner, M. (2020). Applying the bifactor S-1 model to ratings of ADHD/ODD symptoms: a commentary on Burns et al. (2019) and a re-analysis. *Journal of Abnormal Child Psychology*, 48(7), 905-910.
- Kan, K. J., Psychogiopoulos, A., Groot, L. J., de Jonge, H., & Ten Hove, D. (2024). Why do bifactor models outperform higher-order g factor models? A network perspective. *Journal of Intelligence*, 12(2), 18.
- Kittelman, A., Mercer, S. H., McIntosh, K., & Nese, R. N. T. (2021). Development and

- validation of a measure assessing sustainability of Tier 2 and 3 behavior support systems. *Journal of School Psychology*, 85, 140-154. doi:10.1016/j.jsp.2021.02.001
- Kofler, M. J., Singh, L. J., Soto, E. F., Chan, E. S. M., Miller, C. E., Harmon, S. L., & Spiegel, J. A. (2020). Working memory and short-term memory deficits in ADHD: A bifactor modeling approach. *Neuropsychology*, 34(6), 686-698.
- Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1), 1-7.
- Lambert, M. C., January, S. A. A., & Pierce, C. D. (2018). Latent structure of scores from the Emotional and Behavioral Screener. *Journal of Psychoeducational Assessment*, 36(3), 249-260.
- Larsson J (2021). *eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses*. R package version 6.1.1. <https://CRAN.R-project.org/package=eulerr>
- Lecerf, T., & Canivez, G. L. (2022). Exploratory factor analyses of the French WISC-V (WISC-VFR) for five age groups: Analyses based on the standardization sample. *Assessment*, 29(6), 1117-1133.
- Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9), 1989-2001.
- MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, 109(3), 502-511.
- Markon, K. E. (2019). Bifactor and hierarchical models: Specification, inference, and interpretation. *Annual Review of Clinical Psychology*, 15, 51-69.
- Montoya, A. K., & Edwards, M. C. (2021). The poor fit of model fit for selecting number of

- factors in exploratory factor analysis for scale evaluation. *Educational and Psychological Measurement*, 81(3), 413-440.
- Moore, H. L. (1908). The statistical complement of pure economics. *The Quarterly Journal of Economics*, 23(1), 1-33.
- Moore, T. M., Kaczurkin, A. N., Durham, E. L., Jeong, H. J., McDowell, M. G., Dupont, R. M., Applegate, B., Tackett, J. L., Cardenas-Iniguez, C., Kardan, O., Akcelik, G. N., Stier, A. J., Rosenberg, M. D., Hedeker, D., Berman, M. G., & Lahey, B. B. (2020). Criterion validity and relationships between alternative hierarchical dimensional models of general and specific psychopathology. *Journal of Abnormal Psychology*, 129(7), 677-688.
- Morgan, M. S. (2012). *The world in the model: How economists work and think*. Cambridge and New York: Cambridge University Press.
- 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: A Multidisciplinary Journal*, 23(1), 116-139.
- 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
- Muthukrishna, M., & Heinrich, J. (2019). A problem in theory. *Nature Human Behavior*, 3(3), 221-229.
- Myers, N. D., Martin, J. J., Ntoumanis, N., Celimli, S., & Bartholomew, K. J. (2014). Exploratory bifactor analysis in sport, exercise, and performance psychology: A substantive-methodological synergy. *Sport, Exercise, and Performance Psychology*, 3(4), 258-272.
- Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian

- approach. *Psychonomic Bulletin & Review*, 4(1), 79-95.
- Myung, I. J., Pitt, M. A., & Kim, W. 2004. Model evaluation, testing and selection. In K. Lambert & R. Goldstone (Eds.), *The handbook of cognition* (pp. 422-436). Thousand Oaks, CA: Sage.
- Neyman, J. (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. *The Annals of Mathematical Statistics*, 10(1), 35-57.
- Norton, S., Cosco, T., Doyle, F., Done, J., & Sacker, A. (2013). The Hospital Anxiety and Depression Scale: a meta confirmatory factor analysis. *Journal of Psychosomatic Research*, 74(1), 74-81.
- Olderbak, S., Wilhelm, O., & Mokros, A. (2021). Psychopathy checklist: Screening version: A bifactor structure for forensic and community samples. *Psychological Assessment*, 33(11), 1050-1064.
- Panayiotou, M., Santos, J., Black, L., & Humphrey, N. (2022). Exploring the dimensionality of the social skills improvement system using exploratory graph analysis and bifactor-(S- 1) modeling. *Assessment*, 29(2), 257-271.
- Panjeh, S., Pompeia, S., & Cogo-Moreira, H. (2022). Probing different aspects of short and ill-timed sleep in adolescents using the Morningness-Eveningness Scale for Children. *Chronobiology International*, 39(3), 333-345.
- Patrick, C. J., Kramer, M. D., Vaidyanathan, U., Benning, S. D., Hicks, B. M., & Lilienfeld, S. O. (2019). Formulation of a measurement model for the boldness construct of psychopathy. *Psychological Assessment*, 31(5), 643-659.
- Pavlov, G., Maydeu-Olivares, A., & Shi, D. (2021). Using the standardized root mean squared residual (SRMR) to assess exact fit in structural equation models. *Educational and*

Psychological Measurement, 81(1), 110-130.

Pavlović, T., Moskalenko, S., & McCauley, C. (2022). Bifactor analyses provide uncorrelated measures of activism intentions and radicalism intentions. *Dynamics of Asymmetric Conflict*, 15(2), 123-140.

Peterson, E. C., Snyder, H. R., Neilson, C., Rosenberg, B. M., Hough, C. M., Sandman, C. F., ... & Kaiser, R. H. (2022). General and specific dimensions of mood symptoms are associated with impairments in common executive function in adolescence and young adulthood. *Frontiers in Human Neuroscience*, 16, 838645.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, 6(10), 421-425.

Pitt, M. A., Kim, W., & Myung, I. J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review*, 10(1), 29-44.

Preacher, K. J. (2003). *The role of model complexity in the evaluation of structural equation models*. [Doctoral dissertation, The Ohio State University]

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, 41(3), 227-259.

Preacher, K. J., & Yaremych, H. E. (2023). Model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (2nd ed.) (pp. 206-222). New York: Guilford Press.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively

- reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research*, 51(6), 818-838.
- Reise, S. P., Mansolf, M., & Haviland, M. G. (2023). Bifactor measurement models. In R. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 329-348). New York: Guilford Press.
- Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, 73(1), 5-26.
- Remmert, N., Schmidt, K. M. B., Mussel, P., & Eid, M. (2021, August 9). The Berlin Misophonia Questionnaire (BMQ): Development and validation of a symptom-oriented diagnostic instrument for the measurement of misophonia.
<https://doi.org/10.31234/osf.io/mujya>
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5), 465-471.
- Rissanen, J. (1986). Stochastic complexity and modeling. *The Annals of Statistics*, 1080-1100.
- Rissanen, J. J. (1996). Fisher information and stochastic complexity. *IEEE transactions on information theory*, 42(1), 40-47.
- Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, 47(5), 1712-1717.
- Ritzka, D., Petzold, C., Wäßnig, N., Schmalbach, B., & Petrowski, K. (2020). Investigation of the factorial structure and psychometrics of the German version of the Florida Shock Anxiety Scale. *Psychology, Health & Medicine*, 25(3), 344-353.
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107(2), 358-367.

- Roberts, S., & Pashler, H. (2002). Reply to Rodgers and Rowe (2002). *Psychological Review*, *109*(3), 605-607.
- Romeijn, J. W. (2017). Inherent complexity: A problem for statistical model evaluation. *Philosophy of Science*, *84*(5), 797-809.
- Rose, S. A., Markman, B., & Sawilowsky, S. (2017). Limitations in the systematic analysis of structural equation model fit indices. *Journal of Modern Applied Statistical Methods*, *16*(1), 69-85.
- Rosenström, T., Gjerde, L. C., Krueger, R. F., Aggen, S. H., Czajkowski, N. O., Gillespie, N. A., ... & Ystrom, E. (2019). Joint factorial structure of psychopathology and personality. *Psychological Medicine*, *49*(13), 2158-2167.
- Rosenström, T., Torvik, F. A., Ystrom, E., Aggen, S. H., Gillespie, N. A., Krueger, R. F., ... & Reichborn-Kjennerud, T. (2021). Specific antisocial and borderline personality disorder criteria and general substance use: A twin study. *Personality Disorders: Theory, Research, and Treatment*, *12*(3), 228-240.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36.
- Rosseel, Y. (2023, Jan. 9). *lavaan: Latent Variable Analysis*. Retrieved from <https://cran.r-project.org/web/packages/lavaan/lavaan.pdf>
- Schachler, V., Epple, S. D., Clauss, E., Hoppe, A., Slemp, G. R., & Ziegler, M. (2019). Measuring job crafting across cultures: Lessons learned from comparing a German and an Australian sample. *Frontiers in Psychology*, *10*, 991.
- Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*, 53-61.

- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461-464.
- Seong, H., Lee, S., & Chang, E. (2021). Perfectionism and academic burnout: Longitudinal extension of the bifactor model of perfectionism. *Personality and Individual Differences*, 172, 110589.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379-423.
- Simonet, D. V., Miller, K. E., Askew, K. L., Sumner, K. E., Mortillaro, M., & Schlegel, K. (2021). How multidimensional is emotional intelligence? Bifactor modeling of global and broad emotional abilities of the Geneva Emotional Competence Test. *Journal of Intelligence*, 9(1), 14.
- Stanton, K., Khoo, S., Watson, D., Gruber, J., Zimmerman, M., & Weinstock, L. M. (2019). Unique and transdiagnostic symptoms of hypomania/mania and unipolar depression. *Clinical Psychological Science*, 7(3), 471-487.
- Thöne, A.-K., Junghänel, M., Görtz-Dorten, A., Dose, C., Hautmann, C., Jendreizik, L. T., Treier, A.-K., Vetter, P., von Wirth, E., Banaschewski, T., Becker, K., Brandeis, D., Dürrwächter, U., Geissler, J., Hebebrand, J., Hohmann, S., Holtmann, M., Huss, M., Jans, T., . . . Döpfner, M. (2021). Disentangling symptoms of externalizing disorders in children using multiple measures and informants. *Psychological Assessment*, 33(11), 1065-1079.
- Tibubos, A. N., Otten, D., Zöllner, D., Binder, H., Wild, P. S., Fleischer, T., Johar, H., Atasoy, S., Schulze, L., & Ladwig, K.-H. (2021). Bidimensional structure and measurement equivalence of the Patient Health Questionnaire-9: Sex-sensitive assessment of depressive symptoms in three representative German cohort studies. *BMC Psychiatry*, 21(1), 1-13.
- 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(6), 759-773.
- Waldman, I. D., King, C. D., Poore, H. E., Luningham, J. M., Zinbarg, R. M., Krueger, R. F., ... & Zald, D. (2023). Recommendations for adjudicating among alternative structural models of psychopathology. *Clinical Psychological Science*, 11(4), 616-640.
- Waller, N. G. (2017). Generating correlation matrices with specified eigenvalues using the method of alternating projections. *The American Statistician*, 74, 21-28.
- Watrin, L., Schroeders, U., & Wilhelm, O. (2022, August 8). Gc at its boundaries: A cross-national investigation of declarative knowledge. <https://doi.org/10.31234/osf.io/r5ubg>
- Watts, A. L., Greene, A. L., Bonifay, W., & Fried, E. I. (2024). A critical evaluation of the p-factor literature. *Nature Reviews Psychology*, 3(2), 108-122.
- Watts, A. L., Greene, A. L., Ringwald, W., Forbes, M. K., Brandes, C. M., Levin-Aspenson, H. F., & Delawalla, C. (2023). Factor analysis in personality disorders research: Modern issues and illustrations of practical recommendations. *Personality Disorders: Theory, Research, and Treatment*, 14(1), 105.
- Watts, A. L., Poore, H. E., & Waldman, I. D. (2019). Riskier tests of the validity of the bifactor model of psychopathology. *Clinical Psychological Science*, 7(6), 1285-1303.
- Watts, A. L., Lane, S. P., Bonifay, W., Steinley, D., & Meyer, F. A. (2020). Building theories on top of, and not independent of, statistical models: The case of the p-factor. *Psychological Inquiry*, 31(4), 310-320.
- Wehner, C., Maaß, U., Leckelt, M., Back, M. D., & Ziegler, M. (2021). Validation of the Short Dark Triad in a German sample: Structure, nomological network, and an ultrashort version. *European Journal of Psychological Assessment*, 37(5), 397-408.

Wendt, L. P., Jankowsky, K., Schroeders, U., London Personality and Mood Disorder Research

Consortium, Nolte, T., Fonagy, P., Montague, P.R., Zimmerman, J., & Olaru, G. (2022).

Mapping established psychopathology scales onto the Hierarchical Taxonomy of

Psychopathology (HiTOP). *Personality and Mental Health*.

<https://doi.org/10.1002/pmh.1566>

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural

equation modeling. In R.H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp.

209-231). New York: Guilford Press.

Wexler, K. (1978). A review of John R. Anderson's *Language, Memory, and*

Thought. *Cognition*, 6(4), 327-351.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M.

A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological

studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 1832.

Wilkinson, L. (2011). Exact and approximate area-proportional circular Venn and Euler

diagrams. *IEEE Transactions on Visualization and Computer Graphics*, 18(2), 321-331.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, 45, e1.

Yost, K. J., Waller, N. G., Lee, M. K., & Vincent, A. (2017). The PROMIS fatigue item bank has

good measurement properties in patients with fibromyalgia and severe fatigue. *Quality of*

Life Research, 26(6), 1417-1426.

Zhang, B., Sun, T., Cao, M., & Drasgow, F. (2021). Using bifactor models to examine the

predictive validity of hierarchical constructs: Pros, cons, and solutions. *Organizational*

Research Methods, 24(3), 530-571.

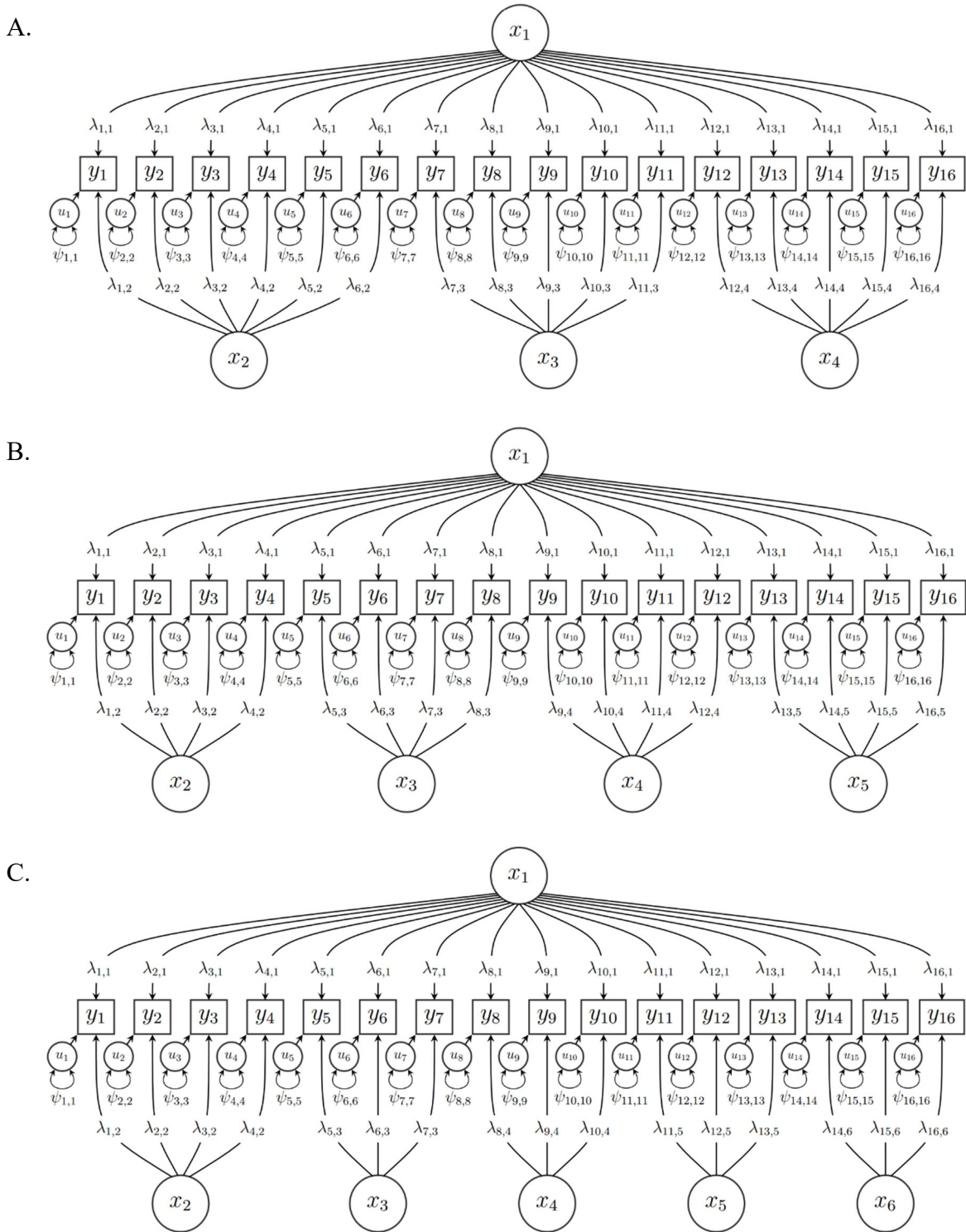


Figure 1. Three factor models with $q = 48$ freely estimated parameters: (A) a bifactor model with three specific factors, (B) a bifactor model with four specific factors, and (C) a bifactor model with five specific factors.

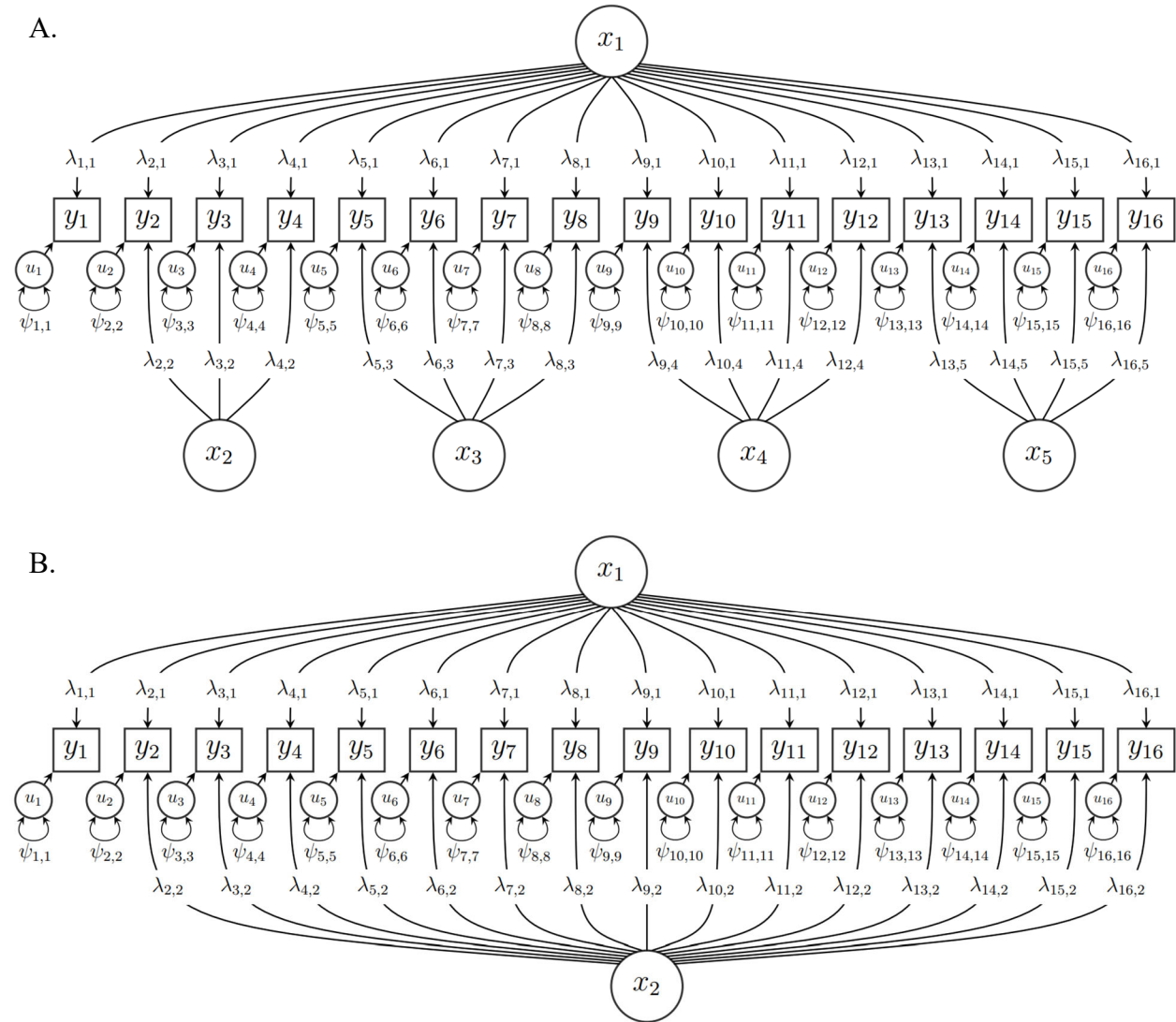


Figure 2. Two factor models with $q = 47$ freely estimated parameters: (A) a “confirmatory” bifactor model and (B) an exploratory factor analysis model.

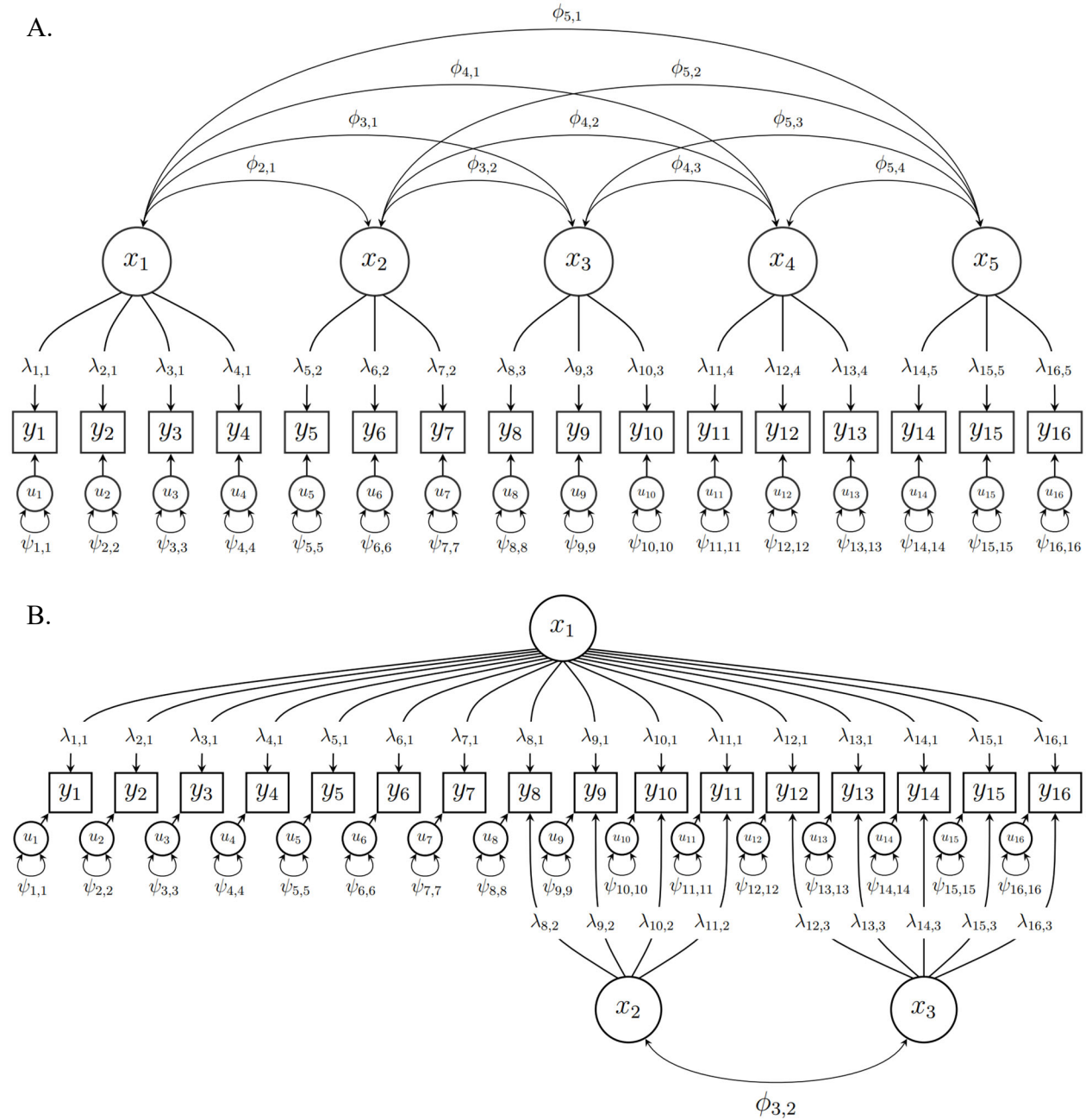


Figure 3. Two factor models with $q = 42$ freely estimated parameters: (A) a correlated-factors model and (B) a bifactor S-1 model.

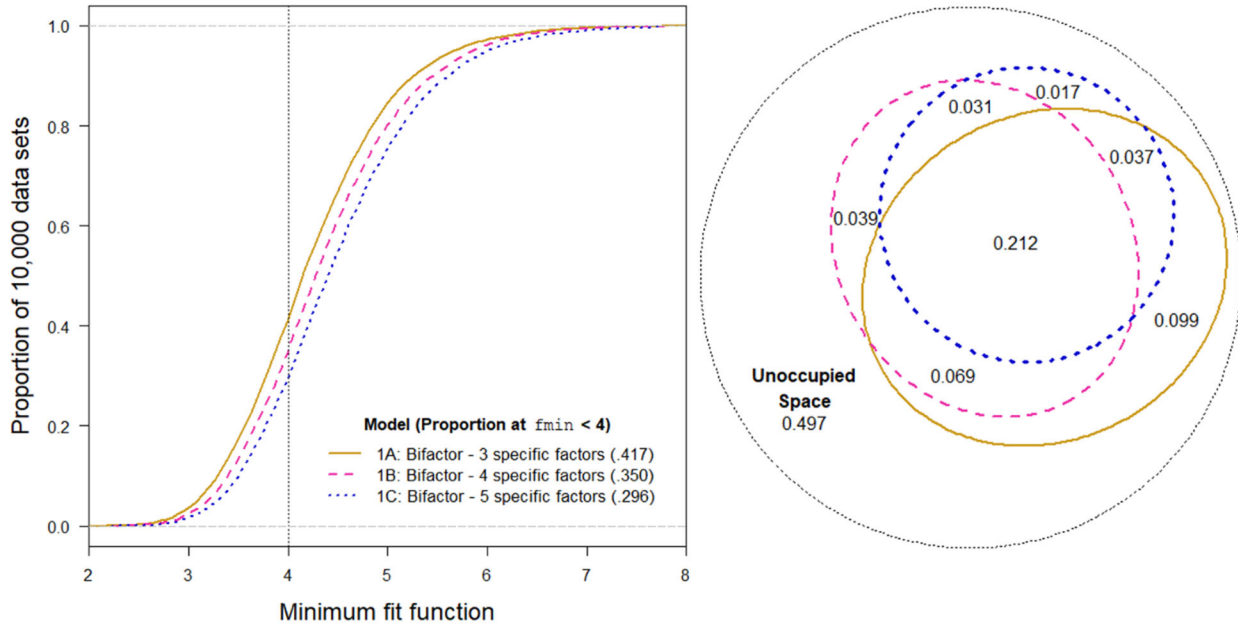


Figure 4. Fitting propensities of Models 1A (a bifactor model with three specific factors), 1B (a bifactor model with four specific factors), and 1C (a bifactor model with five specific factors) relative to 10,000 data sets that were randomly, uniformly sampled from the complete continuous data space, displayed as empirical cumulative distribution functions (left) and an area-proportional Euler plot based on minimum fit function $f_{\min} < 4.0$ (right). All three models include $q = 48$ freely estimated parameters.

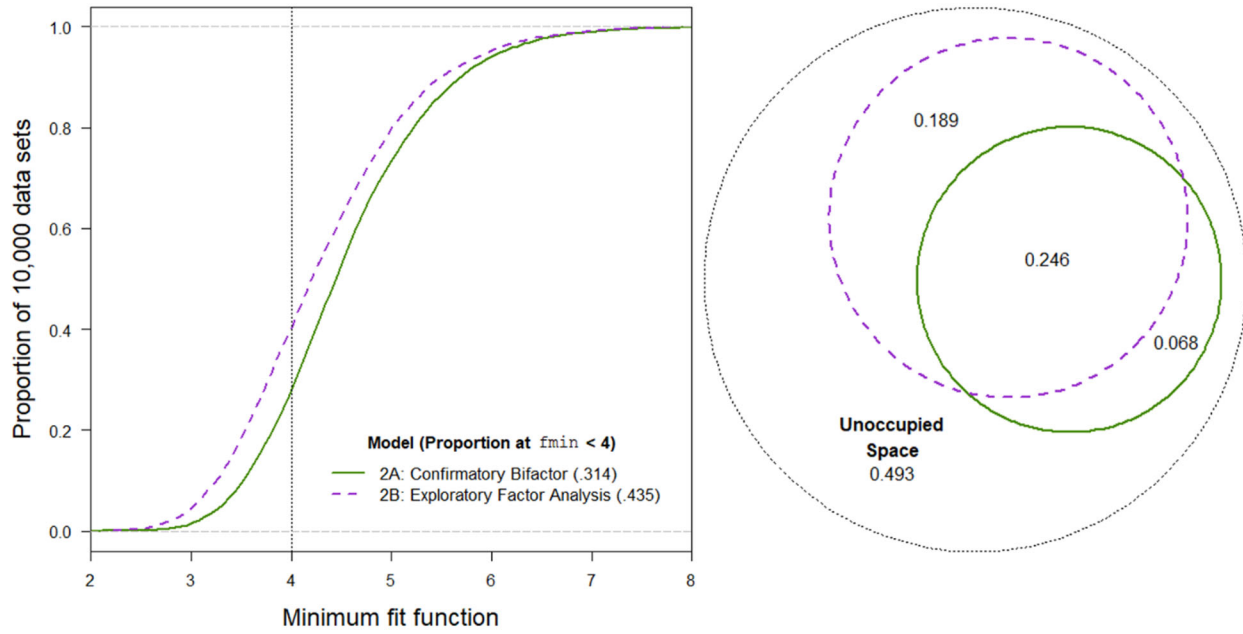


Figure 5. Fitting propensities of Models 2A (a confirmatory bifactor model) and 2B (an exploratory factor analysis model) relative to 10,000 data sets that were randomly, uniformly sampled from the complete continuous data space, displayed as empirical cumulative distribution functions (left) and an area-proportional Euler plot based on minimum fit function $f_{min} < 4.0$ (right). Both models include $q = 47$ freely estimated parameters.

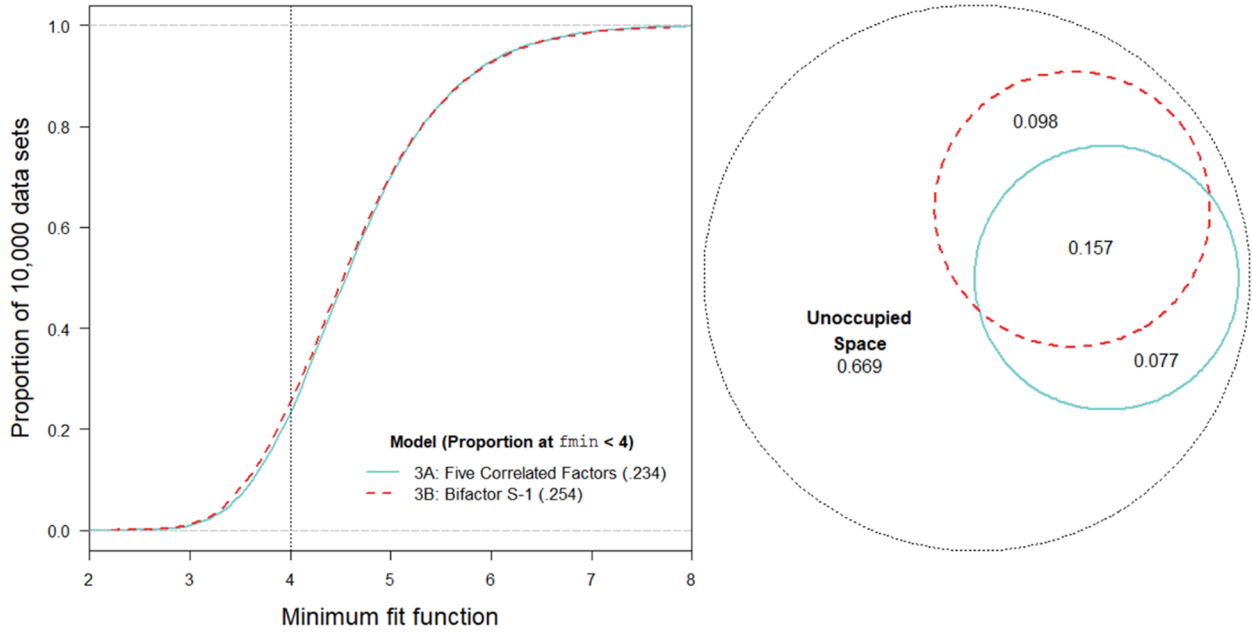


Figure 6. Fitting propensities of Models 3A (a correlated-factors model with five factors) and 3B (a bifactor S-1 model) relative to 10,000 data sets that were randomly and uniformly sampled from the complete continuous data space, displayed as empirical cumulative distribution functions (left) and an area-proportional Euler plot based on minimum fit function $f_{\min} < 4$ (right). Both models include $q = 42$ freely estimated parameters.