Pre-Candidacy Thesis

Bayesian SEM with a General Factor and Cross-loadings: A Sensitivity Study

by

Aleksandr Ivanov

A Research Project submitted in partial satisfaction of the requirements

for the degree of Master of Arts

in Psychological Sciences

Committee Members:

Dr. Haiyan Liu, Chair, Advisor

Dr. Sarah Depaoli

Dr. Meng Qiu

University of California, Merced

May 2025

UNIVERSITY OF CALIFORNIA, MERCED

Pre-Candidacy Thesis: Bayesian SEM with a General Factor and Cross-loadings: a Sensitivity Study

A Research Project submitted in partial satisfaction of the requirements for the degree of Master of Science in Psychological Sciences

by

Aleksandr
Ivanov

The Research Project of Aleksandr Ivanov is approved, and it is acceptable in quality:

Professor Name	Signature	Date

X _____
Haiyan Liu
Chair

X _____
Sarah Depaoli

X _____
Meng Qiu

As bifactor modeling has become increasingly popular, concerns have been raised for its potential overuse due to its enhanced flexibility in fitting data. This issue is particularly pressing given the common but flawed practice of reusing the same data to generate and test structural hypotheses, as traditionally done for model selection. Another issue of bifactor modeling lies in its strong assumptions, such as the orthogonality of factors and the exclusion of cross-loadings. This study addresses both issues by using the Bayesian framework to estimate a confirmatory bifactor model that allows several cross-loadings. We argue that Bayesian estimation, with its flexibility and ability to incorporate prior knowledge, is a promising approach for estimating complex bifactor models with cross-loadings. We assess parameter recovery of such models under varying magnitudes of a general factor, different levels of cross-loadings, sample sizes, and prior specifications. Of particular interest are near-zero priors, which we examine for their ability to test hypotheses about the presence of a general factor. We argue that the near-zero small variance priors allow researchers to evaluate the existence of a general factor within a single modeling step, rather than using the bifactor model as a last resort when other models fail to fit. Additionally, we evaluate the performance of recently developed Bayesian indices of approximate fit to detect meaningful misspecifications. The results suggest good parameter recovery of the true model when diffuse priors are used, and mixed performance of fit indices. Based on the results, we provide several guidelines for applied researchers.

Keywords: bifactor, structural equation modeling, confirmatory factor analysis, Bayesian, cross-loadings, near-zero priors, model fit

Introduction

Social and psychological researchers often study latent variables — unobservable constructs that underlie observed behaviors (Borsboom et al., 2003). Examples include mood, personality traits, attitudes, clinical conditions, and knowledge (Borsboom et al., 2003). Because latent constructs cannot be measured directly, they are inferred from behavioral indicators, whose scores are statistically integrated to form the operational definition of the underlying construct (Griffin & Bartholomew, 1994). These indicators are typically collected through psychometric scales designed to capture the relevant aspects of the latent variable. Scales can range from as few as two items to more than fifty and may target a single latent construct or multiple relevant constructs. Each item is meant to capture a different facet of the latent construct, collectively providing a more comprehensive representation (Lane et al., 2016).

Before a scale can be used in research or applied settings, it must undergo a rigorous statistical evaluation. One primary objective is to ensure that the item scores are meaningfully related — specifically, that they demonstrate sufficient inter-correlations to support internal consistency (Spearman, 1961). Another critical objective is to evaluate the validity of a scale, which refers to the degree to which the scores accurately represent the intended latent construct and the strength of the evidence supporting their use to make inferences about that construct (Kane, 2013; Messick, 1995). Several factors comprise validity, including item content and wording, consequential validity (i.e., whether score differences correspond to meaningful differences in life outcomes), and internal structure (e.g., factor structure) (Cook & Beckman, 2006). While some factors, such as item wording, can be addressed during scale development, others — like internal structure — must be evaluated during data analysis, typically through factor analysis.

Factor analysis is a key method to understand how observed indicators relate to underlying latent constructs (Cattell, 1952). In this approach, researchers specify a model that reflects hypothesized relationships among observed variables and latent factors. The model–implied covariance matrix is then estimated and compared to the observed covariance matrix derived from the data, with model estimation aimed at minimizing the discrepancy between the two. Model parameters and fit indices are subsequently computed to evaluate how well the model fits the data (for details on factor analysis, see Lawley & Maxwell, 1971; P. Kline, 2014; Harrington, 2009).

Psychological data are often highly complex, and statistical models, including factor analysis models, can only approximate this complexity. Consequently, all models should be viewed as simplified abstractions of reality (Box, 1976; MacCallum & Tucker, 1991; Chatfield, 1995; Depaoli, 2021). Moreover, a measurement structure identified in one dataset may differ from that in another, even when the same scale is used. These limitations are evident in applied research. For example, Moreira et al. (2021) conducted a systematic review of the studies using Clayton's Environmental Identity Scale (Clayton, 2003) and found substantial inconsistencies in the reported factor structures, including the disagreement in the number of underlying dimensions. In addition, none of the reported models demonstrated satisfactory fit when retested. These ambiguous results may reflect the inflexibility of theory–based factor analysis models, which may be too restrictive to accurately capture the complexity of real-world psychological constructs (Moreira et al., 2021).

The complexity of real-world data reflects a broader issue in statistical analysis: the discrepancy between exploratory and confirmatory approaches. Like many statistical methods, factor analysis can be applied either from an exploratory or a confirmatory philosophical perspective (Suhr, 2006; Fife & Rodgers, 2022). In the confirmatory factor analysis (CFA), researchers formulate specific hypotheses about

the factor structure and test them against observed data, seeking evidence to support or reject these hypotheses. In contrast, the exploratory factor analysis (EFA) is employed when no or little prior theories exist about the structure, and the primary goal of the analysis is to uncover potential patterns within the data. Therefore, the EFA is largely unconstrained, allowing the data to shape the model, whereas the CFA imposes a pre-determined structure that is tested against the data (Thompson, 2004). The predetermined structure implies that only theoretically meaningful and important parameters are allowed to be freely estimated, and others are fixed at zero or another specific value. When both analyses are used together, the CFA often fails to reproduce the structure uncovered in the EFA. To remedy this problem, applied researchers often retest the model by freeing several fixed parameters one-by-one, based on modification indices. This practice can eventually improve fit of the CFA model, but simultaneously leads to inflated Type I error rates for model parameters and contradicts the philosophy of confirmatory testing (MacCallum et al., 1992; B. Muthén & Asparouhov, 2012; Crede & Harms, 2019). The sole use of the CFA is more common in psychological research involving latent variables, as researchers normally have theoretically driven structural hypotheses about their constructs. However, even in that scenario a model is often retested to find the best-fitting solution.

While the EFA and the CFA differ in how researchers approach model specification, data–driven versus theory–driven, they both rely on the common factor model as their underlying framework. This model serves as the basis for understanding how a set of observed variables can be explained by a smaller number of latent factors. There are several variants of the common factor model, each differing in complexity and theoretical assumptions (Mulaik, 2018). In its simplest form, the model assumes that all indicators are associated with a single underlying

construct. A slightly more complex variant is the correlated factors model, in which each item loads onto one specific latent variable, and the latent variables themselves are allowed to correlate (Mulaik, 2018). More advanced structures introduce hierarchical relationships among latent variables, as seen in the higher-order factor model, where a second-order (or general) factor accounts for the covariation among several first-order factors, which in turn explain the observed indicators (Markon, 2019).

A conceptually related, but structurally distinct model is the bifactor model, originally introduced by Holzinger & Swineford (1937). Like the hierarchical factor model, the bifactor model incorporates both a general and domain-specific dimensions, but it differs in how these dimensions are specified. In the bifactor model, the general factor directly influences all items, while the specific factors account for additional shared variance among subsets of items (Markon, 2019). The total item variance is decomposed into three components: variance common to all items (captured by the general factor), variance shared only among subsets of items (captured by specific factors), and unique item–specific variance. Conceptually, the general factor represents a broad construct that influences all items, whereas the specific factors capture narrower, domain–specific sources of variance. For example, it has been a historical debate about the nature of the "intelligence" construct: either existing as a general ability (Spearman, 1961) or multiple correlated domains of cognitive ability (e.g., verbal, spatial, math Sternberg & Grigorenko, 2002; Humphreys, 1979; Thurstone, 1938). Bifactor modeling accommodates both perspectives by allowing general and domain–specific dimensions to be modeled simultaneously (Reise, 2012). Bifactor modeling has significantly risen in popularity in the last decade (Reise et al., 2023) and has been applied to various constructs, including psychopathy (Patrick et al., 2007), alexithymia (Carnovale et al., 2021),

health outcomes (Reise et al., 2007), well-being (Yeo & Suárez, 2022), and emotional intelligence (Tejada-Gallardo et al., 2022), to name just a few. Empirically testing a bifactor model requires satisfying several assumptions and model specification rules.

In the standard bifactor model, each item loads on a single general factor, which captures the common variance across all items, and on one specific factor, which accounts for additional variance unique to a particular subdomain (Reise, 2012; Markon, 2019). Cross-loadings across specific factors are not permitted in this traditional formulation, as the model is designed to cleanly partition variance into general, domain–specific, and unique components. While the covariance between general factor and specific factors indeed needs to be fixed at zero to maintain theoretical soundness and statistical identifiability, cross-loadings and covariances between specific factors can be allowed (Murray & Johnson, 2013; Eid et al., 2018; Markon, 2019; Fang et al., 2021), with the latter potentially being the manifestation of the former. Flexible bifactor variants — such as those estimated within the Exploratory Structural Equation Modeling (ESEM) framework with target rotation (Asparouhov & Muthén, 2009; Morin et al., 2016) — allow items to load on multiple specific factors. The inclusion of cross-loadings is justified on both empirical and theoretical grounds. Empirically, items may reflect multiple subdomains, and constraining them to load on a single specific factor can distort the underlying structure. Even though in theory the general factor should absorb possible population cross-loadings, fitting a standard bifactor solution to items that have population cross-loadings was shown to result in distorted general factor loadings for corresponding observed variables (Reise et al., 2010; Ximénez et al., 2022; Reise, 2012), reflecting the need to account for substantive cross-loadings even in bifactor models. Theoretically, when constructs span overlapping domains, cross-loadings may offer a more accurate representation, manifesting the reflective nature of latent

variables as the cause of variation in observed variables (Edwards & Bagozzi, 2000). Therefore, more relaxed models can provide greater flexibility and may better capture the complexity inherent in psychological data (MacCallum & Tucker, 1991; Morin et al., 2016, 2020). A recent study by Ximénez et al. (2022) evaluated parameter recovery and model fit in bifactor analysis using maximum likelihood estimation, specifically examining the consequences of omitting cross-loadings. The findings revealed substantial bias in parameter estimates when cross-loadings were moderate or large, and showed that standard SEM fit indices failed to detect the resulting model misfit — often indicating good fit despite considerable misspecification.

Given the limitations of maximum likelihood (ML) estimation, particularly its inability to detect misspecification when cross-loadings are omitted in bifactor models, Bayesian estimation offers a promising alternative. Bayesian estimation is grounded in Bayes' theorem and incorporates prior information about parameters (Ghosh et al., 2007). In this framework, prior beliefs about the values of parameters are expressed through prior distributions, which are then combined with the data–driven likelihood to produce a posterior distribution. While point estimates, such as the posterior mean or median, can be reported similarly to ML estimates, the posterior distribution also quantifies uncertainty directly, eliminating the need for traditional significance testing that relies on asymptotic distributions (Van De Schoot & Depaoli, 2014; Wagenmakers et al., 2018). The introduction of prior information enhances model identification, which can make models that are inestimable in frequentist framework estimable using Bayesian approach (B. Muthén & Asparouhov, 2012). This flexibility is especially valuable in complex bifactor models, where issues of identification, small sample sizes, or omitted cross-loadings can distort ML–based inference (Green & Yang, 2018; Bader et al., 2022; Morin et

al., 2016; Bonifay & Cai, 2017; Eid et al., 2017; Murray & Johnson, 2013).

In Bayesian inference, priors play a central role, as they represent the researcher's beliefs or knowledge about parameter values before observing the data. Priors can vary in both informativeness and accuracy (Lemoine, 2019). An accurate prior is one that is centered at the true population value (or the center of the population distribution), while informativeness refers to the degree of certainty, typically reflected in the spread (variance) of the prior distribution. For example, if a prior is correctly centered at the true population value and has a small variance, it is considered a strongly informative and accurate prior. If it is still centered correctly but has a large variance, it is termed a weakly informative accurate prior. The choice of prior distribution depends on the nature of the parameter being estimated. For instance, factor loadings are often assigned normal priors with specified means and variances, reflecting the expected distribution of loadings in the population.

A natural extension of the role of priors in Bayesian estimation is the use of near-zero small variance priors, which are particularly valuable in the context of bifactor analysis with cross-loadings. The rationale behind near-zero small variance priors is that they do not enforce the strict assumption that a parameter is exactly zero, but instead allow for a degree of flexibility, which encourages estimates to be close to zero unless the data strongly suggest otherwise. This approach avoids model non-identification that can arise from freely estimating too many parameters, while also preventing the omission of potentially important non-zero effects. B. Muthén & Asparouhov (2012) introduced this approach in the confirmatory factor analysis as an alternative to the traditional practice of fixing certain loadings to zero or sequentially freeing them based on modification indices. This framework allowed for simultaneous confirmatory testing and exploratory probing of potential

cross-loadings in a single modeling step, demonstrating the flexibility and efficiency of Bayesian estimation in less constrained confirmatory models.

We argue that this same principle can be extended to general factor loadings in bifactor models, particularly when a researcher is uncertain whether a general factor should be included. Such method appears to potentially be very useful, as the immense growth in the popularity of bifactor models is accompanied with widespread concerns over their misuse due to large fit propensity (Bonifay et al., 2017; Eid et al., 2018) (natural proneness to well-fitting (Preacher, 2006)). The ability of a bifactor model to always appear well-fitting and even fitting random or nonsensical data well is shown in multiple studies (Reise et al., 2016; Bonifay & Cai, 2017; Greene et al., 2019; Watts et al., 2020) and calls for very careful applied use of bifactor models. It can often be the case that an applied researcher has a theoretically grounded interest in testing a bifactor structure and suspects a few important cross-loadings may exist besides multiple small cross-loadings. Rather than fitting a bifactor model with all cross-loadings fixed at zero, they can model only a selected few — either freely or with informative accurate priors — while applying near-zero priors to the general factor loadings. Meaningful cross-loadings will be incorporated in the model and allowed to be freely estimated, while small non-meaningful cross-loadings will be absorbed by the general factor (Murray & Johnson, 2013; Raykov et al., 2019). This practice would allow the model to reveal whether the general factor is truly needed, similar to the exploration of cross-loadings in B. Muthén & Asparouhov (2012). This approach avoids common pitfalls such as excessive model re-specification and repeated use of the same data for hypothesis generation and confirmation, while providing more insight into the structure of the construct, compared to the traditional bifactor model.

An additional challenge in evaluating complex models is that traditional SEM

fit indices often fail to detect substantial model misspecifications and are a function of many other variables besides actual model fit, such as model type, sample size and the magnitude of between-item covariances (Saris et al., 2009; Heene et al., 2011; Tomarken & Waller, 2003; Lai & Green, 2016; Savalei, 2012). Specifically for bifactor models, as we noted above, fit indices, such as CFI and RMSEA, often indicate perfect fit regardless of the true population model. As shown by Ximénez et al. (2022), substantial misspecifications in a complex bifactor model with cross-loadings can be overlooked under standard maximum likelihood-based indices. This limitation can potentially be addressed using Bayesian estimation, as it offers alternative tools for model evaluation. Specifically, posterior predictive checking provides a flexible approach to assessing model-data fit without relying on large-sample theory (B. Muthén & Asparouhov, 2012). In addition, recent advances in Bayesian structural equation modeling have made it possible to estimate approximate fit indices, such as RMSEA and CFI, within the Bayesian framework (Garnier-Villarreal & Jorgensen, 2020; Asparouhov & Muthén, 2021), using more sophisticated Bayesian analogues of the frequentist chi-square (χ^2) metric and the degrees of freedom at each iteration of the estimation algorithm. These developments offer applied researchers a familiar yet more nuanced toolkit for evaluating model fit, particularly when exploring uncertain model components, such as cross-loadings or general factors. Together, near-zero priors and Bayesian fit indices enhance both the flexibility and rigor of model evaluation, helping to ensure that retained model components are both theoretically justified and empirically supported. However, as these indices are relatively new, there is a limited number of methodological studies evaluating their performance in complex Bayesian SEM and CFA models. This is especially important in the context of using near-zero priors. These priors can influence model fit, depending on the correct assignment to

parameters that are actually substantially different from zero in the population (B. Muthén & Asparouhov, 2012).

The discussion above highlights several important gaps in the literature. First, there has been no systematic evaluation of Bayesian estimation performance in bifactor models with cross-loadings, particularly regarding how different types of priors and model misspecifications influence parameter recovery. Second, the utility of near-zero priors for detecting the presence of a general factor remains largely unexplored. Third, it is still unclear how effectively Bayesian fit indices, including posterior predictive p-values (PPP) and Bayesian approximate fit indices, such as BRMSEA and BCFI, can detect model misfit in bifactor analysis, specifically, the omission of general factors or cross-loadings. To address each of these gaps, we will conduct a simulation study designed to systematically examine the impact of prior specification and model misspecification on parameter recovery and model fit assessment in Bayesian bifactor models with cross-loadings. The results will provide insight as to which conditions the model can be applied to and which prior settings to use to achieve convergence, identification, and accuracy of results. This will offer guidance for applied researchers to discern in which conditions it is better to choose the Bayesian bifactor model to better account for model complexity associated with the presence of a non-redundant and meaningful general factor and a couple of important cross-loadings. This will contribute to drawing more valid conclusions based on applied data and helping psychological science move forward.

The remainder of the manuscript is organized as follows. First, we describe the measurement model in SEM and its statistical assumptions. Second, we outline the simulation design, sampling process and assessment of convergence. Third, we overview the measures used to evaluate the results of the simulation study, such as bias and coverage, as well as indices of exact and approximate fit. Fourth, we

analyze the results of the simulation study. The first portion of the results concerns parameter recovery, comprised of bias, mean squared error, Bayesian coverage, and a Bayesian analogue of power to detect a non-zero population parameter. The second portion is devoted solely to the performance of near-zero priors in detecting general factor loadings. In the third portion, we provide the results of testing sensitivity of fit indices to misspecifications in different parts of the model. Finally, we summarize the study, outlining possible limitations and future directions.

Bayesian Structural Equation Modeling

In this section, we provide a brief introduction to Bayesian structural equation modeling to establish the foundation for the rest of the study.

Structural Equation Model

The basic structural equation model (SEM) describes the relationship between observed and latent variables. A SEM model consists of two parts: the structural model and the measurement model. Latent variables with their observed indicators constitute the measurement model. The relations among latent and observed variables that are not indicators constitute the structural model. The inclusion of latent variables with their measurement model and indicators allows the model to account for measurement error. Variables can be exogenous (not regressed on anything) or endogenous, depending on their structural positioning in the model. Structural modeling requires certain underlying assumptions to be satisfied to ensure the accuracy of results — specifically, multivariate normality, no systematic missing data, sufficient sample size obtained via random sampling, and correct model specification. The violation of the multivariate normality assumption leads to underestimated standard errors and overestimated likelihood ratio statistic. The correct model specification implies that no important parameters that are present in

the population are omitted from the model. The mathematical expression for a basic SEM model in the LISREL (Jöreskog & van Thiilo, 1972) notation is as follows:

$$y = \Lambda_y \eta + \epsilon \tag{1}$$

$$x = \Lambda_x \xi + \delta \tag{2}$$

$$\eta = B\eta + \Gamma\xi + \zeta, \tag{3}$$

where y and x are the observed indicators; Λ_y is the matrix of factor loadings for endogenous latent factors, Λ_x — for exogenous latent factors; ϵ and δ are measurement errors; η and ξ are the vectors of endogenous and exogenous latent factors, respectively; B is the matrix connecting endogenous latent factors together; Γ is the matrix connecting observed and latent variables, or exogenous and endogenous latent factors together, if the structure is only based on latent variables; and ζ is the error term. Therefore, Equations 1 and 2 represent the measurement model, and Equation 3 represents the structural model (defines how the latent variables are related to each other and to observed variables that are not indicators for latent variables.

In this study, we will only use the measurement model, which reduces the above equations down to:

$$x = \Lambda_x \xi + \delta, \tag{4}$$

where all the notation remain the same, but there is no difference between exogenous and endogenous variables. The underlying assumption for these parameters is that factor loadings are fixed values consistent across time and individuals, error terms follow a normal distribution, and latent factors follow a multivariate normal distribution.

The measurement model also has a specific covariance structure defined by:

$$\sum(\theta) = \Lambda_x \Phi_\xi \Lambda_x' + \Theta_\delta, \tag{5}$$

where $\sum(\theta)$ is the covariance matrix of observed indicators x, Λ_x is a factor loading matrix, Φ_ξ is the latent factor variance-covariance matrix, and Θ_δ is the error term covariance matrix. There are several assumptions about these parameters. First, the mean of the error is assumed to be zero: $E(\delta = 0)$. Second, δ is assumed to be uncorrelated with ξ. Some constraints are also imposed, such as the Θ_δ is specified as a diagonal matrix, meaning that no covariances are allowed between error terms. Specifically for bifactor models, no covariances between latent factors are allowed, which reduces Φ_ξ down to a diagonal matrix.

Bayesian Structural Equation Model

The term Bayesian Structural Equation model (BSEM) denotes the estimation of structural equation models, such as SEM, CFA and ESEM, using Bayesian estimation. This framework estimates the posterior distribution for all model parameters as a product of likelihood of model parameters and prior knowledge about the distribution of model parameters in the form of priors. In model interpretation, the central tendency measure of the posterior distribution, such as the mean, can be used as a point estimate for the parameter. Prior distribution can be specified for all model parameters, for simplicity, denoted as $\theta = (\eta, \Lambda, \Phi, \Theta)$, where factor loadings constitute the Λ factor matrix, covariances between factors are stored in the Φ matrix, residual variances of latent factors (σ^2_ξ), and residual variances of the indicators ($\sigma^2_{\varepsilon_j}$), both constitute the Θ matrix. The process of

Bayesian estimation is described through the following equation:

$$p(\theta|Y) \propto p(Y|\theta)p(\theta), \tag{6}$$

where $p(\theta|Y)$ is the posterior distribution of model parameters given the observed data Y, $p(Y|\theta)$ is the probability of the observed data given model parameters θ and $p(\theta)$ is the prior for parameters θ. The probability $p(Y|\theta)$ is based on parameters θ estimated using the likelihood function $L(\theta|Y)$. The most widely used function is the maximum likelihood (ML) defined as:

$$F_{ML} = \log|\sum(\theta)| + tr[S\sum^{-1}(\theta)] - \log|S| - q, \tag{7}$$

where S is the sample covariance matrix and $\sum(\theta)$ is the model-implied covariance matrix of structural parameters, both these terms are assumed to be non-singular. Essentially, minimizing this function results in the estimates of the model-implied covariance matrix close to the sample covariance matrix.

As mentioned before, the posterior is formed using both likelihood and prior information. In BSEM, normal priors are commonly used for factor loading parameters and path coefficients between latent variables:

$$\lambda_x \sim \mathcal{N}(\mu_{\lambda_x}, \sigma^2_{\lambda_x}), \tag{8}$$

$$\gamma, \beta_k \sim \mathcal{N}(\mu_{\lambda_{\gamma\beta_k}}, \sigma^2_{\gamma\beta_j}). \tag{9}$$

By default, for loadings and item intercepts, diffuse (non-informative) priors are used. These are normal priors centered at 0 with an infinite variance: $\lambda_x \sim \mathcal{N}(0, \infty)$. By adjusting the mean hyperparameter (μ) and variance hyperparameter (σ^2), we can adjust the normal priors' accuracy (mean) and informativeness (degree of

spread or variance). This achieves flexibility in estimation. For example, instead of restricting cross-loadings to be exactly zero, we can estimate a model where cross-loadings are assigned a highly-informative prior with $\mu = 0$ and $\sigma^2 = 0.01$. If the modeled parameter is substantially different from zero, with sufficient sample size, the data will contribute more to the estimation and the central tendency of the posterior will be close to the population value (Asparouhov et al., 2015).

For the residual variance parameters ($\sigma^2_{\xi/\eta_Y/\varepsilon_j}$), the inverse-Gamma ($\mathcal{IG}$) prior can be utilized:

$$\sigma^2_{\xi/\eta_Y/\varepsilon_j} \sim \mathcal{IG}(\kappa, \nu), \tag{10}$$

where κ and ν are the shape and scale hyperparameters. Shape parameter controls the height, the scale — the spread of the distribution. Modifying these hyperparameters will adjust the informativeness of the $\mathcal{IG}$ prior. The non-informative prior $\sim \mathcal{IG}(-1, 0)$ is used in M*plus* as a default and will be used in this study.

For the variance-covariance matrix, the inverse Wishart prior can be specified (Liu et al., 2016):

$$\Phi_\xi \sim \mathcal{IW}(V, m), \tag{11}$$

where V is the scale matrix, and m is degrees of freedom for a $p \times p$ variance-covariance matrix. The default prior for the variance-covariance matrix parameter in M*plus* is a non-informative inverse Wishart prior: $\sim \mathcal{IW}(0, -p - 1)$, with p being the dimension of the multivariate block of latent variables.

Markov Chain Monte Carlo (MCMC)

We implement the Bayesian estimation of SEM models via M*plus* software in the current study (L. K. Muthén & Muthén, 2017). The data is analyzed using Markov Chain Monte Carlo (MCMC) method. The algorithm used in this study is the M*plus* default Gibbs sampler algorithm (Asparouhov & Muthén, 2010).

By default, two Markov chains are generated for each parameter, each with a user-specified number of iterations, and the initial 50% of these iteration samples (so-called burn-in phase) are discarded to reduce the influence of the initial values and to reach a stable distribution. The remaining MCMC samples after burn-in approximate the posterior distribution $p(\theta|Y)$, on which Bayesian estimates are based. In this study, these default settings will be employed and the posterior mean will be used as a Bayesian point estimate for the parameters.

To assess convergence, the built-in M*plus* Potential Scale Reduction Factor (PSR, PSRF) (also referred to as $\hat{R}$) is used. This criterion assesses the chains for every estimated parameter and calculates the between- and within-chain variance. If the model is reaching convergence, it means two chains are arriving at similar results, which is indicated by a small between-chain variance. Alongside that, the chain has to explore all the posterior distribution, which would be indicated by a large within-chain variance. Therefore, between-chain variance should be smaller than within-chain and total variance. The $\hat{R}$ shows the ratio of those two variances, where the smaller $\hat{R}$ will indicate larger ratio of within-chain variance to between-chain variance, and thus, better convergence. In M*plus*, the largest $\hat{R}$ across all parameters at a given iteration is reported. Therefore, if by the last iteration the reported $\hat{R}$ is below the cutoff value, it indicates that the convergence for all model parameters has been achieved (for the $\hat{R}$ computation, see Brooks & Gelman, 1998).

In the current study, we will use the value $\hat{R} = 1.05$ as the cutoff for convergence. Since there are no guidelines for convergence cutoff for Bayesian bifactor models, this value was chosen as a compromise between a strict criterion ($\hat{R} = 1.01$), recommended specifically for Bayesian CFA models (Brown, 2015; Depaoli & Van de Schoot, 2017) and a more commonly recommended $\hat{R} = 1.1$ (Gelman et al., 2004; Vats & Knudson, 2021).

Simulation Design

The aim of this research is to evaluate the performance of Bayesian inference in estimating a bifactor model with cross-loadings and test the sensitivity of Bayesian fit indices in detecting misspecification in different parts of the model, both using different Bayesian prior settings. This section describes the design of the simulation study with regards to this aim. The population model and design factors with their levels are explained, and the design is further summarized in Table 1.

The design comprises three parts: a population (data generation) model, an analysis model (fitted model) and priors. The population model may vary depending on whether a general factor and cross-loading are present. The analysis model may be with or without a general factor and cross-loadings, and they may also vary in terms of priors.

The Population Model Structure

We specified one population model structure as a basis that can change depending on the specification of population general factor loadings and population cross-loadings, which will be explained in corresponding sections. This population model is a bifactor structure with 12 indicators $x_1 - x_{12}$ that load onto one general factor η_G and 3 specific factors $\eta_{S_1}, \eta_{S_2}, \eta_{S_3}$. This setup is often employed in previous studies, such as a recent study by Ximénez et al. (2022). All indicators load onto a

general factor η_G. Simultaneously, indicators $x_1 - x_4$ load onto the first specific factor η_{S_1}, indicators $x_5 - x_8$ load onto the second specific factor η_{S_2}, and indicators $x_9 - x_{12}$ — onto the third specific factor η_{S_3}. In addition to these loadings, we specified one cross-loading to each specific factor: 1) a cross-loading from x_1 to the second specific factor $\lambda_{C_{1,2}}$, 2) a cross-loading from x_5 to the third specific factor $\lambda_{C_{5,3}}$., and 3) a cross-loading from x_9 to the first specific factor $\lambda_{C_{9,1}}$. This structure is shown in Figure 1. Solid lines represent main loadings (general and specific factor loadings), and dashed lines represent cross-loadings.

General factor loadings

To test how the strength of the general factor loadings affects the performance of Bayesian inference, we consider 5 levels of this factor: 0, 0.10, 0.30, 0.40, 0.60. We restrict all population general factor loadings to be equal. When general factor loadings are 0, data will be generated according to the structure in Figure 1, but the general factor and corresponding loadings will be absent, which will reduce the model down to a correlated factors structure (in Figure 1 the structure with a general factor is shown, so the specific factor correlations are not shown). This level is included to test the sensitivity of fit measures to detect misspecification in the general factor part. The next level is 0.10, which represents a parameter that is close, but still substantially different from 0. Levels 0.30 and 0.40 represent low and moderate general factor loadings, respectively. Level 0.60 represents large general factor loadings, according to previous studies (Guo et al., 2019; Ximénez et al., 2022). Manipulating general factor loadings will result in different magnitude of general factor–related model complexity, usually evaluated based on the estimated factor loadings through Explained Common Variance (ECV). ECV describes the percentage of variance explained by the general factor in the total variance

(Rodriguez et al., 2016; Gu et al., 2023):

$$ECV = \frac{\sum \lambda_g^2}{\sum \lambda_g^2 + \sum \lambda_s^2 + \sum \lambda_c^2}, \tag{12}$$

where λ_g refers to general factor loadings, λ_s — specific factor loadings, and λ_c — cross loadings.

In our study, the population ECV varies with varying the magnitude of general factor loadings, whereas varying cross-loadings does not produce any substantive change in ECV. Specifically, general factor loadings 0.1 result in ECV of 0.02, loadings 0.3 produce ECV of 0.20, and loadings of 0.6 suggest an ECV of 0.50, indicating that 50% of the variance is explained by the general factor. Simulation conditions with smaller ECV reflect the non-significant general factor that may be omitted from the model, while the moderate ECV of 0.50 suggests that both general factor and specific factors are important and should be modeled simultaneously (Rodriguez et al., 2016).

Cross-loadings

To test how the presence and strength of cross-loadings affects the performance of Bayesian inference estimating a bifactor model, we consider 4 levels of this factor: 0, 0.10, 0.20, 0.30. We restrict all population cross-loadings to be equal. When cross-loadings are 0, the data will be generated according to the structure in Figure 1, but the dashed lines (cross-loadings) will be absent, which will reduce the structure down to a regular bifactor structure. This level is included to test the sensitivity of fit indices in detecting misspecification in the cross-loadings part. The other levels correspond to the literature, where a 0.1 cross-loading is considered to be of little importance, a cross-loading of 0.2 of some importance, and a cross-loading of 0.3 of importance (Cudeck & O'dell, 1994). This setup is employed

in similar studies (Guo et al., 2019; Ximénez et al., 2022).

Specific Factor Loadings, Factor Correlations and Residual Variances

Since we are primarily interested in the influence of the strength of general factor and cross-loadings on Bayesian inference, we considered holding specific factor loadings constant and setting their values at 0.6, as well as restricting them to be equal, for all simulation conditions. All factor correlations were set to zero in a bifactor model to satisfy the orthogonality constraint, but in conditions when general factor loadings are zero and this constraint is no longer required, the correlations between factors are set to 0.5 to represent a moderate correlation. Residual variances for each indicator were set using the $1 - \lambda_G^2 - \lambda_S^2 - \lambda_C^2$ formula, to result in unit variance for each indicator.

Sample Size

To test the robustness of Bayesian estimation to sample size, we consider 4 levels for this design factor: 100, 250, 500, 1000. This corresponds to common sample sizes for social sciences and adequate power for the frequentist analogue of our model. The level 100 corresponds to a very small sample size — just about enough to identify a model with 4 indicators per factor and 2 uncorrelated factors (Boomsma, 1985; Marsh & Hau, 1999). The level 250 corresponds to an adequate sample size according to the commonly accepted rule of 10 cases per an indicator (Thorndike, 1995). Since we have 12 indicators, but they are used for both specific and general factors, this becomes about 20 cases per indicator, which gives us 240 participants as a minimum sample. The other two levels are 500 and 1000, which represents twice and 4 times as large as the minimum required sample, respectively.

The Analysis Model

For each generated dataset we fit 3 analysis models: 1) a full model (M_1) with cross-loadings and a general factor; 2) a no cross-loadings model with general factor (M_2); and 3) a no general factor model with cross-loadings (M_3). We fixed factor variances at 1 for identification, allowing factor loadings to be freely estimated. Depending on the presence of general loadings and cross-loadings in the population structure, an analysis model can be true or misspecified. For example, if the general factor loadings are 0.30 and cross-loadings are zero in the population, the analysis model M_1 is an overfit model, the analysis model M_2 is a true model, and M_3 is misspecified by ignoring the true general factor and including extra cross-loadings.

Prior

The models M_1, M_2 and M_3 were estimated with default prior settings in M*plus*, including diffuse priors (centered at zero with a very large variance) $\lambda_x \sim \mathcal{N}(0, \infty)$ for all factor loadings.

To test the influence of accuracy and informativeness of priors on Bayesian estimation of a bifactor model with cross-loadings, as well as the ability of near-zero priors to test the hypotheses about general factor, additional prior settings for general factor loadings were considered. In the first situation, near-zero priors were specified. Near-zero prior is centered at zero with a small variance, such as 0.1: $\lambda_x \sim \mathcal{N}(0, .1)$. In this case, the zero mean of the prior reflects our belief that the parameter is likely zero, but the small variance reflects the uncertainty we have in the parameter being truly zero, by allowing the estimates to fluctuate within a small distance around the mean. Possible variance hyperparameters include: .01, .1 (for more details on BSEM with near-zero priors, see B. Muthén & Asparouhov, 2012). In our study, we considered two near-zero prior settings that differ in variance: 1) a

prior $\lambda_x \sim \mathcal{N}(0, .1)$, reflecting high certainty, and a second $\lambda_x \sim \mathcal{N}(0, .01)$, reflecting very high certainty in the parameter being zero. These settings serve multiple purposes. First, to ensure that we provide enough information for the model with both cross-loadings and a general factor to be identified. Second, to evaluate how these settings perform in detecting the parameter in the situation of uncertainty. To reiterate from the previous parts of the manuscript, if a researcher is not sure that general factor should be included, but is sure that cross-loadings exist, he can assign near-zero priors to general factor loadings while freely estimating several cross-loadings. In this case, if general factor indeed exists, given sufficient sample size, the posterior means will be close to population values, which will give a researcher a comprehensive picture of their data in just one iteration of model fitting.

In the second situation, we considered weakly-informative accurate priors, to reflect the situation where a researcher has a strong theoretical evidence of a presence of a general factor and has a guess about the population values for general factor loadings based on previous studies. The hyperparameters of that prior are defined using the following formula: $\lambda_x \sim \mathcal{N}(true, 0.5 \times true)$, where the prior is centered in the population value (defined by a corresponding design factor) and the variance of the prior equals half the true value of the parameter (Depaoli, 2013).

All in all, this design allows to contrast different magnitude of a general factor with different level of cross-loadings in a BSEM model, using different prior settings to estimate a general factor. For simplicity and to prevent the occurrence of meaningless simulation cells, we treat analysis model and priors as one design factor with six levels. This design is summarized in Table 1. The combination of these design factors yields: $5 \times 4 \times 4 \times 6 = 480$ unique cells (simulation conditions) in total.

MCMC Specifications

Each simulation condition was replicated 500 times with 20,000 iterations. Two chains were used for the posterior. The first 50% of iterations were discarded as a burn-in phase. Before proceeding with the results, the convergence of all simulation cells was assessed according to the procedure described in the Markov Chain Monte Carlo section above.

Simulation Result I: Model Parameter Recovery

In this section we assess the performance of Bayesian approach in estimating the bifactor model with cross-loadings, focusing on accuracy of the central tendency and variability of the posterior distribution for each parameter when the true model is estimated. We also briefly talk about the influence of model misspecification and priors on estimation accuracy. The reported results are based on converged replications.

Evaluation Criterion

We assess the central tendency of the posterior distribution, the coverage rates of the posterior 95% credible intervals, and the Bayesian "power" for detecting the presence of a general factor and cross-loading when they indeed exist in the population model. In the following, we first explain how each statistic is calculated for the evaluation. We then report the results for each statistic. Based on the types of parameters, we organize the results into groups: for general factor loadings, cross-loadings, and specific factor loadings, respectively.

Bias and Mean Squared Error (MSE)

Bias is the measure of how the point estimate differs from the true population value of a parameter. Let θ be a parameter and its true value, and $\hat{\theta}_r$ be the

559 posterior mean in the r^{th} replication, which acts as the point estimate for the
560 parameter θ in r^{th} replication. The notation $\bar{\theta}$ represent the average of parameter
561 estimates across R converged replications and it is calculated as follows:

$$\bar{\theta} = \frac{1}{R} \sum_{r=1}^{R} \hat{\theta}_r, \tag{13}$$

562 where R is the number of converged replications and $\hat{\theta}_r$ is the posterior mean of the
563 parameter θ in the r^{th} replication.

564 In turn, relative bias is a ratio of bias, calculated as the difference between the
565 point estimate presented above and the population value, to the absolute population
566 value,

$$\text{Relative bias}_{\theta} = \begin{cases} \frac{\bar{\theta}-\theta}{|\theta|} & \text{if } \theta \neq 0 \\ (\bar{\theta} - \theta) & \text{otherwise.} \end{cases} \tag{14}$$

567 If the population value is zero, bias is simply a difference between a population
568 value and an estimated value, or so-called "raw bias", otherwise the difference is
569 normed on the absolute population value, resulting in relative bias.

570 In addition to bias, which quantifies the deviation of a point estimate from true
571 value, we can also compute the mean squared error (Morris et al., 2019). This index
572 reflects the stability of an estimate across replications using the variance of the
573 estimated parameter across replications. Let $\hat{\theta}_r$ be the point estimate of a parameter
574 in replication r and $\bar{\theta}$ the mean of the point estimates in the replications. The MSE

is computed as follows

$$MSE = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta)^2 \tag{15}$$

$$= \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \bar{\theta} + \bar{\theta} - \theta)^2 \tag{16}$$

$$= \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \bar{\theta})^2 + \frac{1}{R}\sum_{r=1}^{R}(\bar{\theta} - \theta)^2 \tag{17}$$

$$= s^2 + (Bias)^2, \tag{18}$$

Hence, the MSE for a given parameter is simply the squared raw bias of that parameter estimate with added variance of the point estimates. Essentially, MSE in a simulation study reflects the potential replicability of results in applied research, where each replication reflects a separate applied study.

Bayesian Coverage Rates

A Bayesian credibility interval (CI) represents a range within the posterior distribution that contains a specified percentage of the parameter's values. It is defined by two boundary points, between which the designated proportion of the parameter values is located. In practice, the 95% credibility interval is often employed, within which 95% of parameter values are.

The equal-tailed 95% credible interval is defined such that the posterior probability of the parameter θ lying below the lower bound L_r is 2.5%, and the probability of it lying above the upper bound R_r is also 2.5%. Mathematically, this

589 is expressed as follows:

$$P(\theta \leq L_r \mid \text{data}) = 0.025 \tag{19}$$

$$P(\theta \geq R_r \mid \text{data}) = 0.025 \tag{20}$$

$$P(L_r \leq \theta \leq R_r \mid \text{data}) = 0.95. \tag{21}$$

590 The coverage rate of the 95% credible interval is an index that measures how well
591 the posterior distribution captures the true parameter value across repeated
592 samples. It is calculated as follows:

$$CR = \frac{1}{R}\sum_{r=1}^{R} I\Big(\theta \in [L_r, R_r]\Big). \tag{22}$$

593 Specifically, it represents the proportion of times the true parameter falls within the
594 computed 95% credible interval across all converged replications. A coverage rate
595 close to 95% indicates that the credible interval accurately reflects the variability of
596 the posterior distribution, while severe deviations may indicate inaccuracies in the
597 posterior distribution.

598 ***Bayesian "Power"***

599 To evaluate the effectiveness of Bayesian estimation methods in detecting the
600 true cross-loadings and general factor loadings, we can calculate the proportion of
601 replications in which their 95% credible interval excludes zero. This index quantifies
602 the probability of correctly identifying a non-zero parameter, resulting in a Bayesian
603 analogue of frequentist "power". The Bayesian analogue of "power" is computed as
604 follows:

$$\text{power} = \frac{1}{R}\sum_{r=1}^{R} I\Big(0 \notin [L_r, R_r]\Big). \tag{23}$$

Overall, using these four measures of model recovery, we can assess how well the analysis model recovers the parameters of the corresponding population model, therefore fulfilling the first aim of our study.

Results

This section is organized as follows. First, we evaluate the convergence rate of all conditions. Second, we assess the accuracy of a point estimate reporting (relative) bias and mean squared error for general factor loadings, cross-loadings and specific factor loadings, respectively. Third, we report the Bayesian power for detecting general factor loadings, cross-loadings and specific factor loadings. Finally, coverage rates of the 95% credible intervals are outlined to see how often the true value is within a 95% of the posterior, which ensures reliability of Bayesian inference. We then conclude this section with a summary of the adequacy of the true model recovery.

We plot the results across simulation factors. Rows and columns in the graphs represent population general factor loadings and cross-loadings, respectively. The x-axis differs depending on the model part for which the graph is plotted. By default, all analysis models are shown as follows: the full model with $\Lambda_G, \Lambda_C \sim \mathcal{N}(0, \infty)$; , the full model with $\Lambda_C \sim \mathcal{N}(0, \infty)$, $\Lambda_G \sim \mathcal{N}(0, 0.1)$, the full model with $\Lambda_C \sim \mathcal{N}(0, \infty)$, $\Lambda_G \sim \mathcal{N}(0, 0.01)$, the full model with weakly informative accurate priors $\Lambda_G \sim \mathcal{N}(true, 0.5 \times true)$, $\Lambda_C \sim \mathcal{N}(0, \infty)$, the model without cross loadings $\Lambda_G \sim \mathcal{N}(0, \infty)$ and the model without general factor $\Lambda_C \sim \mathcal{N}(0, \infty)$. However, for plots referring to general factor loadings (such as Figure 3), the no general factor model is not shown, the same with cross-loadings and the no cross-loadings model.

Convergence Rates

Convergence rates are shown in Figure 2. We can see that most simulation conditions, especially full models with informative priors and the no general factor model, demonstrate rates of 95% and higher for any combination of population parameters. The full model with diffuse priors and a model without cross-loadings show lowered convergence rates for certain conditions. Specifically, for the full model with diffuse priors, for population general factor loadings of 0.3 and less, convergence is between 80 and 90 percent. For the no cross-loadings model, convergence rate is between 90 and 80 percent when population cross-loadings are 0, and also when general factor loadings are 0.3. Convergence rate for this model drops even more when cross-loadings in the population are larger than 0.2 and general factor loadings are 0 or 0.1, reaching only 50% rate.

Overall, the analysis models have high convergence rates when informative priors are specified, while the true model with diffuse priors has slightly lower convergence rates. Misspecification in the cross-loadings part leads to a severe drop in the convergence rate. Additional tests increasing number of iterations to 40,000 did not improve convergence for the aforementioned cells and therefore are not reported here. In the further analysis, only converged replications will be used.

Raw and Relative Bias

Bias of General Factor Loadings. First, we note that items 1, 5 and 3 in our study can have cross-loadings, while all other items can only have specific factor loadings and general factor loadings. Observing the pattern of bias, we noticed that items that can have cross-loadings behave differently, compared to other items. This difference is manifested in elevated relative bias of items with cross-loadings, when cross-loadings are omitted from the analysis (underfit) This is shown in Table 2.

The table shows relative bias of general factor loadings for each item in columns, when the no cross-loadings model is fitted. Each row represents different combinations of sample size, magnitude of general factor and cross-loadings in the population. We can see that as the cross-loadings increase, for any sample size, general factor loadings are severely overestimated, with relative bias exceeding 10%. These values are boldface type. To reiterate, this pattern of bias is observed only for items 1, 5 and 9.

The pattern of general factor loading bias for the rest of the items is presented in Figure 3. The bias values are averages across items 2-4, 6-8 and 10-12. Rows and columns represent population general factor loadings and cross-loadings, respectively. The x-axis shows different analysis models and priors. The red dashed lines represent the $\pm 10\%$ cutoffs for relative bias. The first row represents no general factor in the population, and, therefore, raw bias of general factor loadings. No cutoffs for raw bias exist, but we can see that it stays close to zero for all models. The first column represents no cross-loadings in the population, and for that column, the true model is the "no cross" model. We see that this model results in acceptable bias for general factor loadings, except when their population values are small (0.1). In that case, small sample ($N = 100$) leads to overestimation, and large samples ($N > 500$) lead to underestimation. For the rest of rows and columns the true model is a full model with diffuse priors, while the no cross-loadings model is an underfit model. The true model is estimated with acceptable bias of general factor loadings, except when their population values are small (0.1). In that case, small sample ($N = 100$) leads to overestimation, and large samples ($N > 500$) lead to underestimation. As we already saw in Table 2 the underfitting no cross-loadings model overestimates general factor loadings, but only for items that have cross-loadings. However, Figure 3 shows that general factor loadings for the rest of

the items can also be overestimated, but only when population general factor loadings are small and cross-loadings are large (fourth and fifth columns and second row).

Concerning the role of priors, we can see that near-zero small variance priors on general factor loadings lead to underestimation for all population magnitudes of the loadings. The prior with variance 0.1 leads to 50% underestimation, and with variance 0.01 - 100% underestimation. This effect decreases with increasing sample size and population values for general factor loadings.

Considering the stability of results across replications, MSE demonstrated in Figure 4 suggests that the estimation error is relatively large for the highly informative near-zero priors and very small for all other analysis models and population conditions, suggesting the stability of results. In the situation of near-zero priors, the error gets smaller with increasing sample size, just as relative bias.

Overall, the true model results in very low relative bias of general factor loadings, except when the loadings are small (0.1). In this situation, small samples lead to overestimation, and large - to underestimation. Sample size is overall not related to the size of bias, but when an informative inaccurate prior is specified, larger samples reduce the bias. These results are stable and replicable, as suggested by MSE.

Bias of Specific Factor Loadings. As mentioned before, items 1, 5 and 3 can have cross-loadings, while all other items can only have specific factor loadings and general factor loadings. These three items have a different pattern of bias, specifically, elevated bias when cross-loadings are omitted from the analysis (underfitting). This is shown in Table 3. The table shows relative bias of specific factor loadings for each item, when the no cross-loadings model is fitted. Each row

represents different combinations of sample size, magnitude of general factor and cross-loadings in the population. Columns represent the items. We can see that as cross-loadings increase, for any sample size, specific factor loadings are severely underestimated, with relative bias exceeding 10%. These values are boldface type. This pattern is observed only for items 1, 5 and 9.

Figure 5 shows bias computed for specific factor loadings for the rest of the items, averaged across these items. Rows and columns represent population general factor loadings and cross-loadings, respectively. The x-axis shows different analysis models and priors. The red dashed lines represent the $\pm 10\%$ cutoffs for relative bias.

For the first row of Figure 5, the true model is the no general factor model. It results in zero bias for specific factor loadings. For the first column, the true model is the no cross-loadings model. It results in zero bias for specific factor loadings. For the rest of rows and columns the true model is a full model with diffuse priors. No cross-loadings and no general factor are underfitting models. The true model shows average relative bias approaching 0%. The underfitting model omitting cross-loadings shows no influence on the specific factor loadings bias, it stays very close to zero. The underfit model omitting general factor shows extreme overestimation of specific factor loadings with relative bias reaching $30 - 40\%$, increasing as the magnitude of population general factor loadings increases.

Concerning the role of priors, we can see that near-zero small variance priors on general factor loadings lead to the same pattern of inflation of specific factor loadings as simply omitting the general factor when it is present in the population. The prior with variance 0.1 leads up to 20% overestimation, and with variance 0.01 - up to 50% overestimation. This effect decreases with increasing sample size and increases with increasing population values of general factor loadings.

Considering the stability of the results, MSE demonstrated in Figure 6 suggests

that the estimation errors are relatively large only when general factor is omitted or estimated with highly informative inaccurate priors placed on general factor loadings. For all other conditions, MSE is small, suggesting the stability and replicability of results.

Overall, the true model results in very low relative bias for specific factor loadings, for any population values of other model parameters. Underfit in general factor part, either due to simply not modeling it, or specifying a highly informative near-zero prior, leads to overestimation of specific factor loadings and increased MSE. Underfit in the cross-loadings part leads to underestimation of specific factor loadings, but only for items that can have cross-loadings. Sample size is overall not related to the size of bias, but when an informative inaccurate prior is specified for general factor loadings, large samples prevent overestimation of specific factor loadings.

Bias of Cross-loadings. As mentioned previously, in our study, items 1, 5 and 9 can have cross-loadings. Figure 7 demonstrates bias computed for cross-loadings, averaged across these 3 items. Rows and columns represent population general factor loadings and population cross-loadings. The x-axis shows different analysis models and priors. The red dashed lines represent the $\pm 10\%$ cutoffs for relative bias.

The first column represents no cross-loadings in the population, and, therefore, raw bias. For all analysis models, bias revolves around zero. For the first row, the true model is the no general factor model. For this model, relative bias of cross-loadings is zero. For the rest of rows and columns, the true model is the full model with diffuse priors, and an underfit model is the no general factor model. When population cross-loadings are small (0.10), relative bias exceeds 10% cutoff, reducing closer to zero as sample size increases. For larger cross-loadings, true model results in almost zero relative bias. The underfit model results in zero bias

for cross-loadings, because, as we saw in Figure 5, the unmodeled variance from a general factor already inflated specific factor loadings.

The choice of prior on general factor loadings influences the estimation of cross-loadings in certain conditions. Specifically, if we have small population cross-loadings (0.1), large population general factor loadings and specify an informative inaccurate near-zero prior for general factor loadings, cross-loadings are overestimated up to 75%. Importantly, when general factor is simply omitted from the model, we do not observe such bias inflation for cross-loadings.

Considering the stability and replicability of results, MSE demonstrated in Figure 8 suggests that the estimation errors are relatively large when highly informative inaccurate near-zero priors are placed on general factor loadings and the variance of a general factor inflates the cross-loadings. Also, there is a noticeable elevation of MSE for sample size 100, compared to larger samples. For the bias itself, this pattern is not observed, suggesting that small sample of 100 observations does not ensure as much stability of cross-loadings estimation, as larger samples do.

Bayesian "Power"

"Power" of General Factor Loadings. Figure 9 demonstrates Bayesian "Power" to detect a non-zero general factor loading. The values shown are averages across all items. Rows and columns represent population general factor loadings and population cross-loadings, respectively. The x-axis shows different analysis models and priors. The red dashed line represents the $\pm 80\%$ "power" level that we consider sufficient. For the first row, all models are overfitting models. We can see that "power" to detect a general factor loading for all models is zero. In this case, "power" can be interpreted as a Bayesian analogue of Type I error rate. However, Type I error rate increases when no cross-loadings analysis model is fitted when

population cross-loadings are large. This is the reflection of inflated general factor loadings. The same pattern of "power" is observed for the second row that represents small general factor loadings (0.1). In this case, the analysis model does not recognize the population general factor, and the "power" to detect a non-zero loading is zero. For the first column, the true model is the no cross-loadings model. For that model, the "power" to detect general factor loadings depends on 1) sample size and 2) magnitude of population general factor loadings. When both increase, "power" increases. To reach the 80% "power" for population general loadings of 0.3, sample size 500 is enough. To detect 0.4 population loadings – 250 observation is enough. For loadings 0.6 – 100 observations is enough. For the rest of rows and columns, full model with diffuse priors is the true model. The no cross-loadings model is an underfitting model. For the true model, the "power" to detect general factor loadings depends on 1) sample size and 2) magnitude of population general factor loadings. When both increase, "power" increases. To reach the 80% "power" for population general loadings of 0.3, sample size 500 is enough. To detect 0.4 population loadings – 250 observation is enough. For loadings 0.6 – 100 observations is enough. The underfitting model behaves the same way, showing that misspecification in the cross-loadings part does not affect "power" to detect general factor loadings, when they are medium and large magnitude.

Considering the influence of the prior, "power" is very low when the inaccurate near-zero prior $\sim \mathcal{N}(0., 01)$ is specified for large general factor loadings, however, as expected, this prior influence is diminished by a large sample size (N=1000).

"Power" of Specific Factor Loadings. Figure 10 demonstrates Bayesian "power" to detect a non-zero specific factor loading. The values shown are the averages across all items. Rows and columns represent population general factor loadings and cross-loadings, respectively. The x-axis shows different analysis models

and priors. The red dashed line represents the 80% "power" level that we consider sufficient. For the first row, the true model is the no general model. It results in over 80% "power" to detect specific factor loadings. The underfitting no cross-loadings model leads to lowered "power" to detect specific factor loadings, when population cross-loadings are large. Sample sizes over 500 observations cancel this effect. For the first column, the true model is the no cross-loadings model. It results in acceptable "power", but it drops below 80% when population general factor loadings are large (0.4 and higher). For the rest of rows and columns the true model is the full model with diffuse priors. For the true model, the "power" is adequate for all population conditions and sample size 250 and higher, only dropping slightly below 80% when sample size is 100 and general factor loadings are high. The same pattern is observed for the underfitting no cross-loadings model. The underfitting no general factor model results in sufficient "power" regardless of sample size or population parameters. Finally, the choice of priors on other parts of the model causes some fluctuations in "power" to detect specific factor loadings, but it never drops below the desired 80%.

"Power" of Cross-loadings. Figure 11 demonstrates Bayesian "power" to detect a non-zero cross-loading. The values shown are the averages across all items that have cross-loadings. Rows and columns represent population general factor loadings and population cross-loadings, respectively. The x-axis shows different analysis models and priors. The red dashed line represents the 80% power level that we consider sufficient. The first column represents zero population cross-loadings and both full and no cross-loadings models with diffuse priors are overfitting models. Therefore, the "power" turns into the analogue of Type I error rate. We can see that it is zero for all models, meaning correctly detecting cross-loadings as 0, regardless of overfitting. The first row represents no general factor in the population, therefore

the true model is the no general factor model. For that model, "power" to detect cross-loadings depends on 1) sample size and 2) the magnitude of cross-loadings. Sufficient "power" is achieved only when cross-loadings are 0.2 and higher and sample size is 500 and higher. For the rest of rows and columns, the true model is the full model with diffuse priors. For this true model, the "power", again, depends on sample size and the magnitude of population cross-loadings. With cross-loadings 0.1 and smaller, sufficient "power" is never achieved. For cross-loadings 0.2, sample size 500 is enough to reach 80% "power". For cross-loadings 0.3, sample size 250 is enough, but sample of 500 and larger appears to produce more robust results.

Coverage Rate

Coverage of General Factor Loadings. Figure 12 demonstrates Bayesian coverage of general factor loadings. The values shown are the averages across all items. The first row represents no general factor in the population. The underfitting no cross-loadings model reduces the coverage of general factor loadings, especially when population cross-loadings are large. The same pattern is observed for the second row, representing very small population general factor loadings. For the first column, the true model is the no cross-loadings model. Coverage of general factor loadings of that model is above 95%. For the rest of rows and columns, the true model is the full model with diffuse priors. The underfitting model is the no cross-loadings model. The true model shows 95% coverage for all population conditions. The underfit model results in a drop of coverage below the 95% level. This effect increases with the increase of population cross-loadings magnitude and sample size.

Concerning the role of priors, specifying a near-zero prior $\Lambda_G \sim \mathcal{N}(0, 0.01)$ for non-zero population general factor loadings results in a drastic drop of coverage to 0. In contrast, a less informative near-zero prior $\Lambda_G \sim \mathcal{N}(0, 0.1)$ only slightly

decreases coverage, and only when there are less than 250 observations in the data.

Coverage of Specific Factor Loadings. Figure 13 demonstrates Bayesian coverage of specific factor loadings. The first row represents no general factor in the population. The true model is the no general factor model and it results in coverage of 95%. The misspecified model is the no cross-loadings model, and fitting that model results in a drop of coverage below 95%. The same is true for the second row, representing very small general factor loadings in the population. For the first column, the true model is the no cross-loadings model, and it results in perfect 95% coverage. For the rest of rows and columns the true model is the full model with diffuse priors, and it results in perfect 95% coverage. The underfitting no cross-loadings model results in a drop of coverage to 75%, but only when sample size is large and population cross-loadings are large. The underfitting no general factor model results in a severe drop of coverage to 0%, increasing with increasing sample size and population general factor loading values.

Specifying an inaccurate near-zero prior $\Lambda_G \sim \mathcal{N}(0, 0.01)$ for large population general factor loadings leads to a drop of coverage of specific factor loadings to 0%. A more relaxed near-zero prior $\Lambda_G \sim \mathcal{N}(0, 0.1)$, in contrast, does not lead to any drop in coverage.

Coverage of Cross-loadings. Figure 14 demonstrates Bayesian coverage of cross-loadings. The values shown are the averages across all items. For the first row, the true model is the no general factor model. It results in perfect coverage for cross-loadings. All other models also result in perfect 95% coverage. For the first column, all models with diffuse or weakly informative priors result in perfect coverage for cross-loadings. For the rest of rows and columns, the true model is the full model with diffuse priors. The true model results in perfect coverage for cross-loadings. The underfitting model is the no general factor model, and it also

results in perfect coverage for cross-loadings, because, as we saw before, the unmodeled variance already inflated the specific factors part of the model.

Considering the role of priors, informative inaccurate near-zero priors placed on general factor loadings lead to an up to 60% drop of coverage of cross-loadings, reducing with sample size of 1000. This effect is observed only for a near-zero prior with variance of 0.01.

Summary

Overall, the true model is estimated accurately across most conditions, but the accuracy of estimation depends on sample size and the magnitude of population parameters. For example, bias is influenced by sample size. Smaller samples result in larger bias, however, this effect is mitigated by the choice of the prior. Weakly informative accurate priors and weakly informative near-zero priors (variance 0.1) show best estimation results for general factor loadings with smallest bias and almost complete convergence. The priors placed on one part of the model affect other parts of the model. For example, inaccurate near-zero small variance priors placed on general factor loadings result in inflated cross-loadings and overestimated specific factor loadings. The same is true for model misspecifications. Omitting the cross-loadings results in inflated general factor loadings and underestimated specific factor loadings for corresponding items. In many cases, placing a diffuse prior results in a more accurate estimate than a strongly informative inaccurate prior, especially in small samples.

Bayesian "power" to detect general factor loadings and cross-loadings heavily depends on sample size and the magnitude of estimated parameters. With small parameters, even large sample was not enough to achieve our desired "power", which suggests higher rates of analogous Type II error. Specific factor loadings,

however, are detected with sufficient power regardless of the population conditions and analysis model. Bayesian coverage results suggest that all the parameters have sufficient coverage, providing reliable point estimates, and only strongly informative inaccurate near-zero priors reduce coverage substantially, given small samples.

Simulation Result II: Detecting General Factors using Near-Zero Small Variance Prior

This section focuses specifically on the power to detect the presence of a general factor using near-zero small variance priors. The rationale is that when there is no prior information about whether a general factor should be included, one can apply a near-zero prior with small variance to the general factor loadings. This prior represents a belief that no general factor exists. If, in fact, a general factor is present in the population, the posterior distribution will reflect a compromise between the prior belief (of having no general factor in the population) and the data-driven evidence supporting its presence. We therefore will present the Bayesian "power" for detecting the true general factor loadings.

The computation of Bayesian power was outlined in the previous section "Simulation Result I". For clarity, we reiterate the calculation of "power" here. Essentially, Bayesian "power" is the proportion of replications in which their 95% credible interval excludes zero:

$$\text{power} = \frac{1}{R}\sum_{r=1}^{R} I\Big(0 \notin [L_r, R_r]\Big). \tag{24}$$

It should be noted again that when the true population general factor loading is zero, the power essentially shows the rate of falsely detecting a non-existing general factor loadings (analogous to the frequentist Type I error rate).

Results

We tested the "power" in situations where different priors centered at zero are used. These priors include a diffuse prior with mean zero and infinite variance and two near-zero priors varying in a degree of informativeness. Specifically, one prior had a variance of 0.1, and another – 0.01. Figure 15 shows "power" for detecting general factor loadings. The rows represent population general factor loadings, ranging from 0 to 0.6. The x-axis shows different variances of priors centered at zero, from highly informative (0.01) to completely non-informative (an infinite variance). The y-axis represents the magnitude of power, with a red dashed line indicating a power of 80%.

As shown in the figure, when the population general factor loadings are truly zero, all priors centered at zero yield a false rejection rate of zero, which is analogous to a zero Type I error rate in the frequentist framework. When the population general factor loading is as small as 0.1, the model also yields zero power, suggesting that such a small effect is treated as equivalent to the absence of a general factor.

With a more substantial population loading of 0.3, a sample size of 500 is sufficient to achieve a power higher than 0.80 when a diffuse prior is used for general factor loadings. As the prior variance decreases, larger sample sizes are required to attain comparable power. For instance, with a prior variance 0.1, a sample size of 1000 is needed to get power higher than 0.80. When the prior variance is reduced further to 0.01, an even larger sample size is necessary.

When the general factor loadings increase to 0.4, a sample size of 250 is sufficient to attain a power higher than 0.80 with both diffuse priors and and near-zero prior with a variance 0.1. As the variance of priors shrinks, the prior becomes more informative and a relatively larger sample size is required to overcome the influence of prior. For instance, with a variance 0.01, a sample size 1000 is needed to achieve

a power 0.80 to detect a general factor with medium-sized factor loadings.

As the population general factor loadings increase to 0.6, it is more likely to detect their presence. For the priors with a variance 0.1 and an infinite variance, a sample size as small as 100 is sufficient to detect the general factor loadings with power at least 0.80.

Summary

These results suggest that near-zero priors can be useful in detecting non-zero general factor loadings, and can reject the presence of a general factor when it is absent from the population. The sample size required to detect the general factor loadings varies across the magnitude of the general factor loadings and the variance of the near-zero priors. If a researcher has a large sample of 1000 observations and more, variance 0.01 will lead to detecting considerably large general factor loadings. With smaller samples, we recommend choosing a less informative prior with variance 0.1.

Simulation Result III: Detection of misspecification in bifactor Bayesian SEM

This part presents the results of the performance evaluation of different model fit indices in the situation where the adequacy ("goodness") of model fit has to be evaluated. This is especially important in the context of overfitting and underfitting. An overfitting model is an analysis model that includes parameters that are not present in the population, while an underfitting model is an analysis model that omits the parameters that exist in the population. An example of overfitting would be fitting the bifactor model when there is no general factor in the population. An underfitting situation would be, for example, when cross-loadings exist in the population, but are fixed at 0 in the analysis model.

In Bayesian SEM, there are two types of fit indices that can be used for model fit evaluation. The measures of exact fit compare the observed covariance matrix to the model-implied covariance matrix and get an index of exact similarity, usually based on a χ^2 (chi-square) test (West et al., 2012; L.-t. Hu & Bentler, 1999). Three commonly used measures are Bayesian Posterior Predictive P-value (PPP), Bayesian Information Criterion (BIC) and Deviance Information Criterion (DIC). In contrast, indices of approximate fit compare the fitted model to either the saturated or the baseline model to get the continuous measure of approximate adequacy of the model. Those indices can indicate if the model is "good enough" (R. B. Kline, 2023) and can be classified into two types. First type is non-centrality based indices, for example, Bayesian Root Mean Square Error of Approximation (BRMSEA). It compares the model against the perfectly-fitted (saturated) model. Second type is the incremental fit indices, for instance, Bayesian Comparative Fit Index (BCFI) and Bayesian Tucker-Lewis Index (BTLI). These measures compare the model with the worst-fitting model (the model with zero covariances between observed variables), also called a baseline model.

The following section is organized as follows. First, the computation of the measures of exact fit is outlined and the results of their performance in our simulation study are described and summarized. Second, the computation of approximate fit indices is described and the results of their performance in the simulation study are presented and summarized. Third, the overall summary of the performance of both groups of indices is provided.

Exact Fit Indices

Bayesian Information Criterion

The Bayesian Information Criterion is calculated as follows (Schwarz, 1978):

$$\text{BIC} = (-2)\log(f(\theta|\boldsymbol{y})) + q\log n, \tag{25}$$

where $\log f(\theta|\boldsymbol{y})$ is the log-likelihood of the model parameters θ given the data $\boldsymbol{y}$, and includes a sample size penalty term, where q represents the number of parameters and n is the sample size. BIC can be used to compare several competing models, where the model with the lowest BIC value would be the preferred model. A more nuanced and currently recommended approach is to use a cutoff for the BIC difference between competing models. For SEM models, Raftery (1995) recommends $\Delta BIC > 10$ as a very strong evidence of a better fit for the model with a smaller BIC.

Deviance Information Criterion (DIC)

The computation of DIC involves several steps, the first one is to compute the posterior deviance at each iteration of an MCMC algorithm (Spiegelhalter et al., 2002):

$$\text{D}(\theta) = -2\log(f(\boldsymbol{y}|\theta)) + 2\log(h(\boldsymbol{y})), \tag{26}$$

where $h(\boldsymbol{y})$ is a standardized term which is a function of the data $\boldsymbol{y}$. Next, the mean of the posterior deviances can be computed:

$$\overline{D(\theta)} = E_\theta[-2\log(f(\boldsymbol{y}|\theta))] + 2\log(h(\boldsymbol{y})). \tag{27}$$

After that, the posterior mean $\bar{\theta}$ and its deviance can be computed. The effective number of parameters, pD, is then computed by subtracting the deviance of the posterior mean from the mean of the posterior deviances:

$$p_D = \overline{D(\theta)} - D(\bar{\theta}), \tag{28}$$

Finally, the DIC is computed with the following formula:

$$\text{DIC} = \overline{D(\theta)} + pD = D(\bar{\theta}) + 2pD = 2\overline{D(\theta)} - D(\bar{\theta}). \tag{29}$$

DIC, just like BIC, can be used to compare several competing models, where the model with the smallest DIC would be the preferred model. A more nuanced and currently recommended approach is to use a cutoff for the DIC difference between competing models. For large SEM models, Cain & Zhang (2019) recommend $\Delta DIC > 3$ as sufficient evidence of a better fit for the model with smaller DIC. The use of DIC is especially important for Bayesian estimation, because its performance can be strongly influenced by prior specification, which was shown by Liu et al. (2022).

Posterior Predictive P-value (PPP)

PPP for Bayesian SEM is calculated based on the discrepancy statistic T_{ML}. The SEM discrepancy function T_{ML} is defined as:

$$T_{ML} = (n-1)\left[\log|\Sigma(\theta)| + \text{tr}\{S\Sigma(\theta)^{-1}\} - \log|S| - p\right],$$

where n is the sample size, S is the sample covariance matrix, $\Sigma(\theta)$ is the model-implied covariance matrix, p is the number of observed variables, $|\cdot|$ denotes the determinant, and $\text{tr}\{\cdot\}$ denotes the trace of a matrix.

At each MCMC iteration, s, a simulated dataset Y^s_{rep} of the same size as the observed dataset is generated from the model with parameter values set at θ^s, which are the updated parameters of the model at a given iteration s. We then can calculate the $T_{ML}(Y_{obs}, \theta^s)$ and $T_{ML}(Y^s_{rep}, \theta^s)$. The latter being the discrepancy statistic calculated on the dataset replicated from the estimated parameters, and the former being the discrepancy statistic calculated on the observed dataset, using the updated model parameters. Then the following proportion is calculated:

$$PPP = P[T_{ML}(Y^s_{rep}, \theta^s) > T_{ML}(Y_{obs}, \theta^s)], \tag{30}$$

which is the proportion of the posterior predictive discrepancy statistics that are greater than the discrepancy statistic of the current data. If the model fitted corresponds to the data well, both the fit of the observed data and data replicated from the model parameters should be good, which means one can be better than the other only at random (with a 50% probability). Therefore, a perfect-fitting model is expected to have a PPP value centering around 0.5 and spread evenly above and below, indicating that about 50% of the replicated datasets have discrepancy statistics greater than those of the observed data. A low PPP value near zero suggests potential model misspecification (Asparouhov & Muthén, 2010). Multiple cutoffs, such as 0.01, 0.05, and 0.10 have been proposed (Asparouhov & Muthén, 2010; Gelman et al., 1996), but because PPP is not uniformly distributed, it cannot have a theoretical cutoff to maintain a desired Type I error rate like p-values do in the frequentist framework (Hjort et al., 2006). The 0.05 cutoff is used most often to maintain consistency with traditional p-values (Asparouhov & Muthén, 2010). In this study, we will test both the 0.05 cutoff and the 0.10 cutoff for ruling out the worst-fitting models.

Overall, it can be seen that the BIC and DIC use a similar penalty term based

on sample size and number of estimated parameters, but the penalty of BIC is simpler and stricter than the penalty of DIC. The PPP also accounts for sample size and number of parameters in the computation of the discrepancy function, then directly comparing the discrepancy of observed and replicated data across MCMC iterations. There are several cutoffs suggested in the literature for these indices. Next, we will evaluate the possibility of BIC, DIC and PPP to detect misspecification in a bifactor model with cross-loadings using suggested cutoffs.

Results

In this section, the results of the performance evaluation for BIC, DIC and PPP are presented. For each index, we report the selection rates of three different analysis models, namely the full model, the no general factor model, and the no cross-loadings model. For model comparison using BIC, we use the $\Delta BIC > 10$ cutoff and compare the full model against the no cross-loadings and the no general factor model. For the DIC, we use the same comparison principle, but employ the $\Delta DIC > 3$ cutoff.

BIC. Figure 16 shows selection rates using the cutoff $\Delta BIC > 10$. Columns and rows represent population cross-loadings and population general factor loadings, respectively. The x-axis represents the reference model against which the full model is compared. The y-axis shows the percentage of replications where the full model shows better fit than the reference model, with BIC difference higher than $\Delta BIC = 10$. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample size.

For the first column, the true model would be the no cross-loadings model. We can see that for this column the full model never has better fit than this model. For other columns, as the population cross-loadings increase, the proportion of replications where the full model fits the data better than the no cross-loadings

model increases. For the first row, the true model is the no general factor model, and the full model never fits the data better that the true model. For the rest of rows and columns the true model is a full model, but we see that it fits the data worse than the no general factor model for the majority of population conditions. The only situation where the full model fits the data better is when the population cross-loadings and general factor loadings are large (0.3 and 0.6, respectively), thus moderate ECV (0.50), and sample size is large (N = 1000).

Overall, we can see that BIC is sensitive to the omission of cross-loadings in most of population conditions. In contrast, it is sensitive to the omission of a true population general factor only when general factor loadings, cross-loadings, and sample size are large. In all other scenarios, it favors the no general factor model. This can be attributed to the penalty term. If we take the full model as a reference point, excluding the general factor from the model leads to estimating 9 less parameters (the number of general loadings minus the number of fixed factor covariances), whereas the exclusion of cross-loadings only removes 3 parameters (the number of cross-loadings).

DIC. Figure 17 shows selection rates using the cutoff $\Delta DIC > 3$. Columns and rows represent population cross-loadings and population general factor loadings, respectively. The x-axis represents the reference model against which the full model is compared. The y-axis shows the percentage of replications where the full model shows better fit than the reference model, with DIC difference higher than $\Delta DIC = 3$. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample size.

For the first column, the true model would be the no cross-loadings model. We can see that for this column the full model never has better fit than the no cross-loadings model. For other columns, as the population cross-loadings increase,

the proportion of replications where the full model fits the data better than no cross-loadings model also increases. For the first row, the true model is the no general factor model, but full model shows better fit most of the time. For the rest of rows and columns the true model is a full model, but it fits the data better than the no general factor model only when the population cross-loadings and general factor loadings are larger than 0.2 and 0.4, respectively, indicating moderate ECV, and sample size is large (N $>$ 500).

Overall, we can see that DIC is sensitive to the omission of cross-loadings for most of population conditions. In contrast, it is sensitive to the omission of a true population general factor only when general factor loadings, cross-loadings, and sample size are relatively large, in other words, when ECV suggest the substantive importance of a general factor. This results is similar to the BIC results, but DIC appears more informative. This can be attributed to a more nuanced penalty term than the one of the BIC.

PPP. Evaluating the performance of PPP, we followed the recommendations of using the PPP in a fashion similar to p-value, testing the hypothesis of a good fit of the model. We explored different cutoffs for PPP in this way. Figure 19 shows the proportion of replications for each simulation cell where the PPP is smaller than 0.05, indicating the proportion of replication where the null hypothesis — model having good fit, was rejected. Columns and rows represent population cross-loadings and population general factor loadings, respectively. The x-axis shows models with diffuse priors: full, no cross-loadings and no general factor. For the first column, the true model is the no cross-loadings model, and it never is rejected based on PPP. However, other models also do not get rejected. Moving to next columns, as the size of population cross-loadings increases, the no cross-loadings model gets rejected more often, reaching 100% rejection rate when the population cross-loadings are 0.2

and higher and sample size is large. From the second to the last row, the no general factor model is a misspecified model, but it gets rejected only when general factor loadings are large (0.6) and population cross-loadings are large, indicating substantive importance of a general factor. For all rows and columns, except for the first row and column, the true model is the full model, and it never gets rejected.

Overall, it can be seen that the PPP cutoff 0.05 successfully identifies the misspecification in the cross-loadings part, when population cross-loadings are moderate and large. Misspecification in the general factor part leads to model rejection only when general factor loadings are large and sample size is large (N > 500), in other words, when general factor explains an amount of variance comparative to specific factors. For any scenario with cross-loadings less than 0.2 and any general factor loadings, this cutoff is not informative. Figure 20 shows the rates of misfit using the 0.10 cutoff for the PPP. We can see that the results are essentially the same as using the 0.05 cutoff.

Although this practice is not common, we decided to evaluate the closeness of the PPP for each model to 0.50 and the smallest deviance from 0.50 as the indicator of the best fit among competing models. Figure 18 shows the results of this procedure. Columns and rows represent population cross-loadings and population general factor loadings, respectively. The x-axis represents the fitted models (full, no cross-loadings, and no general factor). The data points show the proportion of replications with the PPP value closest to .50 among the three models. The line type and line color represent different sample sizes.

For the first column, the true model is the no cross-loadings model, but it shows worse or similar fit to other models. For the first row, the true model is the no general factor model, and it shows better fit to other models, given that sample size is 250 and larger. For the rest of rows and columns, the true model is the full

model. It shows consistently better fit to the data than other models, and the precision of this effect increases with increasing population general factor loadings, cross-loadings and sample size.

Summary

The indices of exact fit show mixed results in terms of model selection. When using cutoffs suggested in the literature, BIC, DIC and PPP are all sensitive to model misspecification in the cross-loadings part, especially when large cross-loadings are omitted. At the same time, the omission of a general factor is only detected when ECV is large and suggests the substantive importance of a general factor. This suggests that researchers should rely on these indices in making decisions about retaining the general factor only after evaluating the loadings' estimates and calculating ECV. The use of the PPP proximity to the 0.5 as a criterion for model selection unexpectedly leads to correct model acceptance and rejection, but with large sample size (ideally, 500 and more). However, we could not find any examples of using PPP in this way in the literature.

Bayesian Approximate Fit Indices for SEM

Bayesian approximate fit indices have been developed to mimic the frequentist indices, using the Bayesian discrepancy function as an analogue for the chi-square statistic, and the estimated number of parameters to compute the analogue for degrees of freedom. These indices can use discrepancy functions differently, and in this study, the version suggested by Garnier-Villarreal & Jorgensen (2020) is used. Hu and Bentler (L. Hu & Bentler, 1998) suggested the following cutoffs for frequentist fit indices to detect reasonably well-fitting models: RMSEA <0.06, TLI >0.95,CFI >0.95, the same cutoffs proved to also be adequate for the Bayesian versions of these indices (Garnier-Villarreal & Jorgensen, 2020).

Bayesian RMSEA

The Bayesian RMSEA is calculated as follows:

$$\text{BRMSEA}_s = \sqrt{\max\left(0, \frac{T_{ML}(Y_{obs}, \theta^s) - p^*}{(p^* - pD)n}\right)}, \tag{31}$$

where $T_{ML}(Y_{obs}, \theta^s)$ is the discrepancy function for the observed data at iteration s, n is a sample size, p^* is the number of parameters in the H_1 model, and pD is the estimated number of parameters in the H_0 model. H_1 model refers to the fully unrestricted model, which is highly complex and allows for all possible relationships between all variables. Basically, it is the observed sample variance-covariance matrix. H_0 model is the model that we are trying to fit, with the number of parameters pD. Thus, the denominator $p^* - pD$ is just the analogue of frequentist degrees of freedom computed by subtracting the number of estimated parameters q from the number of non-redundant sample moments p^*. Misfit captured by the BRMSEA is the discrepancy at iteration s, rescaled by the number of observed sample moments (Garnier-Villarreal & Jorgensen, 2020). A small value of BRMSEA implies good model fit.

Bayesian CFI

The Bayesian CFI is computed as follows:

$$\text{BCFI}_s = 1 - \frac{T^T_{ML}(Y_{obs}, \theta^s) - p^*}{T^B_{ML}(Y_{obs}, \theta^s) - p^*}, \tag{32}$$

where $T^T_{ML}(Y_{obs}, \theta^s)$ is the discrepancy function (for the observed data) of the fitted model and $T^B_{ML}(Y_{obs}, \theta^s)$ is the discrepancy of the baseline model, both at iteration s. Generally, the baseline model is a model in which all variances are estimated but all covariances are fixed at zero (R. B. Kline, 2023). The larger values of BCFI

indicate better fit.

Bayesian TLI

The Bayesian TLI is computed as follows:

$$\text{BTLI}_s = \frac{\frac{T^B_{ML}(Y_{obs}, \theta^s) - pD_B}{p^* - pD_B} - \frac{T^T_{ML}(Y_{obs}, \theta^s) - pD_T}{p^* - pD_T}}{\frac{T^B_{ML}(Y_{obs}, \theta^s) - pD_B}{p^* - pD_B} - 1}, \tag{33}$$

where pD_B is the model complexity term for the baseline model, and pD_T is the model complexity term for the target model under evaluation. The larger values of BTLI indicate better fit.

Overall, the discrepancy function is evaluated and the approximate fit indices are calculated using the parameters at each iteration of the MCMC, which indicates that they can be influenced by the prior distribution. Moreover, the inclusion of informative priors changes the estimated number of parameters (Garnier-Villarreal & Jorgensen, 2020), which is used in the computation of the indices. Therefore, it is essential to investigate how different prior specifications impact these fit indices.

Results

In this section, the results of the performance evaluation of BRMSEA, BCFI and BTLI in detecting model misspecification are presented. We show the selection rates for each model with diffuse priors (the full model, the no general factor model, the no cross-loadings model), evaluate the Bayesian credibility intervals of the indices and report the prior sensitivity analysis.

BRMSEA. Figure 21 shows the proportion of replications for which the computed BRMSEA falls into the acceptable range (below 0.06). Columns and rows are population cross-loadings and general factor loadings, respectively. The x-axis

represents the three fitted models with diffuse priors (full, no cross-loadings, and no general factor). Line type and line color represent sample size.

For the first column, the true model is the no cross-loadings model, and it shows good fit 100% of the time. However, the other two models show the same result, despite being misspecified. For the first row, the true model is the no general factor model. It shows good fit 100% of the time. However, an overfitting full model also shows good fit 100% of the time. For the rest of rows and columns, the true model is the full model, and it shows good fit 100% of the time. The underfitting no general factor model also shows good fit 100% of the time. The no-cross loadings model is an underfitting model, as it omits the non-zero population cross-loadings. When those cross-loadings are large (0.3), the no cross-loadings model shows smaller selection rates, reaching 0% when general factor loadings and sample size are large.

Overall, when sample size and population parameters are small, BRMSEA accepts all models, even misspecified. Concerning misspecifications, BRMSEA seems to be sensitive to cross-loadings and not sensitive to a general factor. It correctly flags the no cross-loadings model as poorly fitting, when population cross-loadings are large. This effect increases with increasing population general factor loadings.

Instead of comparing the BRMSEA point estimate to an established cutoff (0.06 for BRMSEA), in Bayesian analysis, we can assess the whole distribution of BRMSEA based on credible intervals. If the established cutoff is within a 90% credible interval, then the model fit is inconclusive. If the entire credible interval is below 0.6, the fit can be classified as conclusively good. Figure 22 presents the results of this fit evaluation procedure. Stacked bar charts depict the proportion of replications classified as “good”, “inconclusive”, or “poor” model fit based on the 90% credible intervals of BRMSEA values. The second-order columns are sample sizes, the first-order columns are population cross-loadings, and rows are population

general factor loadings. The y-axis shows the percentage of replications. The x-axis shows the fitted models: the full model, the no general factor model, the no cross-loadings model. We can see that for sample size 100, from 60 to 80% of replications for all three models have inconclusive fit, suggesting that BRMSEA is not informative for this sample size. For larger samples, all three models are flagged as good fit across the majority of replications. Only when sample size is 1000, population cross-loadings are 0.3 and population general factor loadings are 0.4, the no cross-loadings model is flagged as a conclusively bad–fitting model for about 80% of the replications. Therefore, using credible intervals has similar results to using the BRMSEA point estimate.

Figure 23 shows the results of prior sensitivity testing for the BRMSEA. On the x-axis the prior settings are shown as following: $\Lambda_G, \Lambda_C \sim \mathcal{N}(0, \infty)$; $\Lambda_C \sim \mathcal{N}(0, \infty)$, $\Lambda_G \sim \mathcal{N}(0, 0.1)$; $\Lambda_C \sim \mathcal{N}(0, \infty)$, $\Lambda_G \sim \mathcal{N}(0, 0.01)$; and $\Lambda_G \sim \mathcal{N}(true, 0.5true)$, $\Lambda_C \sim \mathcal{N}(0, \infty)$. Rows and columns represent population general factor loadings and population cross-loadings, respectively. The x-axis represents prior variances for the prior placed on general factor loadings. The first three priors are centered at zero, and the last prior is centered at the true value. From the results we can observe that near-zero priors placed on general factor loadings affect the ability of BRMSEA to detect a well-fitting model and this effect is mitigated by sample size and population parameters. Specifically, when general factor loadings are 0.3 and less in the population, both models with near-zero priors fit the data well, especially the small variance prior $\sim \mathcal{N}(0, 0.1)$. These population loadings are small and the prior does not cause huge discrepancy with the data. As the size of general factor loadings in the population increases, reaching 0.4 and 0.6, the near-zero prior $\Lambda_G \sim \mathcal{N}(0, 0.01)$ becomes inaccurate, which leads the BRMSEA to flag this model as poorly fitting. This effect is mitigated by sample size. However, the model with

$\Lambda_G \sim \mathcal{N}(0, 0.1)$ prior still fits the data well in those conditions.

BCFI. Figure 24 shows the proportion of replications for which the computed BCFI falls within the acceptable range (above 0.95). Columns and rows are population cross-loadings and population general factor loadings, respectively. The x-axis represents the three fitted models (full, no cross-loadings, and no general factor). Line type and line color represent different sample sizes.

For the first column, the true model is the no cross-loadings model, and it shows good fit 100% of the time. However, the other two models show the same result, despite being misspecified. For the first row, the true model is the no general factor model. It shows good fit 100% of the time. However, an overfitting full model also shows good fit 100% of the time. For the rest of rows and columns, the true model is the full model, and it shows good fit 100% of the time. The underfitting no general factor model also shows good fit 100% of the time. The no-cross loadings model is also an underfitting model, as it omits the non-zero population cross-loadings. When those cross-loadings are large (0.3), the no cross-loadings model shows smaller selection rates, reaching 0% when general factor loadings are small and sample size is large.

Therefore, when sample size and population parameters are small, BCFI accepts all models, even misspecified. Concerning misspecifications, BCFI seems to be sensitive to cross-loadings and not sensitive to a general factor. It correctly flags the no cross-loadings model as poorly fitting, when population cross-loadings are large and general factor loadings are small.

Instead of comparing a BCFI point estimate to a cutoff (0.95 for BCFI), in Bayesian analysis, we can assess the entire distribution of BCFI directly, based on credible intervals. If the cutoff is within a 90% credible interval, then the fit is inconclusive. If the entire credible interval is above 0.95, the fit can be classified as

conclusively good. Figure 25 presents stacked bar charts that depict the proportion of replications classified as "good", "inconclusive", or "poor" model fit based on the 90% credible intervals of BCFI values. The second-order columns are sample sizes, the first-order columns are population cross-loadings, and rows are population general factor loadings. The y-axis shows the percentage of replications. The x-axis shows the fitted models (full, no general factor, no cross-loadings). We can see that for sample size 100, from 60 to 100% of replications for all three models have inconclusive fit, suggesting that BCFI is not informative for this sample size. For larger samples, all three models are flagged as good fit across the majority of replications. Only when sample size is 1000, population cross-loadings are 0.3 and population general factor loadings are 0 and 0.1, the no cross-loadings model is flagged as a conclusively bad–fitting model for about 60% of the replications. Therefore, using credible intervals has similar results to using the BCFI point estimate.

Figure 26 shows the results of prior sensitivity testing for the BCFI. On the x-axis the prior settings are shown as following: $\Lambda_G\Lambda_C \sim \mathcal{N}(0, \infty)$; $\Lambda_C \sim \mathcal{N}(0, \infty)$, $\Lambda_G \sim \mathcal{N}(0, 0.1)$; $\Lambda_C \sim \mathcal{N}(0, \infty)$, $\Lambda_G \sim \mathcal{N}(0, 0.01)$; and $\Lambda_G \sim \mathcal{N}(true, 0.5true)$, $\Lambda_C \sim \mathcal{N}(0, \infty)$. Rows and columns represent population general factor loadings and population cross-loadings, respectively. The x-axis represents different variances of a prior placed on general factor loadings. The first three priors are centered at zero and the last prior is centered at a true value. From the results we can observe that when general factor loadings are 0.3 and less in the population, both near-zero priors specified for these loadings are accurate and capture the population values well, especially the $\sim \mathcal{N}(0, 0.1)$ prior, but only for sample sizes larger than 100. As the size of general factor loadings increases, reaching 0.4 and 0.6, the strongly informative near-zero prior $\Lambda_G \sim \mathcal{N}(0, 0.01)$ becomes inaccurate, which leads the

BCFI to flag this model as poorly fitting. This effect, however, is mitigated by sample size. Sample sizes of 500 and 1000 observations place more weight to the data, than to the prior, which leads BCFI to indicating good fit. In contrast, the less informative near-zero prior does not cause lack of fit.

BTLI. Figure 27 shows the proportion of replications for which the computed BTLI falls into the acceptable range (above 0.95). Columns and rows are population cross-loadings and general factor loadings, respectively. The x-axis represents the three fitted models with diffuse priors (full, no cross-loadings, and no general factor). Line type and line color represent sample size.

For the first column, the true model is the no cross-loadings model, and it shows good fit 100% of the time. However, the other two models show the same result, despite being misspecified. For the first row, the true model is the no general factor model. It shows good fit 100% of the time. However, an overfitting full model also shows good fit 100% of the time. For the rest of rows and columns, the true model is the full model, and it shows good fit 100% of the time. The underfitting no general factor model also shows good fit 100% of the time. The no-cross loadings model is an underfitting model, as it omits the non-zero population cross-loadings. When those cross-loadings are large (0.3), the no cross-loadings model shows smaller selection rates, reaching 0% when general factor loadings and sample size are large.

Overall, when sample size and population parameters are small, BTLI accepts all models, even misspecified. Concerning misspecifications, BTLI seems to be sensitive to cross-loadings and not sensitive to a general factor. It correctly flags the no cross-loadings model as poorly fitting, when population cross-loadings are large.

Instead of comparing the BTLI point estimate to an established cutoff (0.06 for BTLI), in Bayesian analysis, we can assess the whole distribution of BTLI based on credible intervals. If the established cutoff is within a 90% credible interval, then

the model fit is inconclusive. If the entire credible interval is above 0.95, the fit can be classified as good. Figure 28 presents the results of this fit evaluation procedure. Stacked bar charts depict the proportion of replications classified as "good", "inconclusive", or "poor" model fit based on the 90% credible intervals of BTLI values. The second-order columns are sample sizes, the first-order columns are population cross-loadings, and rows are population general factor loadings. The y-axis shows the percentage of replications. The x-axis shows the fitted models: the full model, the no general factor model, the no cross-loadings model. We can see that for sample size 100, from 60 to 80% of replications for all three models have inconclusive fit, suggesting that BTLI is not informative for this sample size. For larger samples, all three models are flagged as good fit across the majority of replications. Only when sample size is 500 and larger, population cross-loadings are 0.3 and population general factor loadings are 0.3 and larger, the no cross-loadings model is flagged as a conclusively bad–fitting model for 70 to 90% of the replications. Therefore, using credible intervals has similar results to using the BTLI point estimate.

Figure 29 shows the results of prior sensitivity testing for the BTLI. On the x-axis the prior settings are shown as following: $\Lambda_G\Lambda_C \sim \mathcal{N}(0, \infty)$; $\Lambda_C \sim \mathcal{N}(0, \infty)$, $\Lambda_G \sim \mathcal{N}(0, 0.1)$; $\Lambda_C \sim \mathcal{N}(0, \infty)$, $\Lambda_G \sim \mathcal{N}(0, 0.01)$; and $\Lambda_G \sim \mathcal{N}(true, 0.5true)$, $\Lambda_C \sim \mathcal{N}(0, \infty)$. Rows and columns represent population general factor loadings and population cross-loadings, respectively. From the results we can observe that near-zero priors placed on general factor loadings affect the ability of BTLI to detect a well-fitting model and this effect is mitigated by sample size and the population parameters. Specifically, when population general factor loadings are 0.3 and less, both near-zero priors specified for these loadings are accurate and capture the population values well, especially the $\sim \mathcal{N}(0, 0.1)$ prior, but only for sample sizes

larger than 100. For sample size 100 the selection rate is smaller regardless of the prior. As the size of population general factor loadings increases, reaching 0.4 and 0.6, the near-zero prior $\Lambda_G \sim \mathcal{N}(0, 0.01)$ becomes strongly informative inaccurate, which leads the BTLI to flag this model as poorly fitting. This effect, however, is mitigated by sample size. Similar effect is observed for the $\Lambda_G \sim \mathcal{N}(0, 0.1)$ prior, but sample size larger than 250 is already sufficient to ignore the prior inaccuracy.

Summary

In our study, approximate fit indices are less sensitive to misspecification than BIC, DIC and PPP, when diffuse priors are used on all loadings in the model. Specifically, these indices fail to detect underfit in the general factor part, even when the ECV indicates that general factor is important and needs to be modeled. It seems like the presence of cross-loadings is enough to regard the model as well–fitting, even when the population general factor is present. However, the omission of cross-loadings is almost always detected and results in lower selection rates for no the cross-loadings model, compared to other models. The use of near-zero priors on general factor loadings leads to more precise results, as when near-zero prior is accurate, the full model is selected correctly, but when it is inaccurate, the full model is rejected with a 100% rate, unless sample size is sufficient enough to ignore the prior.

To understand the pattern of selection rates when the full model and the no general factor model results in the same selection rate, we can take a look at the estimated number of parameters pD. This parameter is used to compute the Bayesian analogue of degrees of freedom in the computation of Bayesian indices of approximate fit, which serves as a penalty term for model complexity. The distribution of that parameter across simulation conditions is presented in Figure 30. We can see that consistently across multiple conditions, especially conditions

having a large ECV, such as population general loadings 0.6, the no general factor model has the smallest number of parameters, which makes it less penalized than the other two models, resulting in good fit. In contrast, we know that the full model includes a general factor, which makes it less parsimonious and more penalized, but at the same time, inherently, such models fit data better, balancing the penalty term. Potentially, this can lead to similar results of fit evaluation for the full model and the no general factor model. The full model loses fit due to less parsimony, but acquires it due to a general factor, and the no general factor model loses fit due to lack of general factor, but gains it due to parsimony. In contrast, the no cross-loadings might result in poor fit when the population cross-loadings are large, because it loses fit due to higher discrepancy, while the number of estimated parameters is not substantially different from the full model, not providing enough parsimony to compensate for elevated discrepancy.

Discussion and Future Directions

In this study, we evaluated the performance of Bayesian approach in estimating a bifactor model with cross-loadings. The simulation design considered a bifactor structure with 12 indicators, three specific factors, as well as three cross-loadings. Both the general factor and the cross-loadings were allowed to be either present or absent and could vary in magnitude. We systematically manipulated the magnitude of general factor loadings and cross-loadings, including conditions fixing their population values at zero, while holding specific factor loadings constant. These combinations were designed to reflect scenarios in which the general factor was either non-meaningful or equally meaningful as the specific (group) factors. In addition, we examined the influence of different priors for general factor loadings, varying their degrees of informativeness and accuracy — including highly informative near-zero small variance priors. A unique contribution of this study is

assessing the utility of near-zero small variance priors as a tool in detecting the presence of a general factor. This directly addresses a broader psychological issue — the common but problematic practice of using the same data for generating and testing structural hypotheses, often by fitting a bifactor model only after a traditional common factor model fails to achieve acceptable fit. We propose that testing the presence of a general factor and estimating other model parameters, including several meaningful cross-loadings, can be done in a single modeling step by using near-zero small variance priors.

Regarding simulation study results, we first evaluated the accuracy of parameter estimation by assessing the bias of estimation accompanied with mean squared error, Bayesian analogue of frequentist power and the Bayesian coverage. These measures gave us insight on how accurately and stably the parameters are estimated, what can be done to achieve high statistical "power" and how reliable the estimates are. In the second portion of results, we focused specifically on testing the utility of near-zero priors in testing the presence of a population general factor using Bayesian analogue of power. In the third portion of the results, we looked at the misspecification sensitivity of recently developed Bayesian versions of indices of approximate fit for SEM, namely, BRMSEA, BTLI and BCFI. We also evaluated the performance of classic indices of exact fit (BIC, DIC and PPP) in misspecification detection, using the cutoffs suggested in the literature.

The results assessing estimation accuracy suggested that the true model was estimated accurately across most conditions. Even when population factor loadings were small, relative bias was within the acceptable range, accompanied with small mean squared error. In turn, model misspecifications in one part of the model produced biased estimates in other parts of the model. First, omitting large population cross-loadings inflated general factor loadings and at the same time led

to underestimation of specific factor loadings for the corresponding observed variables. These findings replicate previous studies (Ximénez et al., 2022; Murray & Johnson, 2013; Morin et al., 2016). Second, omitting large general factor loadings resulted in inflated specific factor loadings for all observed variables, while specifying an inaccurate near-zero prior led to the inflation of both specific factor loadings and cross-loadings. The power of detecting general factor loadings and cross-loadings was highly related to sample size and population parameter magnitude, with increasing sample size and having larger population parameters altogether leading to higher power. Importantly, we recommend aiming for at least 500 observations for a model similar to the model we estimated, as in many cases, smaller samples were not enough even to detect a moderate to large population loading. Finally, good coverage was achieved for the true model across most conditions, only fluctuating with model misspecifications and inaccurate priors. These results suggest that even using diffuse priors, we could accurately recover the population cross-loadings and general factor loadings, as well as specific factor loadings, using a confirmatory bifactor model estimated in Bayesian framework. We note that the inclusion of meaningful cross-loadings was crucial for achieving convergence, therefore, we recommend including them in the model, given substantive theoretical or empirical justification.

The second portion of results was devoted specifically to testing the ability of near-zero small variance priors to detect the presence of a population general factor. We used the Bayesian analogue of frequentist power to analyze these results. As expected, near-zero priors worked as informative accurate priors when population loadings were zero or small, resulting in accurate estimates. With increasing population loadings, the discrepancy between the data and the prior also increased, resulting in distorted estimates and large errors, as well as diminished “power” and

coverage. However, these effects were ignorable for the prior with variance 0.1 for most of sample sizes, and for prior with variance 0.01 with sample size 1000. Those findings indicate the importance of conducting prior sensitivity analysis as a robustness check for the results (Depaoli et al., 2020). Highly informative priors can distort the estimation and produce misleading results, which will remain undetected if only one set of prior specifications is used. If several sets of priors with different hyperparameters are used, we can directly see how the estimates change, and if the change is large, place more weight in our conclusions to less informative priors.

Concerning the third portion of the results, we discovered multiple important patterns. First, indices of approximate fit were non-informative for sample size 100, indicating inconclusive or good fit for virtually all analysis models based on both credible intervals and point estimates of fit indices. A similar non-informativeness was observed irrespective of sample size for combinations of small population parameters, such as cross-loadings 0.1 and general factor loadings 0.1, and was somewhat expected due to little substantive importance of such small loadings and corresponding factors. Second, both indices of exact and Bayesian approximate fit were sensitive to the omission of cross-loadings. Additionally, indices of exact fit were somewhat sensitive to the omission of a general factor, when it is large, cross-loadings are large and sample size is large, reflecting the substantive importance of a general factor. In contrast, Bayesian indices of approximate fit were largely insensitive to the omission of a general factor, resulting in almost equally good fit for a model with and without a general factor, even when a general factor was present in the population and ECV was approaching 50%, suggesting that general factor is as important as specific factors. We attribute this phenomenon to the way these indices are computed. They use the estimated number of parameters of the fitted model and the discrepancy function. Small discrepancy and small

number of parameters lead to better fit. We argue that a full model has a high selection rate across most conditions, because naturally, a model with a general factor fits any data better than any other model, which was noted in many studies (Reise et al., 2016; Bonifay & Cai, 2017; Greene et al., 2019; Watts et al., 2020). However, it loses some fit due to an inflated penalty term. At the same time, the no general factor model has a much smaller number of parameters, which improves fit through the smaller model complexity term, but loses fit in the discrepancy function part. Therefore, both of these models gain and lose fit in different sources, potentially arriving at similar results. In comparison, the no cross-loadings model is not much different from the full model in the number of parameters, but loses fit due to a bigger discrepancy, because large cross-loadings are omitted. Therefore, none of the places where this model could gain fit contribute to it, compared to the full and the no general factor model. However, testing these arguments systematically is beyond the scope of the current study and is subject to future research.

Future research can also address more complex population models, where loadings, at least for specific factors, are not restricted to have the same magnitude, and some specific factors can be slightly weaker than the others. Another potential limitation of this study is that we did not explore more combinations of population parameters, such as larger general factor loadings and, thus, larger ECV. To reiterate, our maximum ECV was 50%, indicating equal importance of both general and specific factors. Larger ECV values suggest that general factor is more important, potentially suggesting unidimensionality and, thus, redundancy of specific factors (Rodriguez et al., 2016). It would be interesting to look at the performance of fit indices, especially indices of approximate fit, in these conditions and compare them with models with moderate and small ECV. Finally, we only allowed for several meaningful cross-loadings, while still keeping specific factors

uncorrelated, under the assumption that potential factor correlations stem from unmodeled cross-loadings, and accounting for substantial cross-loadings would eliminate the need to model specific factor covariances in a bifactor model. Further research using Bayesian estimation may look into estimating factor covariances alongside with cross-loadings.

Based on the results, we developed several recommendations for applied researchers. *First*, we recommend applied researchers to fit bifactor models only when there is a theoretical justification for the presence of a general factor. *Second*, when the decision is to fit a bifactor model, we recommend employing Bayesian estimation to model the general factor using near-zero priors, even when there is strong theory justifying the general factor. *Third*, when fitting a Bayesian bifactor model with near-zero priors, we recommend recruiting a sample of at least 500 participants, if the model is similar to our full model in terms of complexity. For the near-zero priors, we recommend choosing variances based on available sample size and testing different variances (different degree of informativeness). If a researcher is able to recruit a sample size of 1000 and higher, the variance can be as small as 0.01, but a larger variance of 0.1 can also be tested as a robustness check. With smaller samples, variance of 0.1 is more preferred to avoid falsely rejecting the presence of a general factor and biasing the estimates of cross-loadings and specific factor loadings. Overall, we do not recommend estimating similar models on sample sizes smaller than 500. *Fourth*, along with testing the presence of a general factor, we suggest including a couple of meaningful cross-loadings, while fixing other cross-loadings at zero. Smaller cross-loadings will get absorbed in the general factor, if they are present, while meaningful cross-loadings will be estimated and can be interpreted, enriching the theory about the construct. *Fifth*, we recommend evaluating model fit by first computing the Explained Common Variance (ECV)

based on the estimates for all factor loadings in a model. The formula for ECV is provided in the Simulation Design section of the current paper. If the ECV is relatively large, with values close to 0.5 and higher, then the model fit can be evaluated by using the indices of exact fit, as they will capture the misspecification in the general loadings part if the researcher compares a Bayesian bifactor model with a correlated factor model. Currently, based on the performance of the indices of approximate fit, we cannot recommend them for evaluating fit of a model similar to our model. In line with our findings and previous research (Forbes et al., 2021; Watts et al., 2019; Bornovalova et al., 2020; Murray & Johnson, 2013; Rodriguez et al., 2016), we recommend evaluating the parameter estimates, such as the size of factor loadings and standard errors, as well as uniformity of loadings within a general factor, instead of relying on model fit indices, when estimating a bifactor model with cross-loadings in the Bayesian framework.

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: a Multidisciplinary Journal*, *16*(3), 397–438.

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis using mplus: Technical implementation.*

Asparouhov, T., & Muthén, B. (2021). Advances in bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 1–14. Retrieved from https://doi.org/10.1080/10705511.2020.1764360

Asparouhov, T., Muthén, B., & Morin, A. J. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. *Journal of Management*, *41*(6), 1561–1577. Retrieved from https://doi.org/10.1177/0149206315591075

Bader, M., Jobst, L. J., & Moshagen, M. (2022). Sample size requirements for bifactor models. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(5), 772–783. Retrieved from https://doi.org/10.1080/10705511.2021.2019587

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

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three concerns with applying a bifactor model as a structure of psychopathology. *Clinical Psychological Science*, *5*(1), 184–186. Retrieved from https://doi.org/10.1177/21677026166570

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in lisrel maximum likelihood estimation. *Psychometrika*, *50*(2), 229–242. Retrieved from https://doi.org/10.1080/0267257X.1994.9964263

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. doi: 10.1016/biopsych.2020.01.013

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203. doi: 10.1037/0033-295X.110.2.203

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*, 791-799. Retrieved from https://api.semanticscholar.org/CorpusID:3906689

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. Retrieved from https://doi.org/10.1080/10618600.1998.10474787

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Cain, M. K., & Zhang, Z. (2019). Fit for a Bayesian: An evaluation of PPP and DIC for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 39–50. doi: 10.1080/10705511.2018.1490648

Carnovale, M., Taylor, G. J., Parker, J. D., Sanches, M., & Bagby, R. M. (2021). A bifactor analysis of the 20-item toronto alexithymia scale: Further support for a

general alexithymia factor. *Psychological Assessment*, *33*(7), 619. doi: 10.1037/pas0001000

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

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *158*(3), 419–444. Retrieved from https://doi.org/10.2307/2983440

Clayton, S. (2003). Environmental identity: A conceptual and an operational definition. *Identity and the Natural Environment: The Psychological Significance of Nature*, 45–65. Retrieved from https://doi.org/10.7551/mitpress/3644.001.0001

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *The American Journal of Medicine*, *119*(2), 166–e7. doi: 10.1016/j.amjmed.2005.10.036

Crede, M., & Harms, P. (2019). Questionable research practices when using confirmatory factor analysis. *Journal of Managerial Psychology*, *34*(1), 18–30. Retrieved from https://doi.org/10.1108/JMP-06-2018-0272

Cudeck, R., & O'dell, L. L. (1994). Applications of standard error estimates in unrestricted factor analysis: significance tests for factor loadings and correlations. *Psychological Bulletin*, *115*(3), 475. doi: 10.1037/0033-2909.115.3.475

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, *18*(2), 186. doi: 10.1037/a0031609

Depaoli, S. (2021). *Bayesian structural equation modeling*. Guilford Publications.

Depaoli, S., & Van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The wambs-checklist. *Psychological Methods*, *22*(2), 240. doi: 10.1037/met0000065

Depaoli, S., Winter, S. D., & Visser, M. (2020). The importance of prior sensitivity analysis in Bayesian statistics: demonstrations using an interactive Shiny App. *Frontiers in Psychology*, *11*, 608045.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155. doi: 10.1037/1082-989x.5.2.155

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

Eid, M., Krumm, S., Koch, T., & Schulze, J. (2018). Bifactor models for predicting criteria by general and specific factors: Problems of nonidentifiability and alternative solutions. *Journal of Intelligence*, *6*(3), 42. doi: 10.3390/jintelligence6030042

Fang, G., Guo, J., Xu, X., Ying, Z., & Zhang, S. (2021). Identifiability of bifactor models. *Statistica Sinica*, *31*, 2309–2330. Retrieved from `https://doi.org/10.5705/ss.202020.0386`

Fife, D. A., & Rodgers, J. L. (2022). Understanding the exploratory/confirmatory data analysis continuum: Moving beyond the "replication crisis". *American Psychologist*, *77*(3), 453. Retrieved from `https://doi.org/10.1037/amp0000886`

Forbes, M. K., Greene, A. L., Levin-Aspenson, H. F., Watts, A. L., Hallquist, M., Lahey, B. B., ... others (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. doi: 10.1037/abn0000533

Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to Maximum Likelihood. *Psychological Methods*, *25*(1), 46. doi: 10.1037/met0000224

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Texts in statistical science: Bayesian data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 733–760.

Ghosh, J. K., Delampady, M., & Samanta, T. (2007). *An introduction to Bayesian analysis: theory and methods*. Springer Science & Business Media.

Green, S., & Yang, Y. (2018). Empirical underidentification with the bifactor model: A case study. *Educational and Psychological Measurement*, *78*(5), 717–736. doi: 10.1177/0013164417719947

Greene, A. L., Eaton, N. R., Li, K., Forbes, M. K., Krueger, R. F., Markon, K. E., ... others (2019). Are fit indices used to test psychopathology structure biased? A simulation study. *Journal of Abnormal Psychology*, *128*(7), 740. doi: 10.1037/abn0000434

Griffin, D. W., & Bartholomew, K. (1994). Models of the self and other: Fundamental dimensions underlying measures of adult attachment. *Journal of*

Personality and Social Psychology, *67*(3), 430. Retrieved from https://doi.org/10.1037/0022-3514.67.3.430

Gu, H., Wen, Z., & Hau, K.-T. (2023). Bifactor exploratory structural equation models versus traditional approaches in predicting external criteria. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(5), 778–788. Retrieved from https://doi.org/10.1080/10705511.2023.2165496

Guo, J., Marsh, H. W., Parker, P. D., Dicke, T., Lüdtke, O., & Diallo, T. M. (2019). A systematic evaluation and comparison between Exploratory structural equation modeling and Bayesian structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 529–556. Retrieved from https://doi.org/10.1080/10705511.2018.1554999

Harrington, D. (2009). *Confirmatory factor analysis*. Oxford university press.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: a cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319. doi: 10.1037/a0024917

Hjort, N. L., Dahl, F. A., & Steinbakk, G. H. (2006). Post-processing posterior predictive p values. *Journal of the American Statistical Association*, *101*(475), 1157–1174. Retrieved from https://doi.org/10.1198/016214505000001393

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. doi: /10.1037/1082-989X.3.4.424

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

Humphreys, L. G. (1979). *The construct of general intelligence* (Vol. 3) (No. 2). Elsevier. Retrieved from https://doi.org/10.1016/0160-2896(79)90009-6

Jöreskog, K. G., & van Thiilo, M. (1972). LISREL: A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables. *ETS Research Bulletin Series*, *1972*(2), i-71. doi: https://doi.org/10.1002/j.2333-8504.1972.tb00827.x

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. Retrieved from https://doi.org/10.1111/jedm.12000

Kline, P. (2014). *An easy guide to factor analysis*. Routledge.

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.

Lai, K., & Green, S. B. (2016). The problem with having two watches: Assessment of fit when RMSEA and CFI disagree. *Multivariate Behavioral Research*, *51*(2-3), 220–239. Retrieved from https://doi.org/10.1080/00273171.2015.1134306

Lane, S., Raymond, M. R., Haladyna, T. M., et al. (2016). *Handbook of test development* (Vol. 2). Routledge New York, NY.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (Vol. 18). Butterworths London.

Lemoine, N. P. (2019). Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, *128*(7), 912–928. doi: https://doi.org/10.1111/oik.05985

Liu, H., Depaoli, S., & and, L. M. (2022). Understanding the Deviance Information Criterion for SEM: Cautions in prior specification. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(2), 278–294. doi: 10.1080/10705511.2021.1994407

Liu, H., Zhang, Z., & and, K. J. G. (2016). Comparison of Inverse Wishart and Separation-Strategy priors for Bayesian estimation of covariance parameter matrix in Growth Curve Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 354–367. doi: 10.1080/10705511.2015.1057285

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

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. Retrieved from `https://doi.org/10.1037/0033-2909.109.3.502`

Markon, K. E. (2019). Bifactor and hierarchical models: Specification, inference, and interpretation. *Annual Review of Clinical Psychology*, *15*(1), 51–69. Retrieved from `https://doi.org/10.1146/annurev-clinpsy-050718095522`

Marsh, H. W., & Hau, K.-T. (1999). Confirmatory factor analysis: Strategies for small sample sizes. *Statistical Strategies for Small Sample Research*, *1*, 251–284.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741. doi: https://doi.org/10.1037/0003-066X.50.9.741

Moreira, P., Loureiro, A., Inman, R., & Olivos-Jara, P. (2021). Assessing the unidimensionality of Clayton's Environmental Identity scale using confirmatory factor analysis (CFA) and bifactor exploratory structural equation modeling (bifactor ESEM). *Collabra: Psychology*, *7*(1), 28103. doi: 10.1525/collabra.28103

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. Retrieved from https://doi.org/10.1080/10705511.2014.961800

Morin, A. J., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques: Bifactor models, exploratory structural equation modeling (ESEM), and bifactor-ESEM. *Handbook of Sport Psychology*, 1044–1073. Retrieved from https://doi.org/10.1002/9781119568124.ch51

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. Retrieved from https://doi.org/10.1002/sim.8086

Mulaik, S. A. (2018). Fundamentals of common factor analysis. *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, 209–251.

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

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313. doi: 10.1037/a0026802

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

Patrick, C. J., Hicks, B. M., Nichol, P. E., & Krueger, R. F. (2007). A bifactor approach to modeling the structure of the psychopathy checklist-revised. *Journal of Personality Disorders*, *21*(2), 118–141. doi: 10.1521/pedi.2007.21.2.118

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

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 111–163. Retrieved from https://www.jstor.org/stable/271063

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

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

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the rosenberg self-esteem scale. *Multivariate Behavioral Research*, *51*(6), 818–838. doi: 10.1080/00273171.2016.1243461

Reise, S. P., Mansolf, M., & Haviland, M. G. (2023). Bifactor measurement models. *Handbook of Structural Equation Modeling*, 329–348.

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

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*, 19–31. doi: 10.1007/s11136-007-9183-7

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

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*(4), 561–582. Retrieved from `https://doi.org/10.1080/10705510903203433`

Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models.

Educational and Psychological Measurement, *72*(6), 910-932. doi: 10.1177/0013164412452564

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.

Spearman, C. (1961). "General intelligence" objectively determined and measured.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series b (Statistical Methodology)*, *64*(4), 583–639. Retrieved from https://doi.org/10.1111/1467-9868.00353

Sternberg, R. J., & Grigorenko, E. L. (2002). *The general factor of intelligence: How general is it?* Psychology Press. Retrieved from https://doi.org/10.4324/9781410613165

Suhr, D. (2006). *Exploratory or confirmatory factor analysis?* SAS Institute. Retrieved from https://books.google.com/books?id=c7B5AQAACAAJ

Tejada-Gallardo, C., Blasco-Belled, A., Torrelles-Nadal, C., & Alsinet, C. (2022). How does emotional intelligence predict happiness, optimism, and pessimism in adolescence? investigating the relationship from the bifactor model. *Current Psychology*, *41*(8), 5470–5480. Retrieved from https://doi.org/10.1007/s12144-020-01061-z

Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications. *Washington, DC*, *10694*(000), 3. Retrieved from https://doi.org/10.1037/10694-000

Thorndike, R. M. (1995). Book review: psychometric theory by Jum Nunnally and Ira Bernstein New York: McGraw-hill, 1994, xxiv+ 752 pp. *Applied Psychological Measurement*, *19*(3), 303–305.

Thurstone, L. L. (1938). Primary mental abilities: Psychometric monographs no. 1. In *The measurement of intelligence* (pp. 131–136). Springer.

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

Van De Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *European Health Psychologist*, *16*(2), 75–84.

Vats, D., & Knudson, C. (2021). Revisiting the Gelman–Rubin diagnostic. *Statistical Science*, *36*(4), 518–529. doi: 10.1214/20-STS812

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... others (2018). Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & review*, *25*, 35–57. doi: 10.3758/s13423-017-1343-3

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. doi: 10.1080/1047840x.2020.1853476

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. Retrieved from `https://doi.org/10.1177/2167702619855035`

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In *Handbook of structural equation modeling.* Guilford, New York.

Ximénez, C., Revuelta, J., & Castañeda, R. (2022). What are the consequences of ignoring cross-loadings in bifactor models? a simulation study assessing parameter recovery and sensitivity of goodness-of-fit indices. *Frontiers in Psychology*, *13*, 923877. doi: 10.3389/fpsyg.2022.923877

Yeo, Z. Z., & Suárez, L. (2022). Validation of the mental health continuum-short form: The bifactor model of emotional, social, and psychological well-being. *PLoS ONE*, *17*(5), e0268232. doi: 10.1371/journal.pone.0268232

Table 1
Summary of Simulation Factors and Prior Specifications

Simulation Factor	**Levels/Specifications**
General factor	0 (no general factor in the population) 0.10 (very small) 0.30 (small) 0.40 (moderate) 0.60 (large)
Cross-loadings	0 (no cross-loadings in the population) 0.10 (small) 0.20 (moderate) 0.30 (big)
Sample Size	$n = 100, 250, 500, 1000$
Model	**Full:** $\Lambda_G/\Lambda_C \sim \mathcal{N}(0, \infty)$ **Full:** $\Lambda_C \sim \mathcal{N}(0, \infty)$ $\Lambda_G \sim \mathcal{N}(0, 0.01)$ **Full:** $\Lambda_C \sim \mathcal{N}(0, \infty)$ $\Lambda_G \sim \mathcal{N}(0, .1)$ **Full:** $\Lambda_G \sim \mathcal{N}(true, 0.5 \times true)$ $\Lambda_C \sim \mathcal{N}(0, \infty)$ **No cross:** $\Lambda_G \sim \mathcal{N}(0, \infty)$ **No general:** $\Lambda_C \sim \mathcal{N}(0, \infty)$

Table 2

General factor loading inflation. The first three columns represent sample size, population general factor loadings and cross-loadings, respectively. Other columns represent general factor loadings for items x_1–x_{12}

Sample	Gen	Cross	No crossloadings model											
			λ_{G1}	λ_{G2}	λ_{G3}	λ_{G4}	λ_{G5}	λ_{G6}	λ_{G7}	λ_{G8}	λ_{G9}	λ_{G10}	λ_{G11}	λ_{G12}
100	0.4	0	-3.0	0.0	-2.0	0.4	-1.4	-2.8	-0.8	1.3	-4.9	-3.3	-5.6	-4.5
100	0.4	0.05	5.7	1.2	-0.9	1.8	7.2	-1.3	0.4	2.7	3.9	-1.8	-4.2	-3.3
100	0.4	0.1	**14.4**	2.3	0.2	3.1	**16.0**	0.5	1.9	4.4	**12.8**	-0.3	-2.6	-1.7
100	0.4	0.2	**31.0**	4.6	2.6	5.6	**33.0**	4.0	5.2	7.5	**29.6**	3.0	1.0	1.6
100	0.4	0.3	**45.7**	6.6	4.8	7.7	**47.8**	6.6	7.4	10.1	**44.8**	6.0	4.1	4.7
250	0.4	0	-1.8	-1.6	-2.4	-0.2	-0.2	-1.5	-0.8	-1.2	-2.9	-3.0	-3.1	-3.8
250	0.4	0.05	7.3	-0.6	-1.6	0.7	8.6	-0.7	-0.3	-0.4	6.1	-2.2	-2.2	-3.0
250	0.4	0.1	**16.1**	0.5	-0.7	1.7	**17.0**	0.0	0.3	0.2	**14.7**	-1.3	-1.2	-2.1
250	0.4	0.2	**32.1**	2.6	1.2	3.7	**32.1**	1.4	1.6	1.7	**30.9**	1.0	1.0	0.0
250	0.4	0.3	**45.9**	4.5	3.0	5.5	**45.5**	3.0	3.1	3.2	**45.1**	3.3	3.1	2.2
500	0.4	0	0.2	0.8	-0.5	0.7	0.5	-0.6	-0.1	0.6	-2.4	-3.1	-2.7	-3.3
500	0.4	0.05	9.3	1.0	-0.3	1.1	9.4	-0.2	0.3	0.8	6.9	-2.6	-1.9	-2.8
500	0.4	0.1	**17.8**	1.3	0.0	1.5	**18.0**	0.4	0.9	1.3	**15.6**	-2.0	-1.2	-2.2
500	0.4	0.2	**33.5**	2.4	1.2	2.8	**33.3**	1.8	2.3	2.5	**31.3**	-0.5	0.3	-0.7
500	0.4	0.3	**47.3**	3.9	2.9	4.5	**46.5**	3.4	3.7	3.8	**44.8**	1.1	1.8	0.9
1000	0.4	0	-1.0	-0.2	-1.1	0.0	0.6	-0.3	0.8	0.7	-0.8	-1.5	-1.5	-1.6
1000	0.4	0.05	8.2	0.0	-0.8	0.3	9.6	-0.1	0.9	0.7	8.3	-1.2	-1.2	-1.3
1000	0.4	0.1	**16.9**	0.4	-0.4	0.7	**18.1**	0.2	1.2	0.9	**16.9**	-0.8	-0.8	-0.9
1000	0.4	0.2	**32.6**	1.4	0.8	1.8	**33.5**	1.3	2.2	1.8	**32.5**	0.4	0.4	0.3
1000	0.4	0.3	**46.2**	2.8	2.3	3.2	**46.9**	2.7	3.7	3.1	**46.1**	1.9	1.9	1.7

Table 3

Specific factor loading underestimation. The first three columns represent sample size, population general factor loadings and cross-loadings, respectively. Other columns represent specific factor loadings for items x_1–x_{12} for corresponding factors S_1–S_3

Sample	Gen	Cross	No cross											
			λ_{1S1}	λ_{2S1}	λ_{3S1}	λ_{4S1}	λ_{5S2}	λ_{6S2}	λ_{7S2}	λ_{8S2}	λ_{9S3}	λ_{10S3}	λ_{11S3}	λ_{12S3}
100	0.4	0	2.3	1.2	0.4	1.7	0.6	0.9	0.2	2.3	3.4	3.4	4.5	3.7
100	0.4	0.05	-1.0	0.7	-0.2	1.1	-2.8	0.4	-0.3	1.7	0.2	2.9	4.0	3.2
100	0.4	0.1	-4.5	0.1	-0.9	0.5	-6.3	-0.4	-1.0	1.0	-3.2	2.1	3.2	2.5
100	0.4	0.2	**-11.6**	-1.6	-2.6	-1.1	**-13.4**	-2.3	-2.6	-0.8	**-10.0**	0.2	1.3	0.7
100	0.4	0.3	**-18.4**	-3.4	-4.6	-2.7	**-20.3**	-4.2	-4.2	-2.7	**-16.8**	-1.9	-0.6	-1.2
250	0.4	0	0.3	0.1	-0.3	0.0	-0.6	-0.4	-0.5	0.3	1.8	0.7	1.3	1.2
250	0.4	0.05	-3.6	-0.3	-0.6	-0.4	-4.4	-0.7	-0.8	0.0	-1.9	0.4	0.9	0.9
250	0.4	0.1	-7.5	-0.8	-1.0	-0.8	-8.1	-1.1	-1.2	-0.3	-5.6	-0.1	0.3	0.5
250	0.4	0.2	**-15.2**	-2.0	-2.1	-2.0	**-15.3**	-2.1	-2.2	-1.2	**-13.1**	-1.5	-1.1	-0.7
250	0.4	0.3	**-22.7**	-3.6	-3.7	-3.6	**-22.2**	-3.4	-3.6	-2.6	**-20.6**	-3.6	-2.8	-2.5
500	0.4	0	0.0	-0.4	-0.3	-0.3	-0.1	-0.1	0.1	-0.2	1.6	1.0	0.5	1.5
500	0.4	0.05	-4.0	-0.5	-0.2	-0.4	-4.1	-0.2	0.0	-0.3	-2.4	0.8	0.2	1.4
500	0.4	0.1	-7.9	-0.7	-0.3	-0.6	-8.0	-0.5	-0.4	-0.5	-6.2	0.5	-0.2	1.1
500	0.4	0.2	**-15.4**	-1.4	-1.0	-1.5	**-15.5**	-1.4	-1.4	-1.3	**-13.5**	-0.4	-1.1	0.4
500	0.4	0.3	**-23.2**	-2.7	-2.3	-3.0	**-22.8**	-2.7	-2.8	-2.5	**-20.7**	-1.8	-2.5	-0.8
1000	0.4	0	0.6	0.3	0.7	0.1	-0.1	-0.2	-0.2	-0.2	0.8	0.8	0.6	0.9
1000	0.4	0.05	-3.4	0.2	0.6	0.0	-4.1	-0.2	-0.2	-0.2	-3.2	0.7	0.4	0.8
1000	0.4	0.1	-7.3	0.0	0.4	-0.2	-8.0	-0.4	-0.4	-0.3	-7.0	0.5	0.2	0.6
1000	0.4	0.2	**-14.6**	-0.6	-0.2	-0.9	**-15.2**	-1.0	-1.1	-0.9	**-14.3**	-0.2	-0.5	-0.1
1000	0.4	0.3	**-21.6**	-1.5	-1.2	-1.9	**-22.3**	-2.0	-2.2	-1.8	**-21.4**	-1.3	-1.6	-1.1

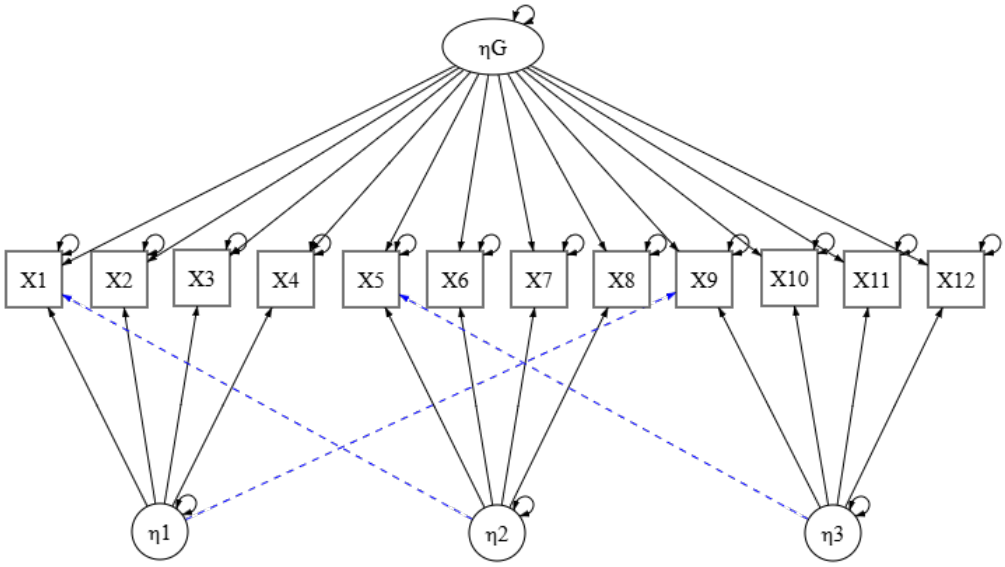
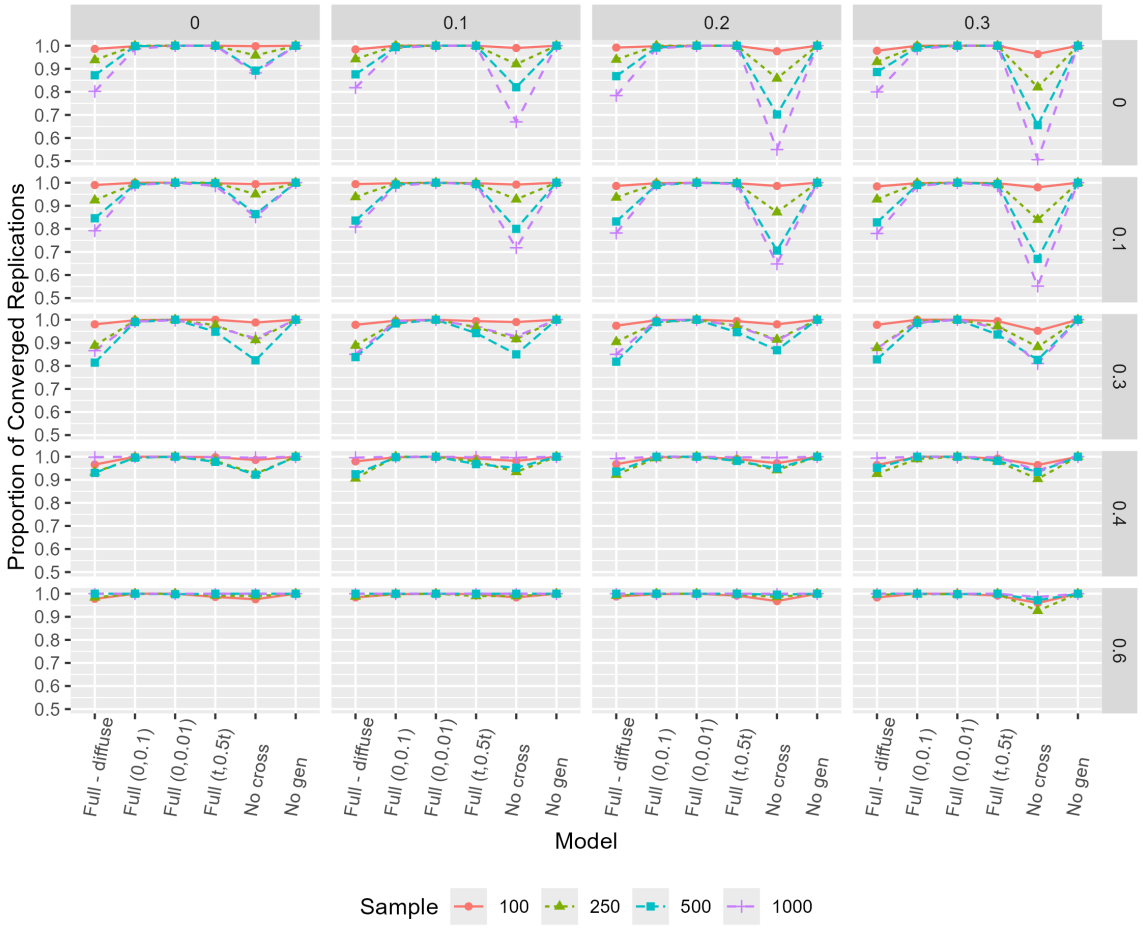


Figure 1

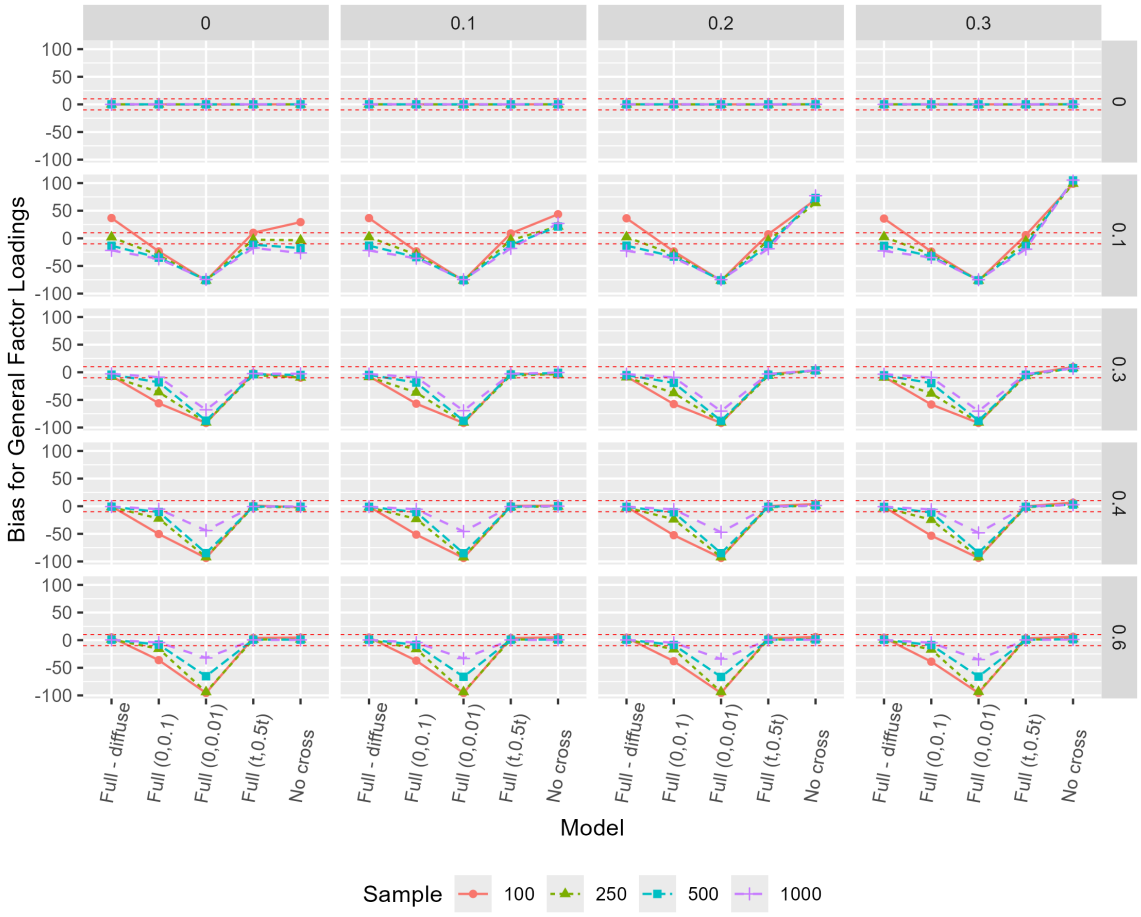
The data generation model. The specific factors are η_1, η_2 and η_3, η_G is the general factor, x_{1-12} are the observed indicators. The black arrows represent general and specific factor loadings. The blue dashed arrows represent cross-loadings.



;

Figure 2

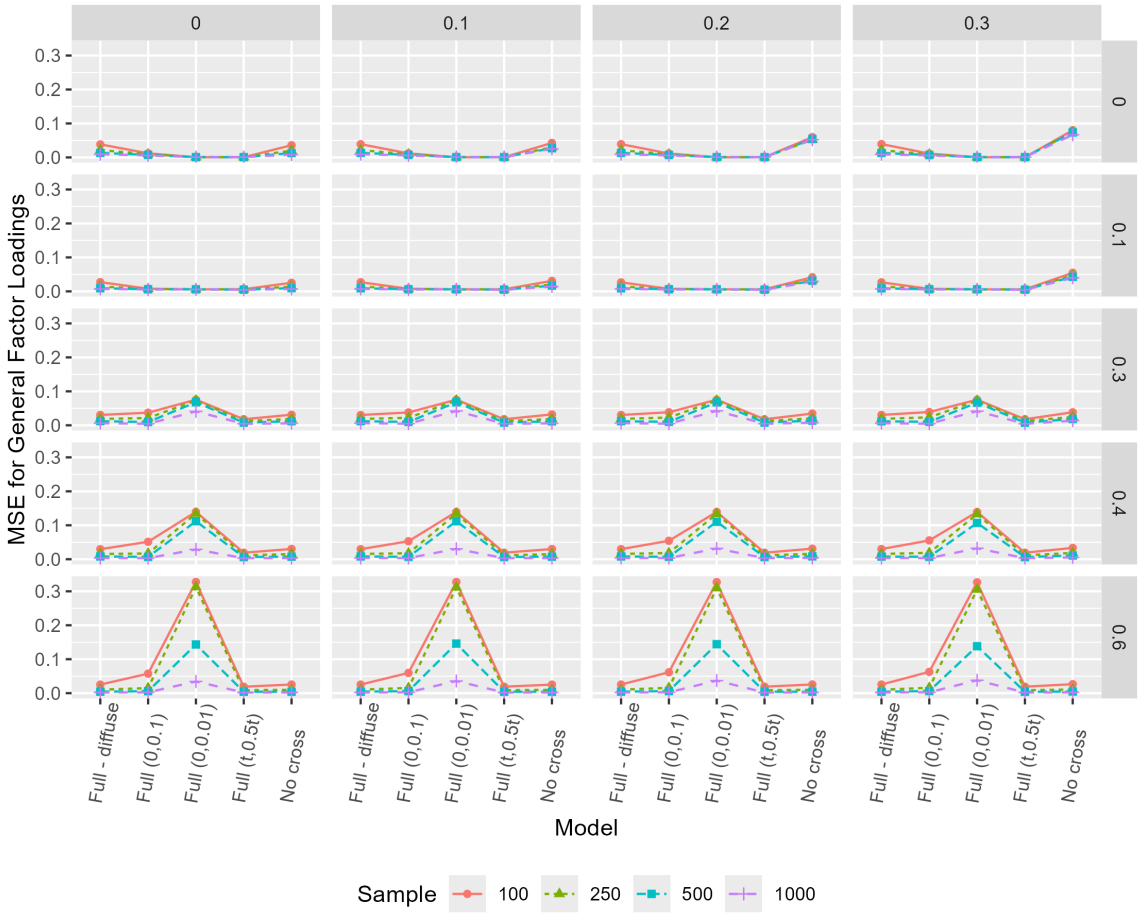
The proportion of converged replications. The x-axis represents the fitted models, including full model with different prior settings. The y-axis shows the proportion of replications with largest $\hat{R} < 1.05$*. The rows and columns represent populations general factor loadings and cross-loadings, respectively.*



;

Figure 3

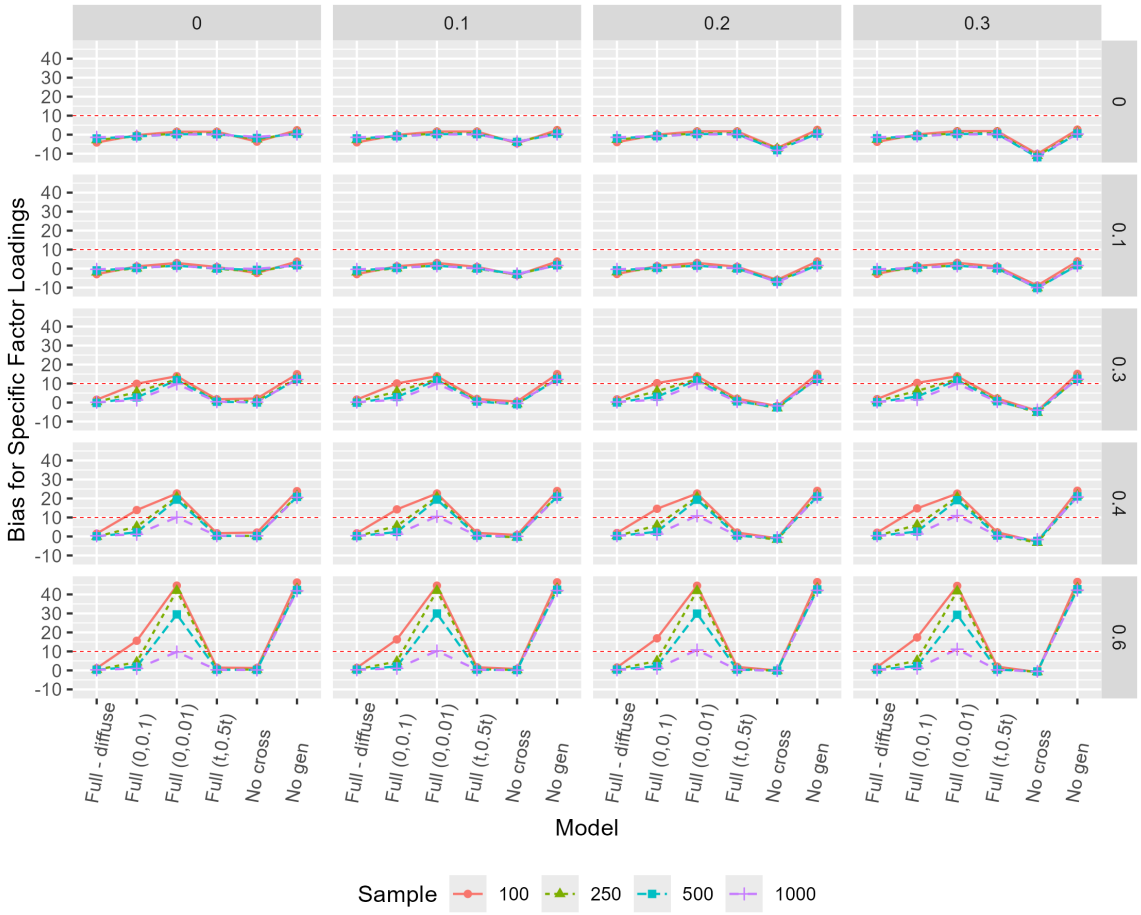
Bias computed for general factor loadings. The x-axis represents the fitted models, including full model with different prior settings. The y-axis shows the percentage of bias. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Raw bias is shown in the first row, relative bias is shown in remaining rows. Line type and line color represent sample sizes. The red dashed horizontal lines indicate the acceptable range of bias $\pm 10\%$.



;

Figure 4

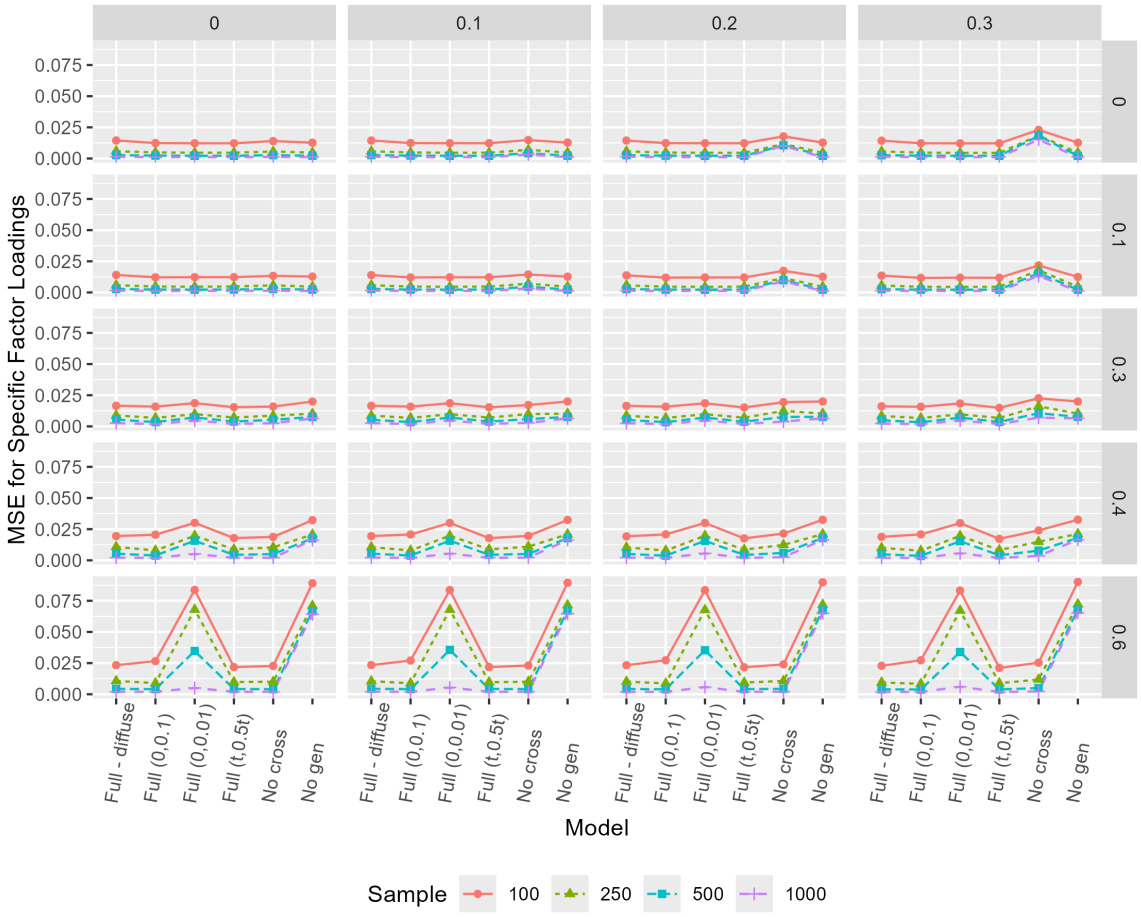
MSE computed for general factor loadings. The x-axis represents the fitted models, including the full model with different prior settings. The y-axis shows the MSE magnitude. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample size.



;

Figure 5

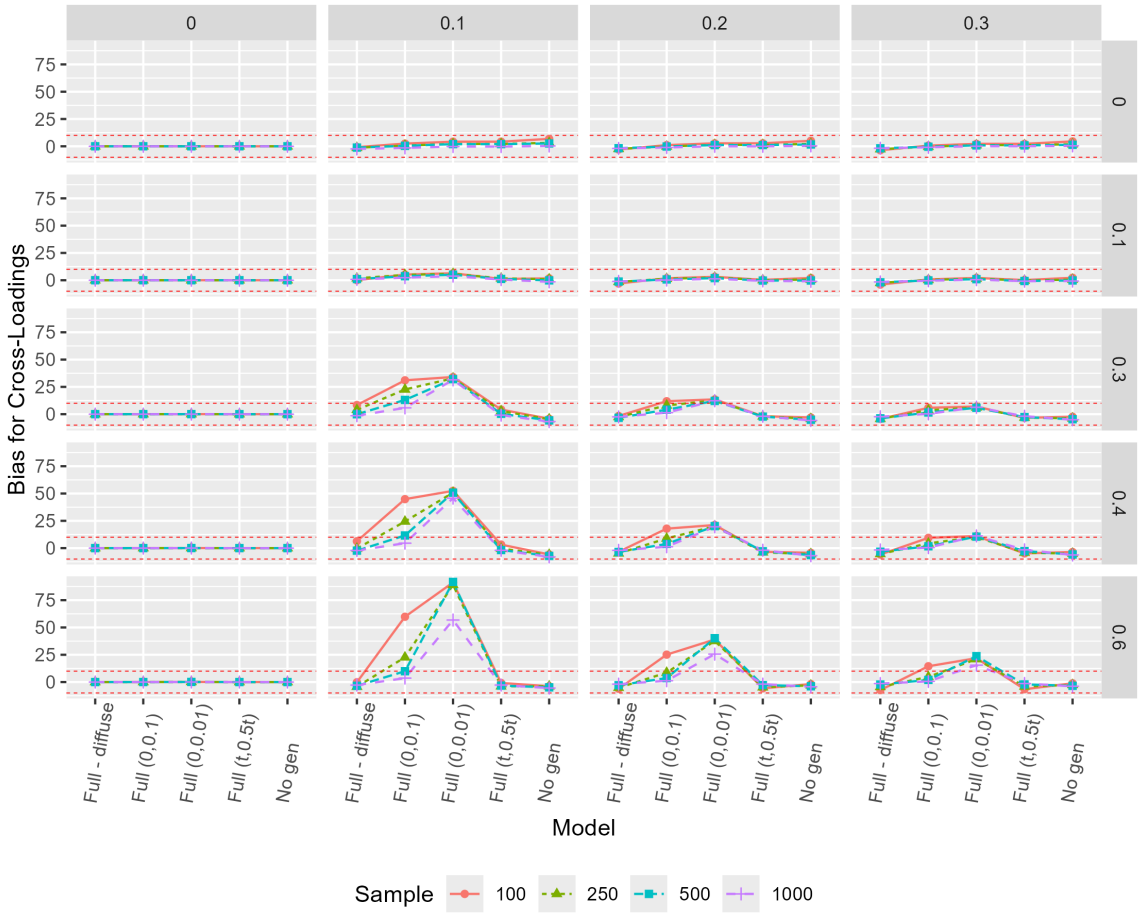
Relative bias computed for specific factor loadings. The x-axis represents the fitted models, including full model with different prior settings. The y-axis shows the percentage of relative bias. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Linetype and line color represent sample sizes. The red dashed horizontal lines indicate the acceptable range of bias $\pm 10\%$.



;

Figure 6

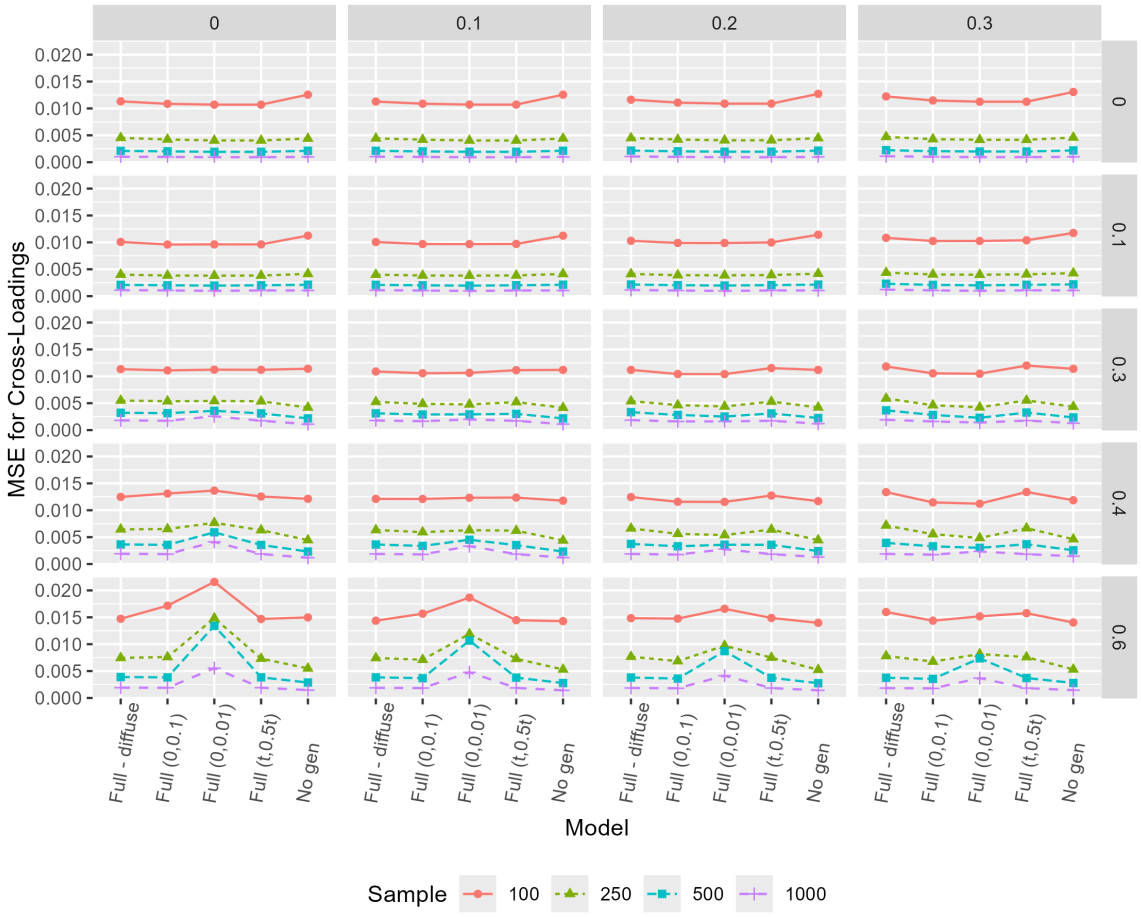
MSE computed for specific factor loadings. The x-axis represents the fitted models, including full model with different prior settings. The y-axis shows the magnitude of MSE. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample size.



;

Figure 7

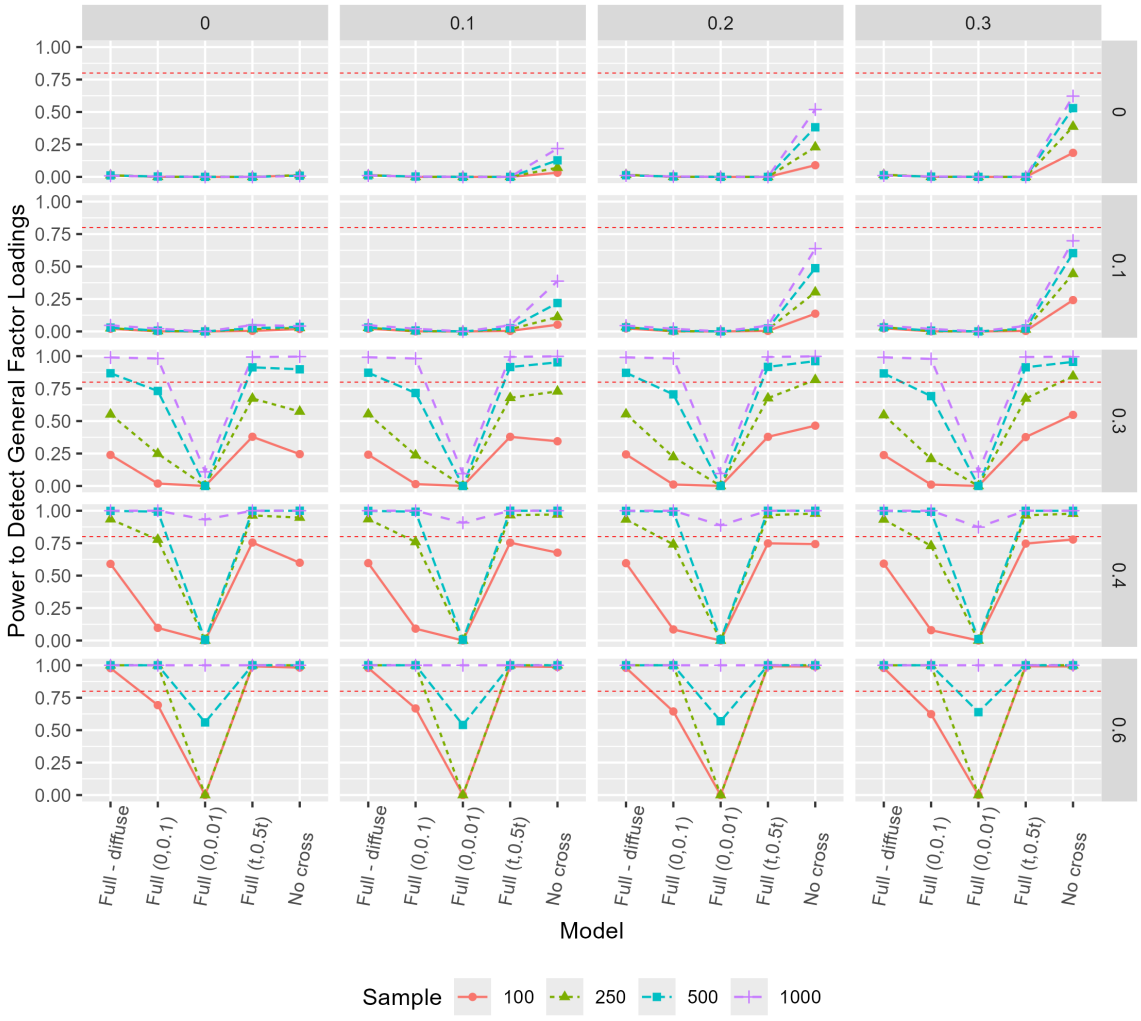
Bias computed for cross-loadings. The x-axis represents the fitted models, including full model with different prior settings. The y-axis shows the percentage of bias. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Raw bias is shown in the first column, relative bias is shown in remaining columns. Line type and line color represent sample sizes. The red dashed horizontal lines indicate the acceptable range of bias $\pm 10\%$.



;

Figure 8

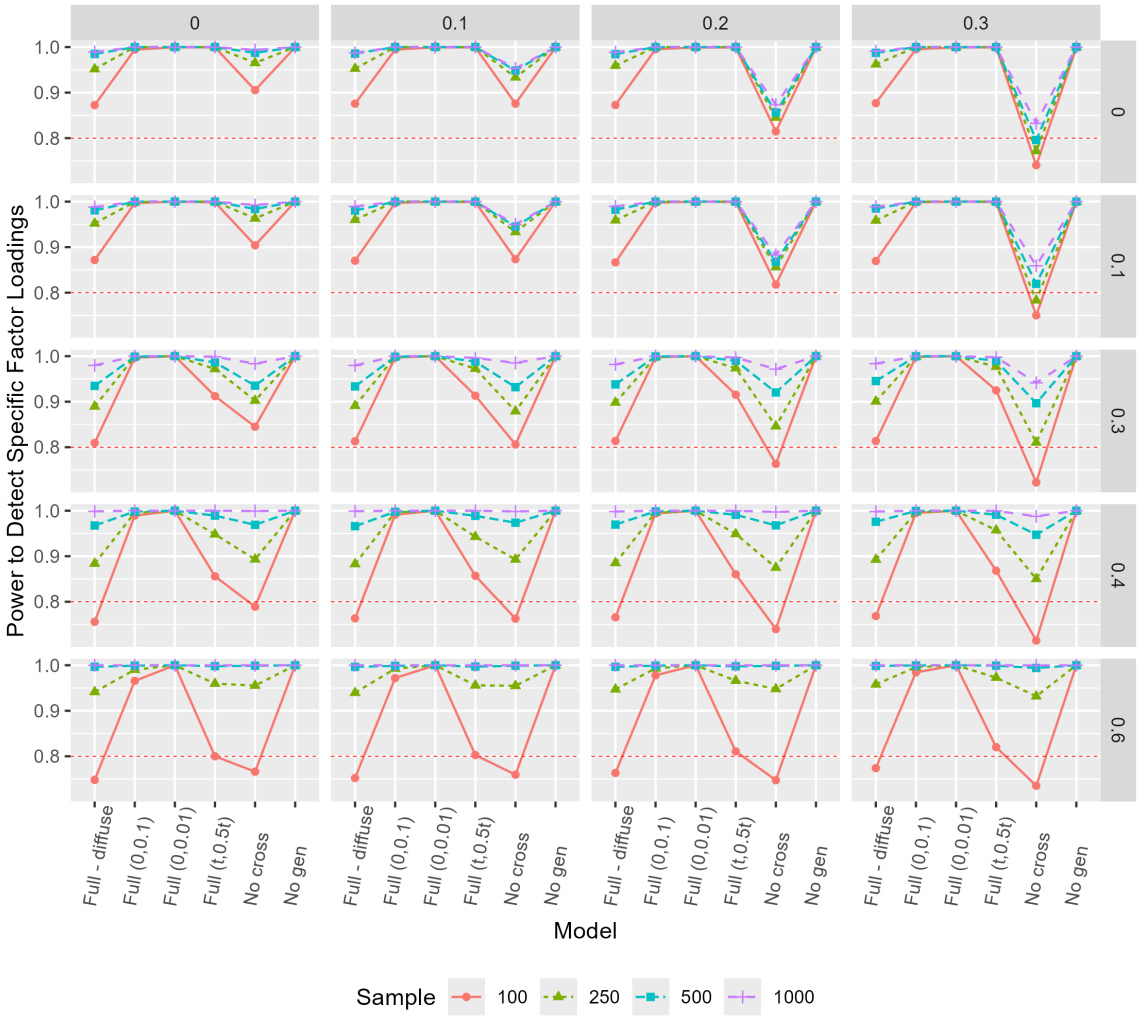
MSE computed for cross-loadings. The x-axis represents the fitted models, including full model with different prior settings. The y-axis shows the magnitude of MSE. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample size.



;

Figure 9

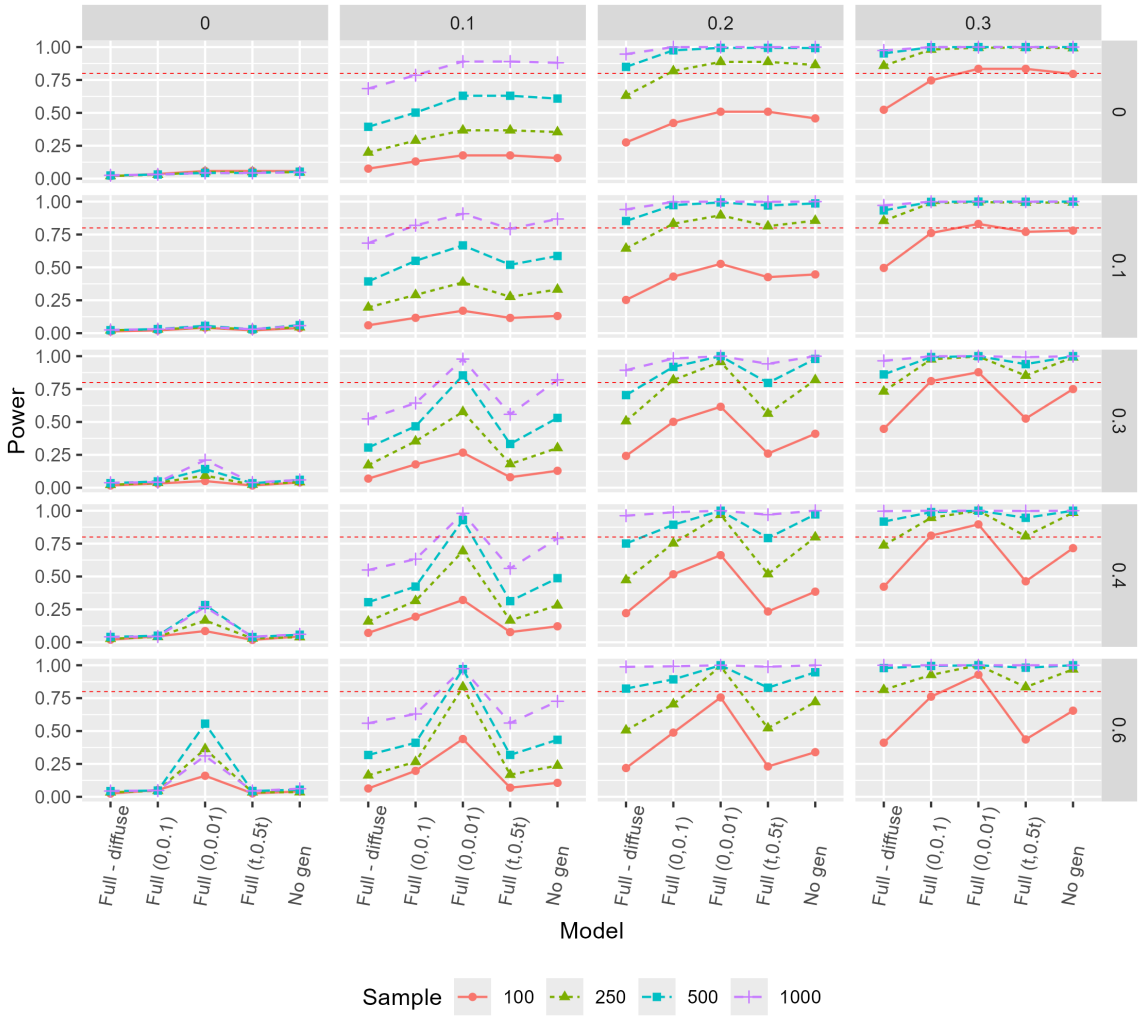
The Bayesian "Power" of detecting general factor loadings. The x-axis represents the fitted models, including the full model with different prior settings. The y-axis shows Bayesian "Power". The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes. The red dashed horizontal line indicates the 80% power



;

Figure 10

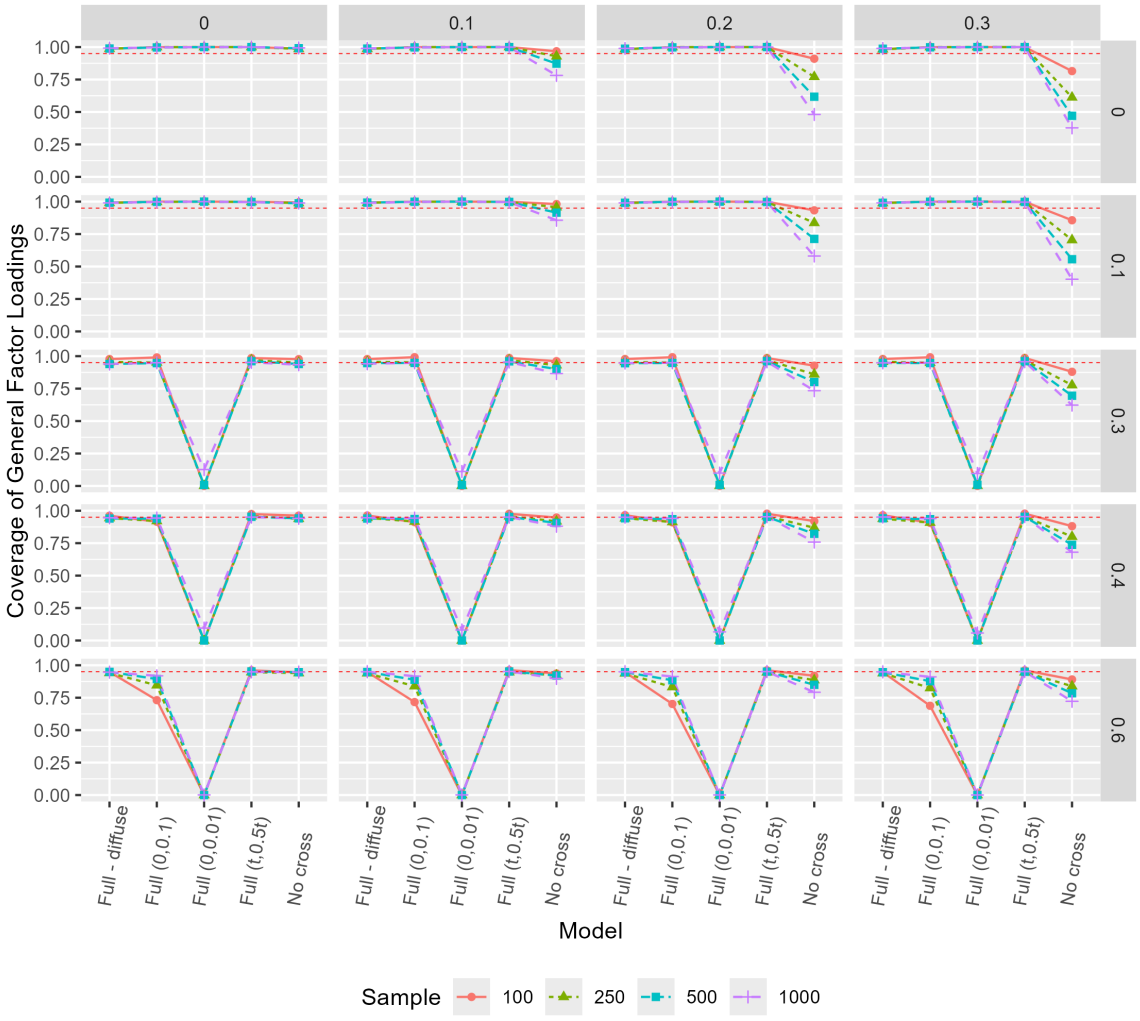
Bayesian "Power" computed for specific factor loadings. The x-axis represents the fitted models, including the full model with different prior settings. The y-axis shows Bayesian "Power". The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes. The red dashed horizontal line indicates the 80% power



;

Figure 11

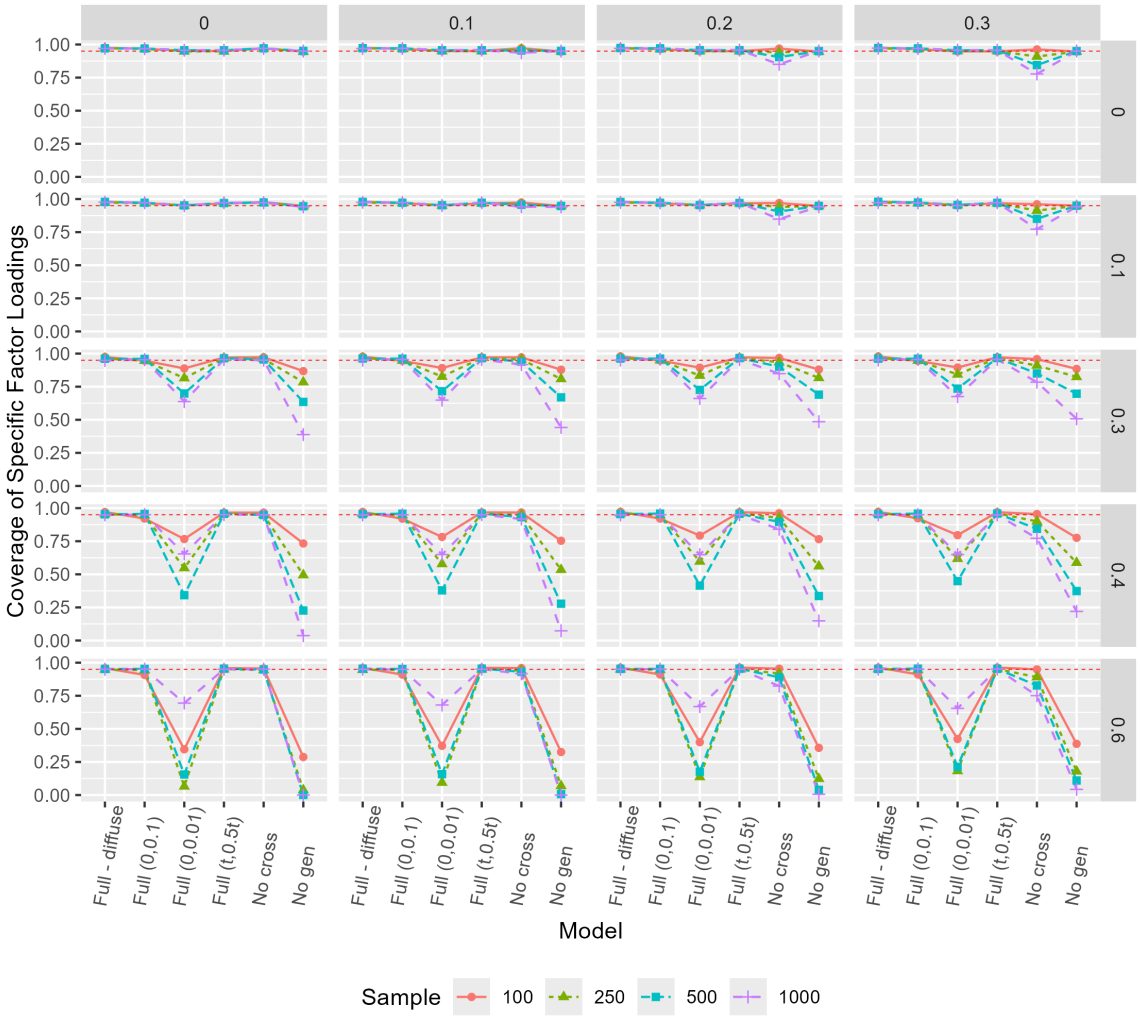
Bayesian "Power" computed for cross-loadings. The x-axis represents the fitted models, including full model with different prior settings. The y-axis shows Bayesian "Power". The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes. The red dashed horizontal line indicates the 80% power



;

Figure 12

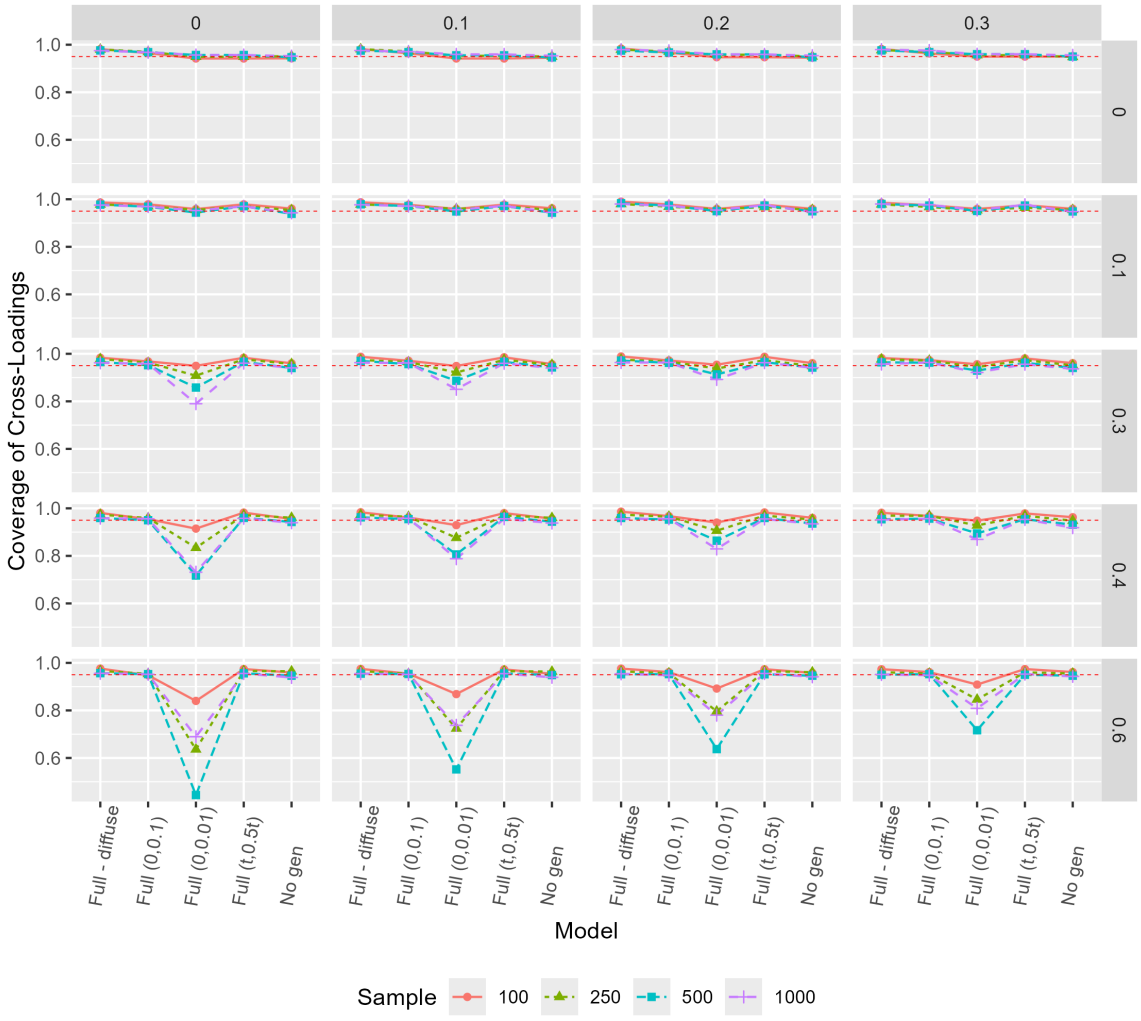
Bayesian coverage computed for general factor loadings. The x-axis represents the fitted models, including the full model with different prior settings. The y-axis shows Bayesian coverage rates. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes. The red dashed horizontal line represents the 0.95 coverage rate



;

Figure 13

Bayesian Coverage computed for specific factor loadings. The x-axis represents the fitted models, including full model with different prior settings. The y-axis shows Bayesian coverage rates. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes. The red dashed horizontal line represents the 0.95 coverage rate



;

Figure 14

Bayesian Coverage computed for cross-loadings. The x-axis represents the fitted models, including full model with different prior settings. The y-axis shows Bayesian coverage rates. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes. The red dashed horizontal line represents the 0.95 coverage rate

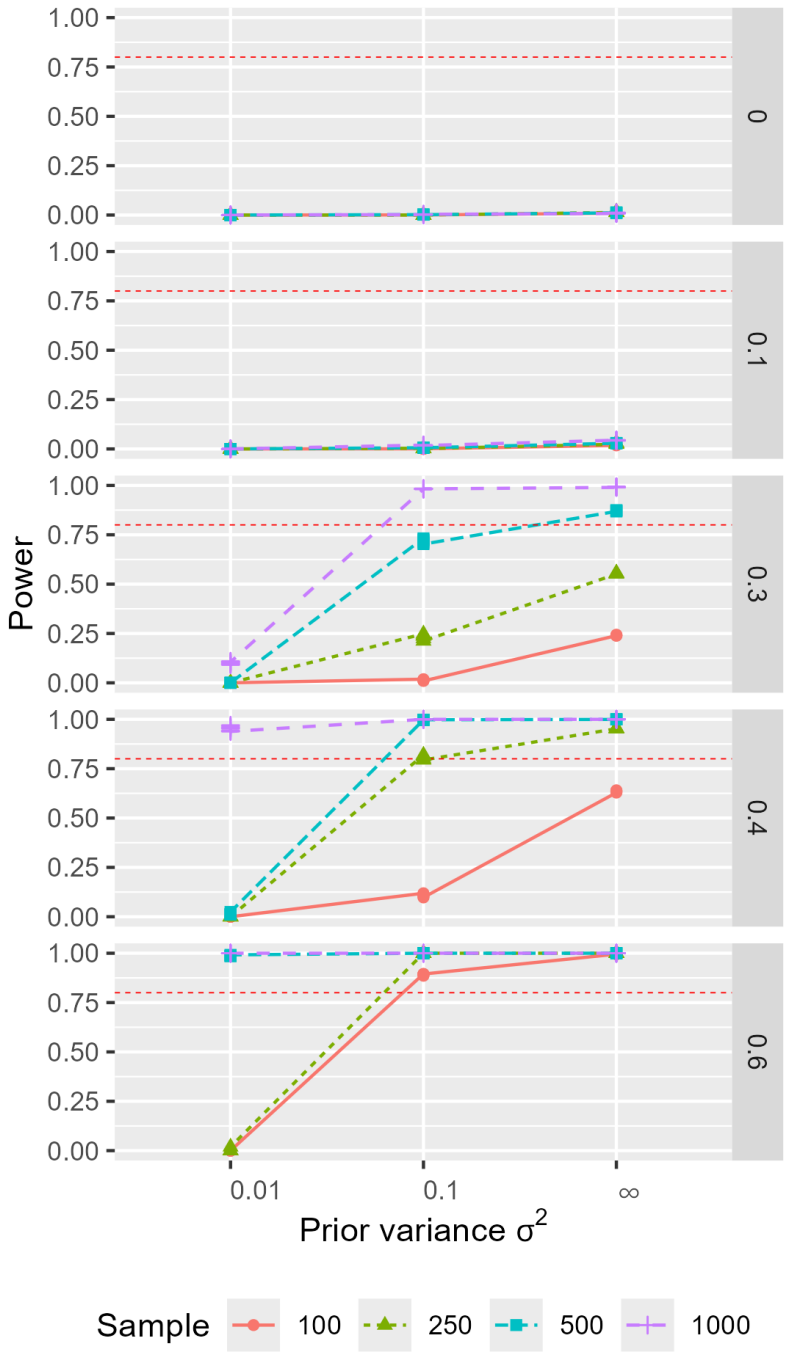
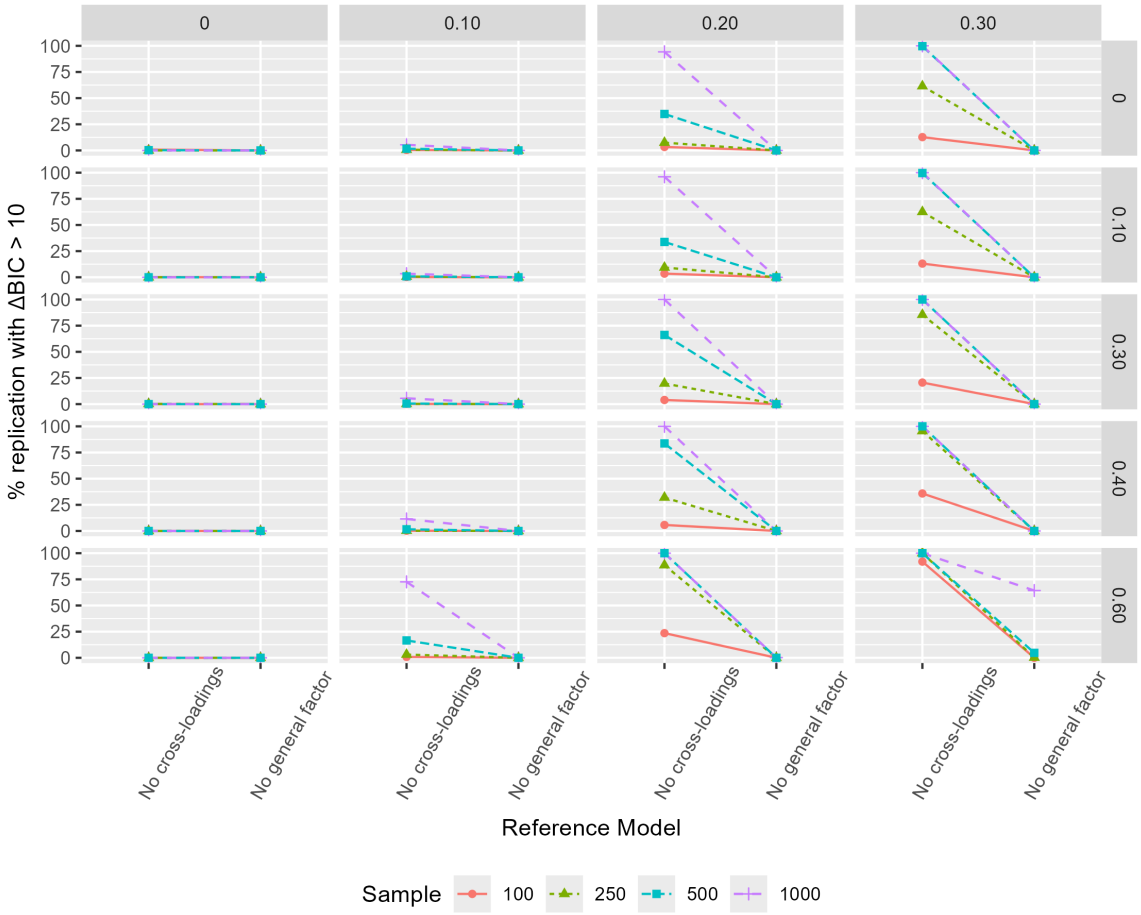


Figure 15

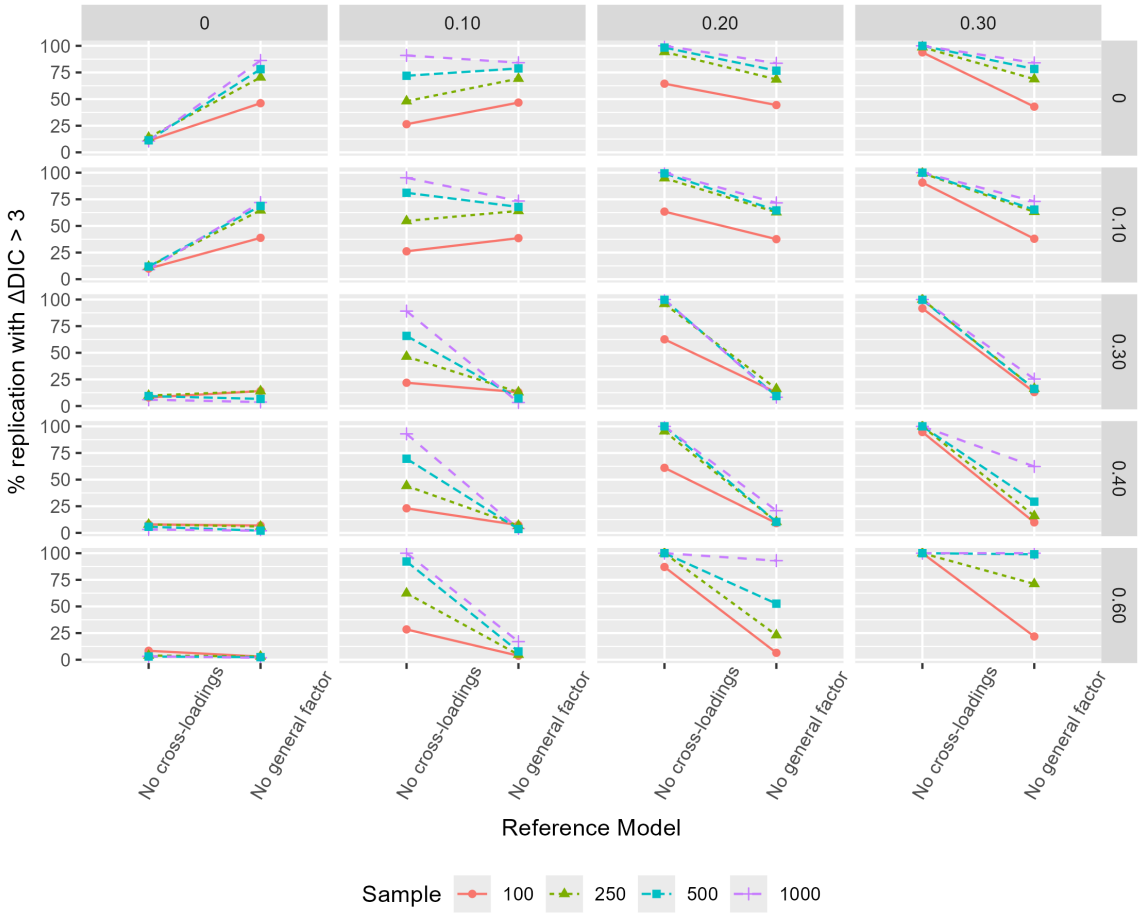
Bayesian "Power" to detect population general factor loadings depending on their assigned prior. All priors are centered at 0. The x-axis represents prior's variance parameter. The y-axis shows Bayesian "Power" to detect a non-zero general factor loading, or "Type I error rate", when the population loading is zero. The rows represent populations general factor loadings. The line type and line color represent sample sizes.



;

Figure 16

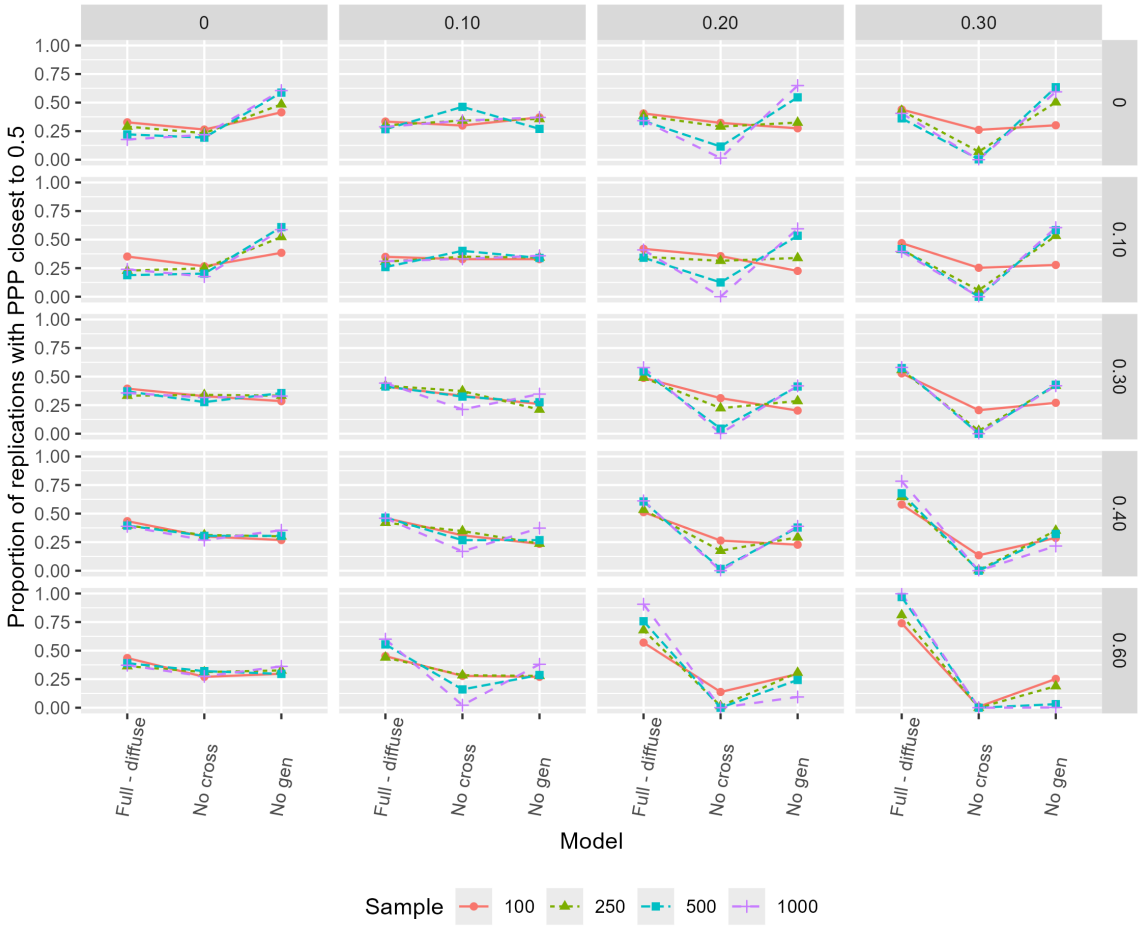
The BIC selection rates based on a cutoff for BIC difference of competing models. The x-axis represents the reference model against which the full model is compared. The y-axis shows the percentage of replications where the full model shows better fit than the reference model, with BIC difference higher than $\Delta BIC = 10$*. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample size.*



;

Figure 17

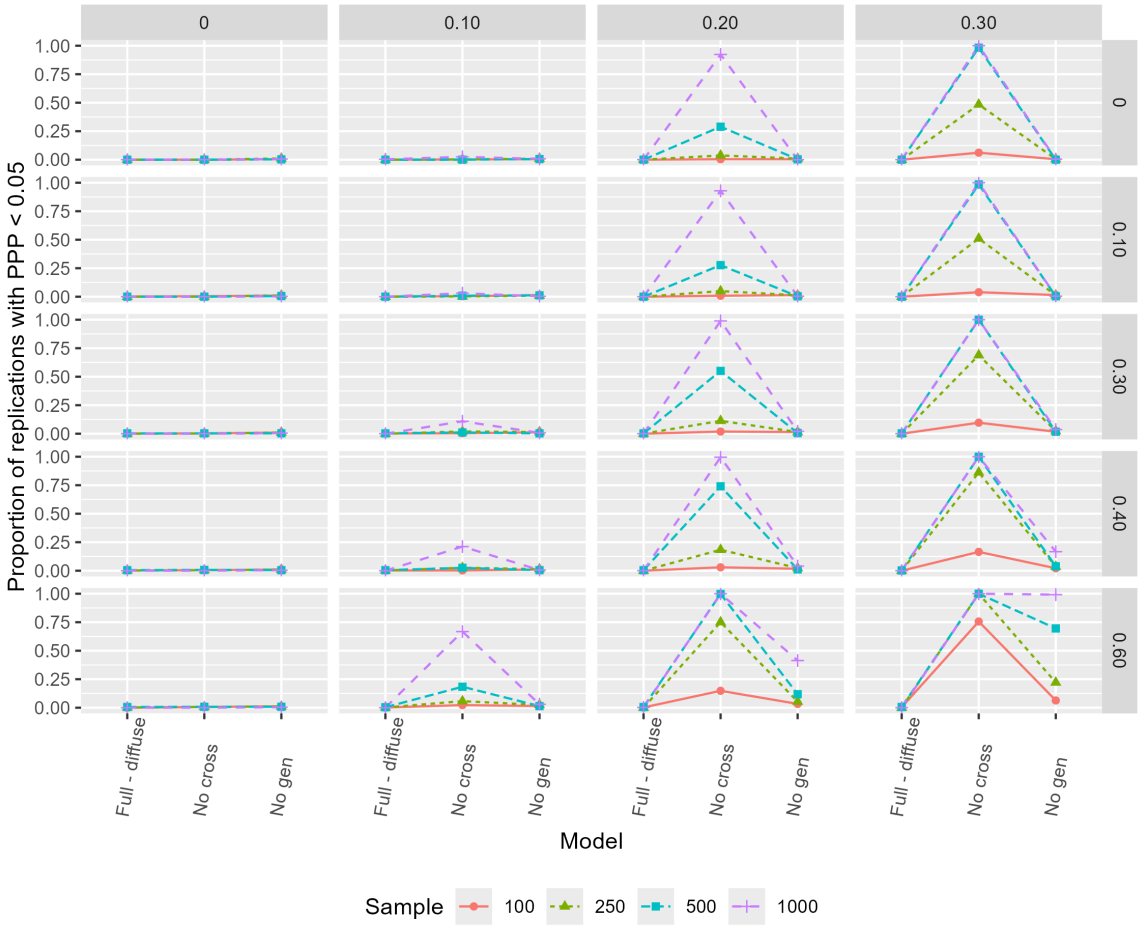
The DIC selection rates based on a cutoff for DIC difference of competing models. The x-axis represents the reference model against which the full model is compared. The y-axis shows the percentage of replications where the full model shows better fit than the reference model, with DIC difference higher than $\Delta DIC = 3$. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample size.



;

Figure 18

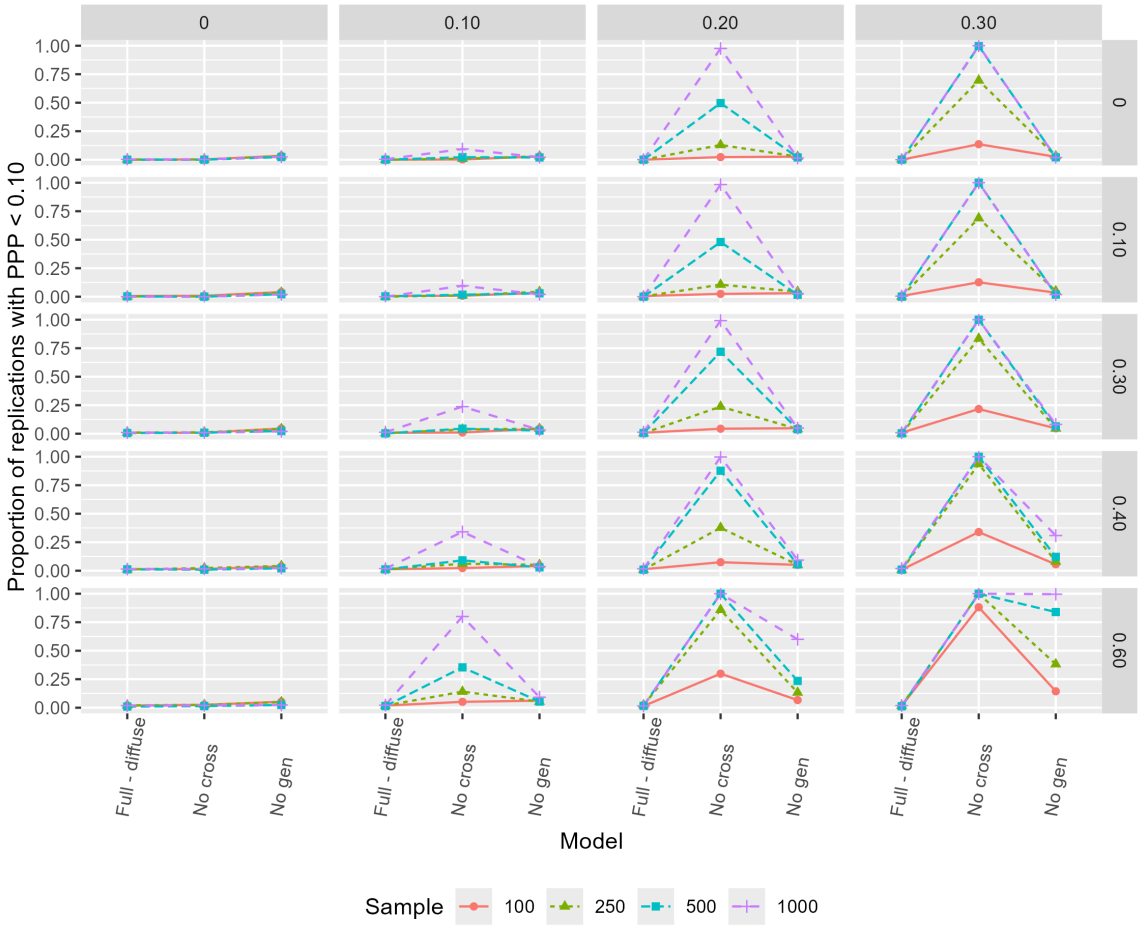
The selection among the three fitted models based on the PPP. The y-axis shows the percentage of replications where a given model had PPP closest to 0.5 among the three models. The x-axis represents the three fitted models. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes.



;

Figure 19

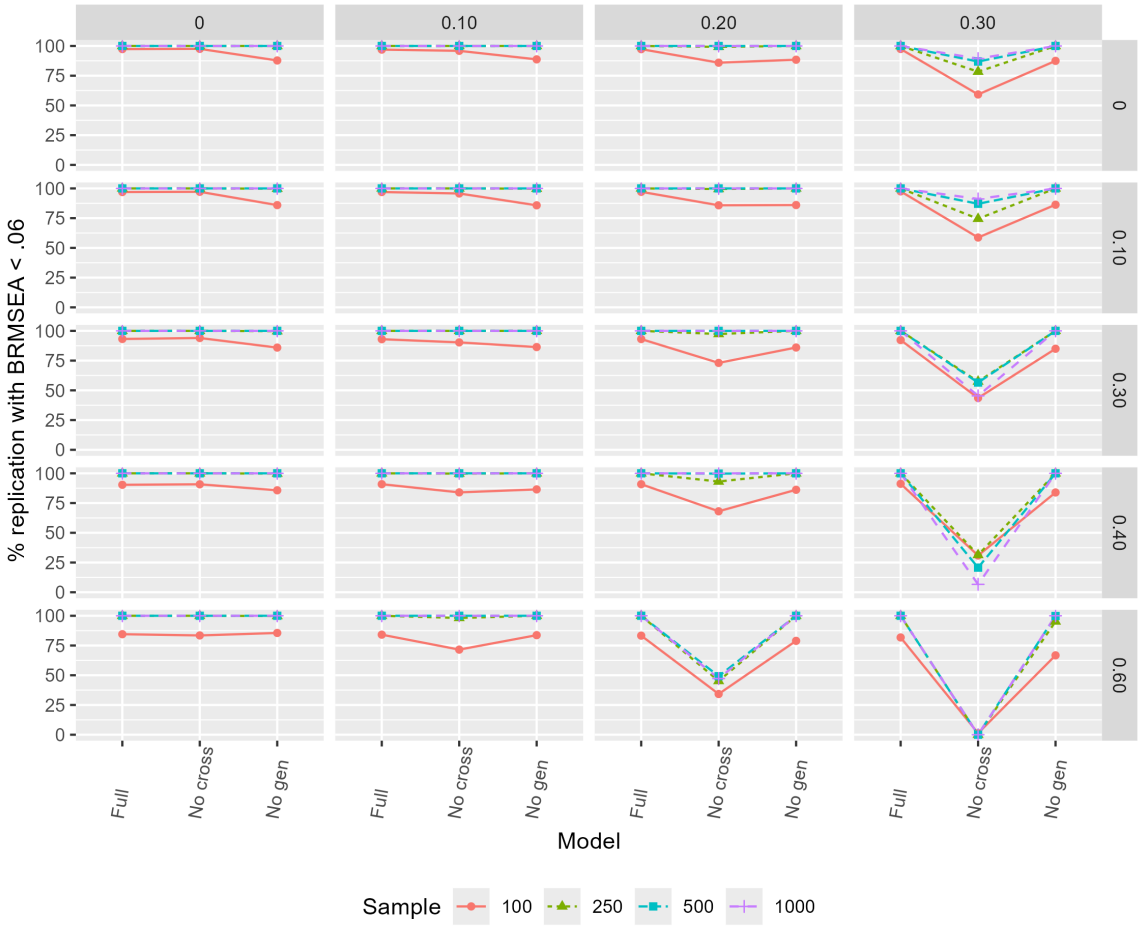
The PPP selection rates (0.05 cutoff). The y-axis shows the proportion of replications where PPP was below the 0.05 cutoff, indicating poor fit. The x-axis represents the fitted models. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes.



;

Figure 20

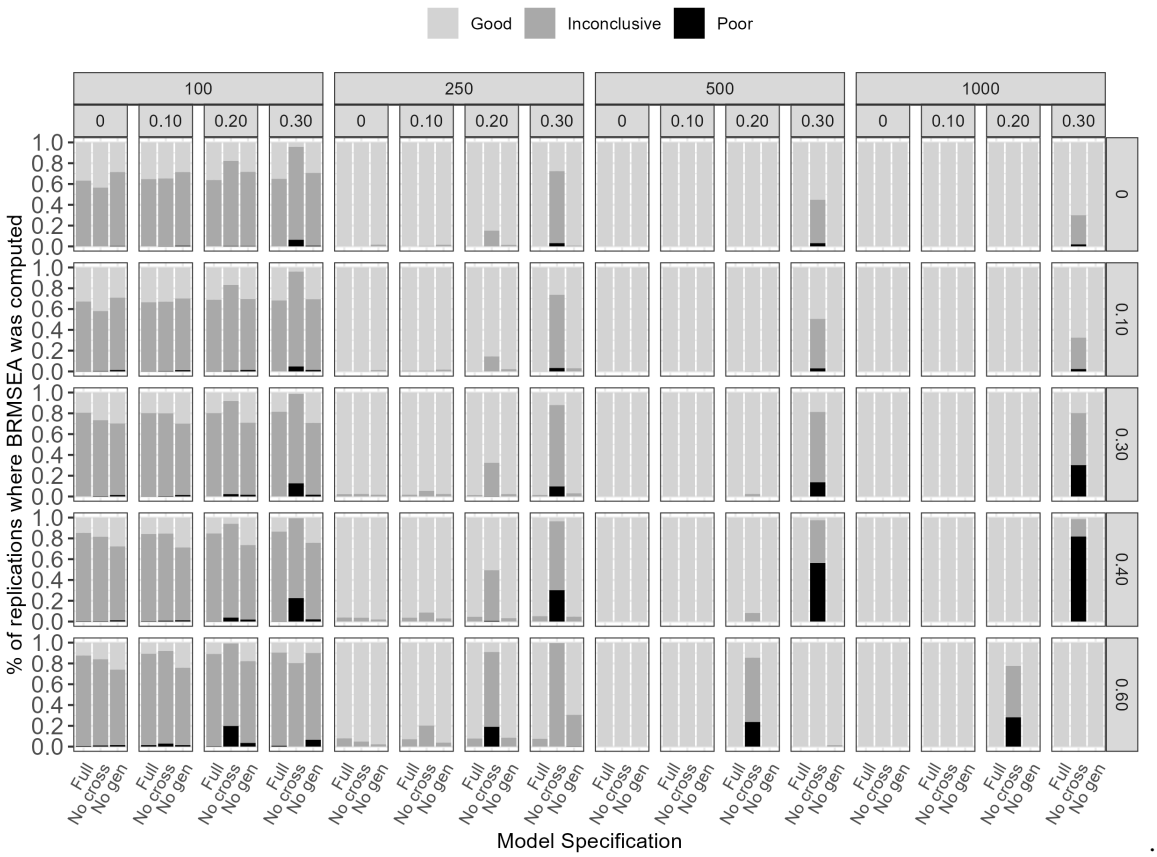
The PPP selection rates (0.10 cutoff). The y-axis shows the proportion of replications where PPP was below the 0.10 cutoff, indicating poor fit. The x-axis represents the fitted models. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes.



;

Figure 21

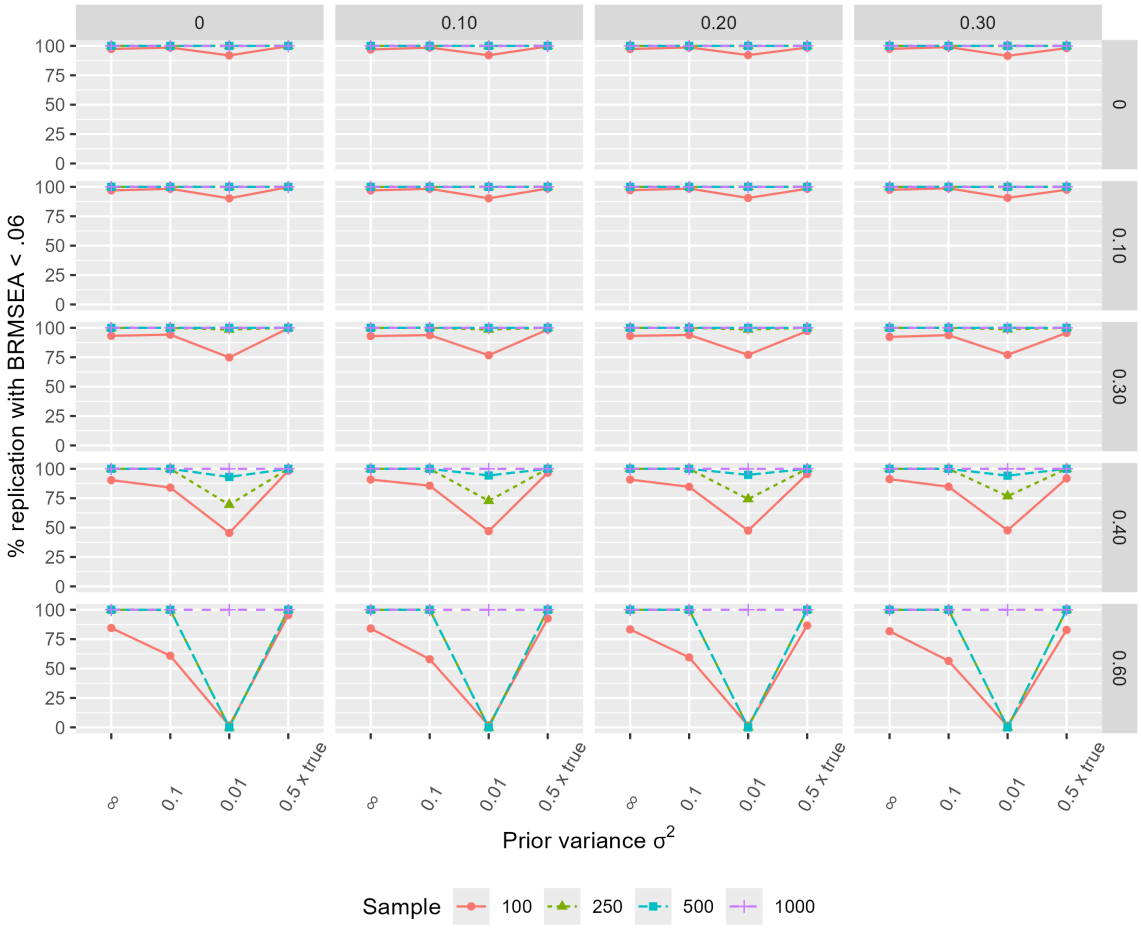
The BRMSEA Selection Rates. The y-axis shows the percentage of replications where BRMSEA was below the 0.6 cutoff, indicating good fit. The x-axis represents the fitted models. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes.



;

Figure 22

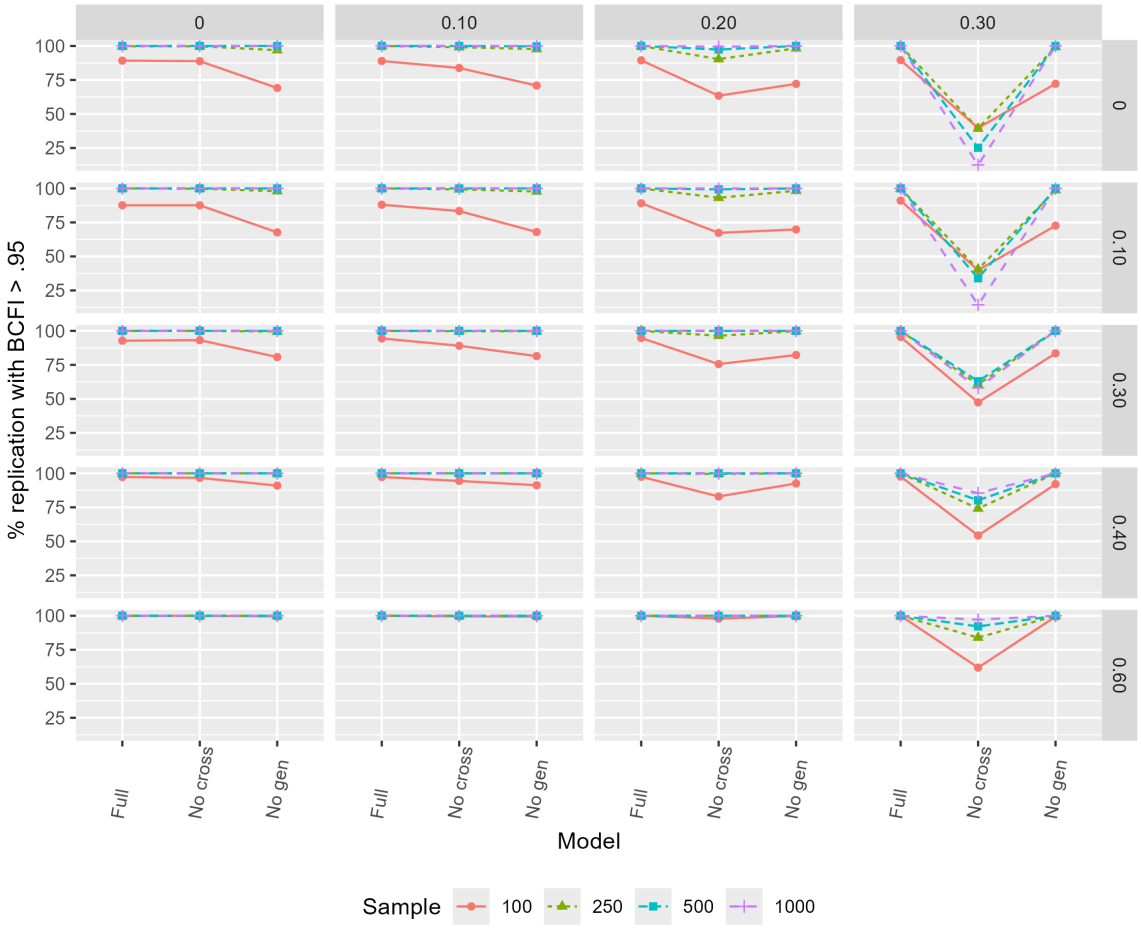
CI-based BRMSEA Selection Rates. The y-axis shows the proportion of replications. The x-axis represents fitted models. The colors of bars represent the decision about the fit: good, bad or inconclusive. Rows and first-order columns represent population general factor loadings and cross-loadings, respectively. Second-order columns represent sample size.



;

Figure 23

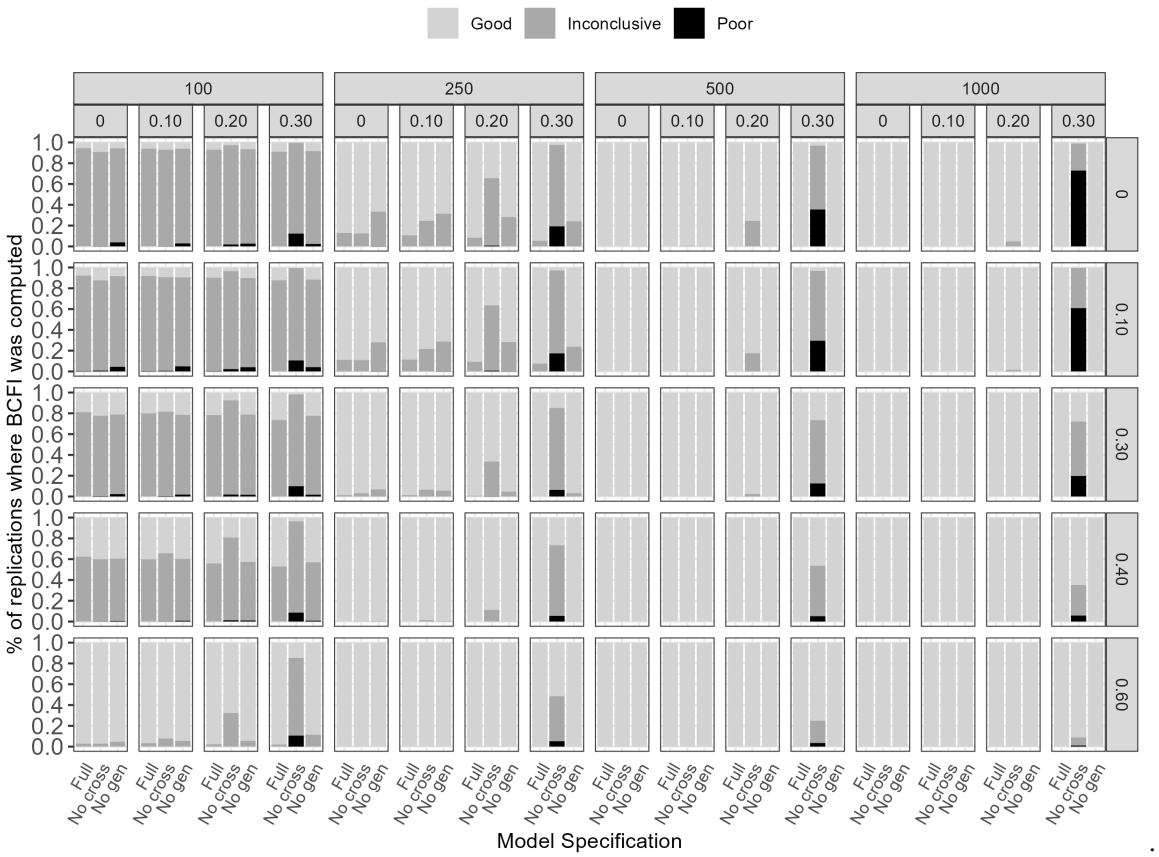
Prior Sensitivity of BRMSEA. The y-axis shows the percentage of replications where BRMSEA was below the 0.6 cutoff, indicating good fit. The x-axis represents variance of priors placed on general factor loadings. The last prior is centered at the true value, other priors are centered at zero. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes.



;

Figure 24

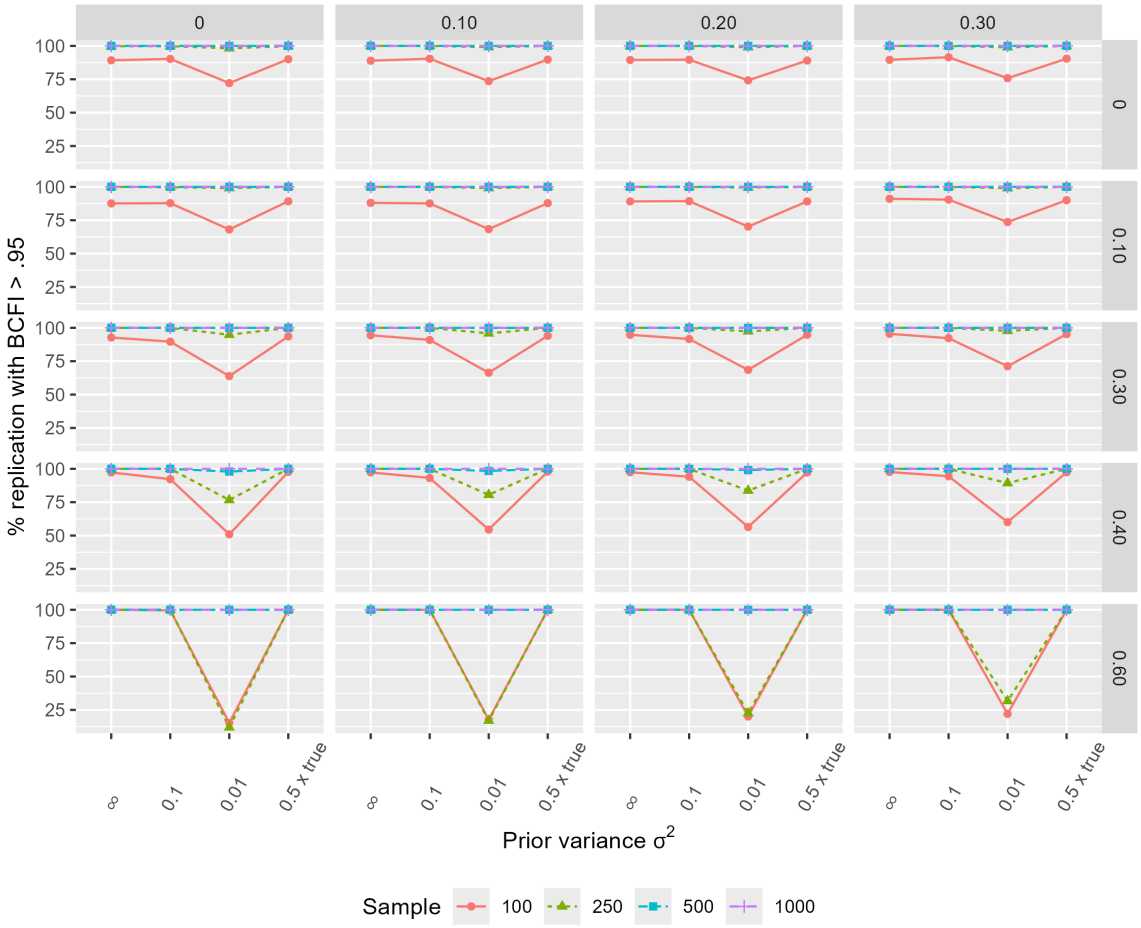
The BCFI Selection Rates. The y-axis shows the percentage of replications where BCFI was above the 0.95 cutoff, indicating good fit. The x-axis represents the fitted models. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes.



;

Figure 25

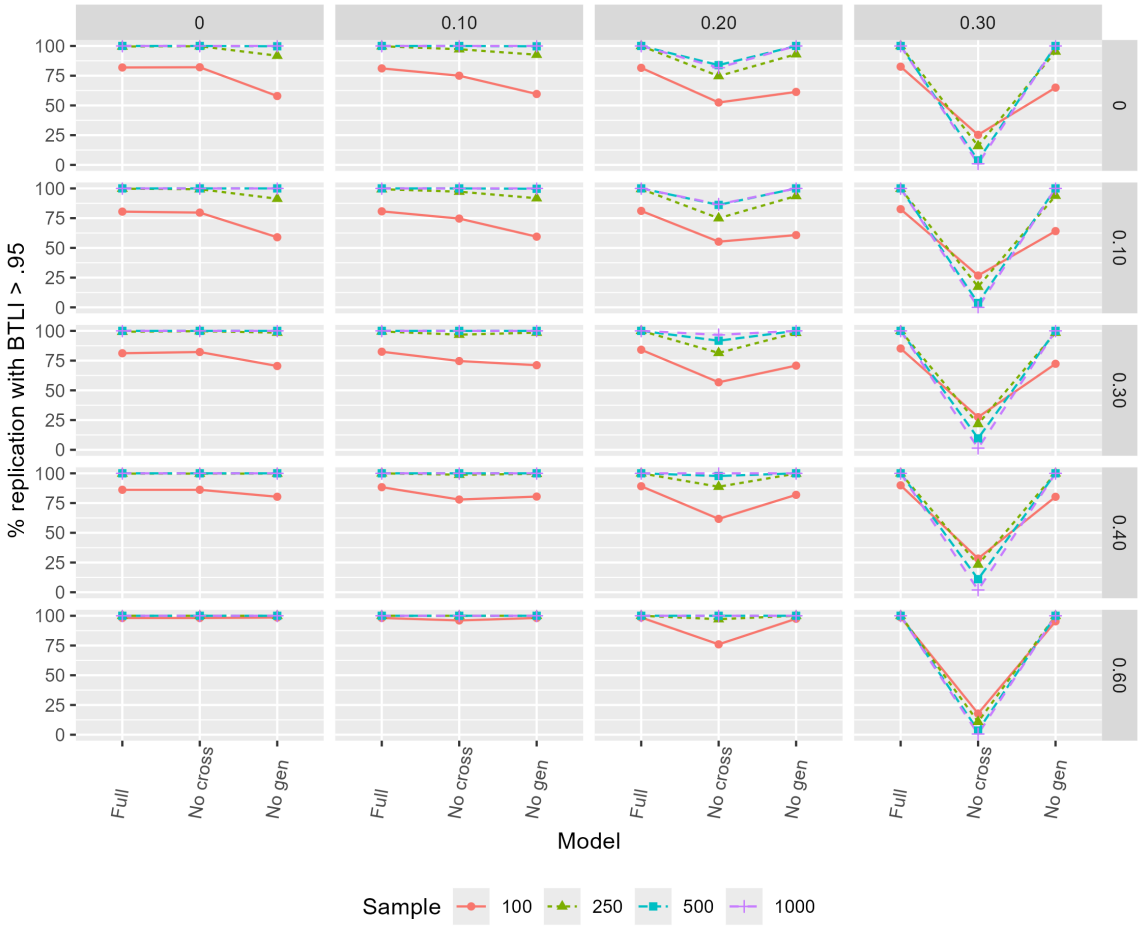
CI-based BCFI selection rates. The y-axis shows the proportion of replications. The x-axis represents fitted models. The colors of bars represent the decision about the fit: good, bad or inconclusive. Rows and first-order columns represent population general factor loadings and cross-loadings, respectively. Second-order columns represent sample size.



;

Figure 26

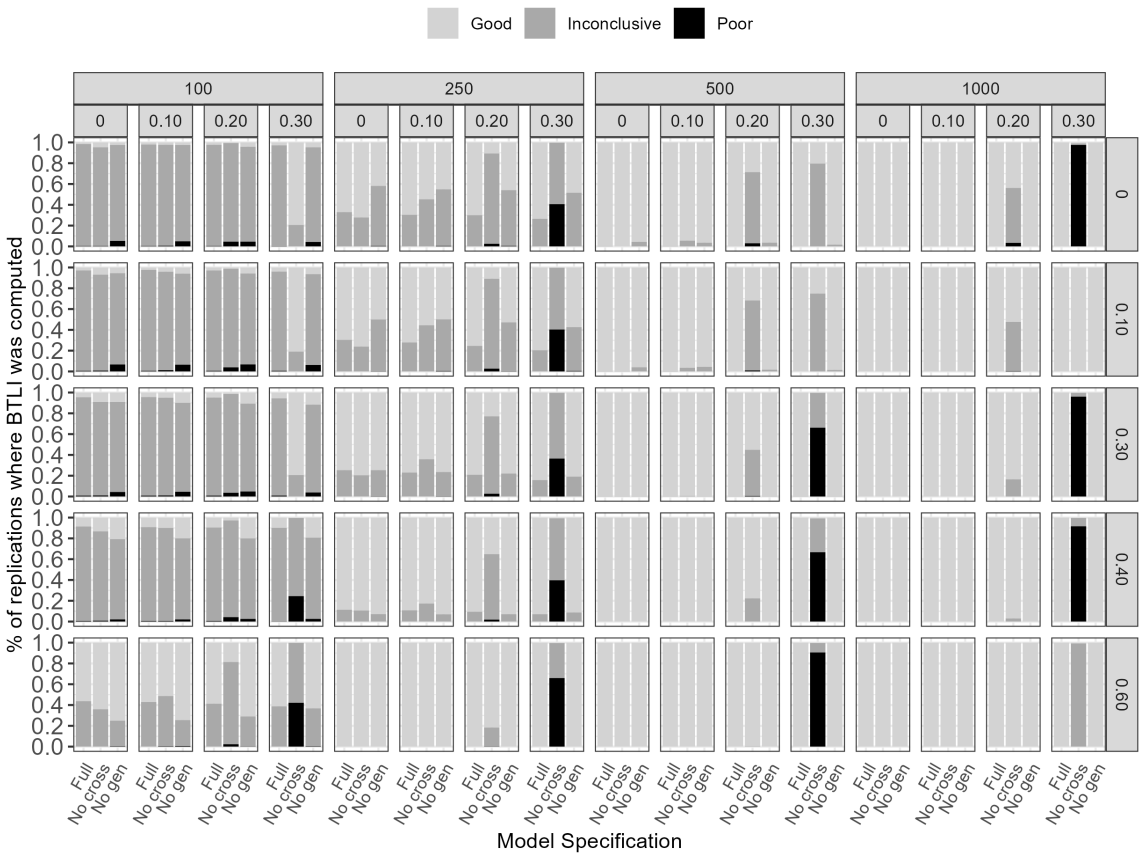
Prior sensitivity of BCFI. The y-axis shows the percentage of replications where BCFI was above the 0.95 cutoff, indicating good fit. The x-axis represents variance of priors placed on general factor loadings. The last prior is centered at the true value, other priors are centered at zero. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample size.



;

Figure 27

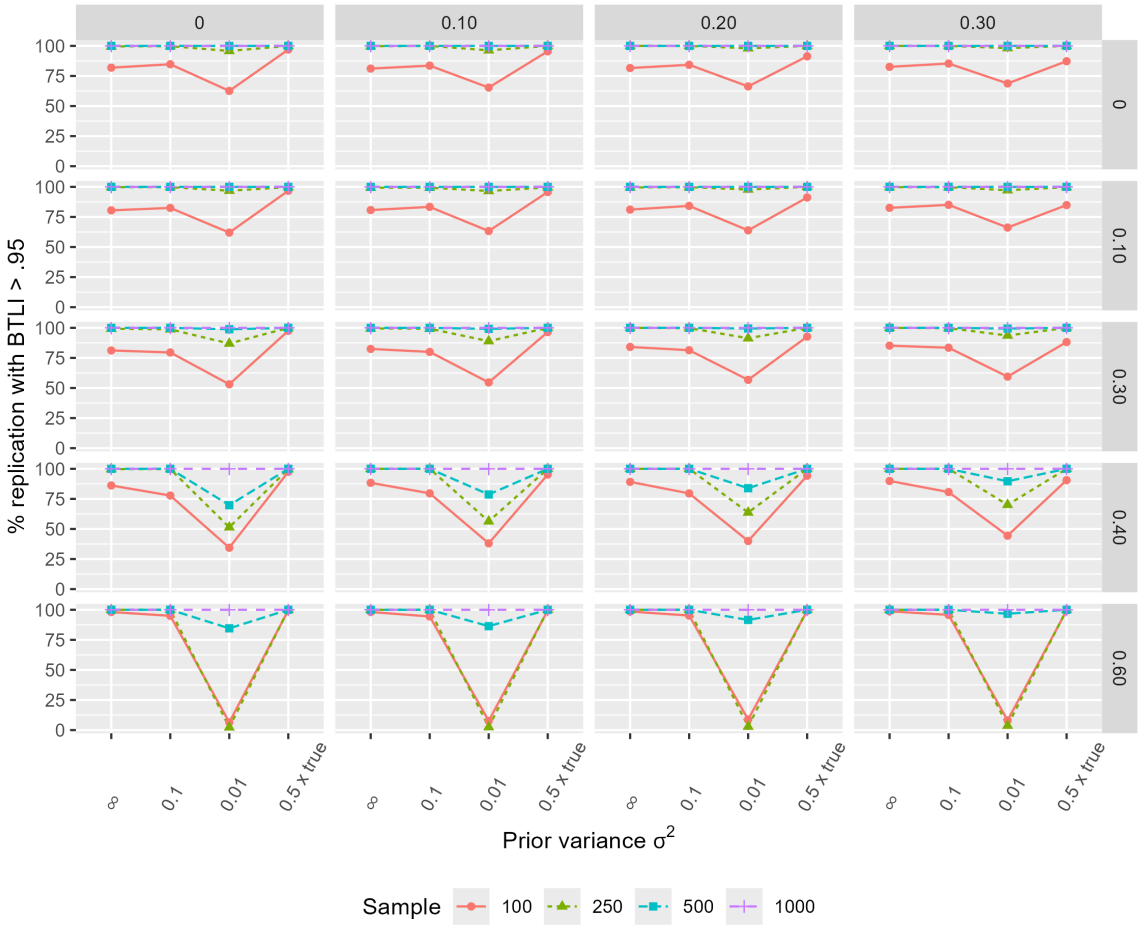
The BTLI selection rates. The y-axis shows the percentage of replications where BTLI was above the 0.95 cutoff, indicating good fit. The x-axis represents the fitted models. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes.



;

Figure 28

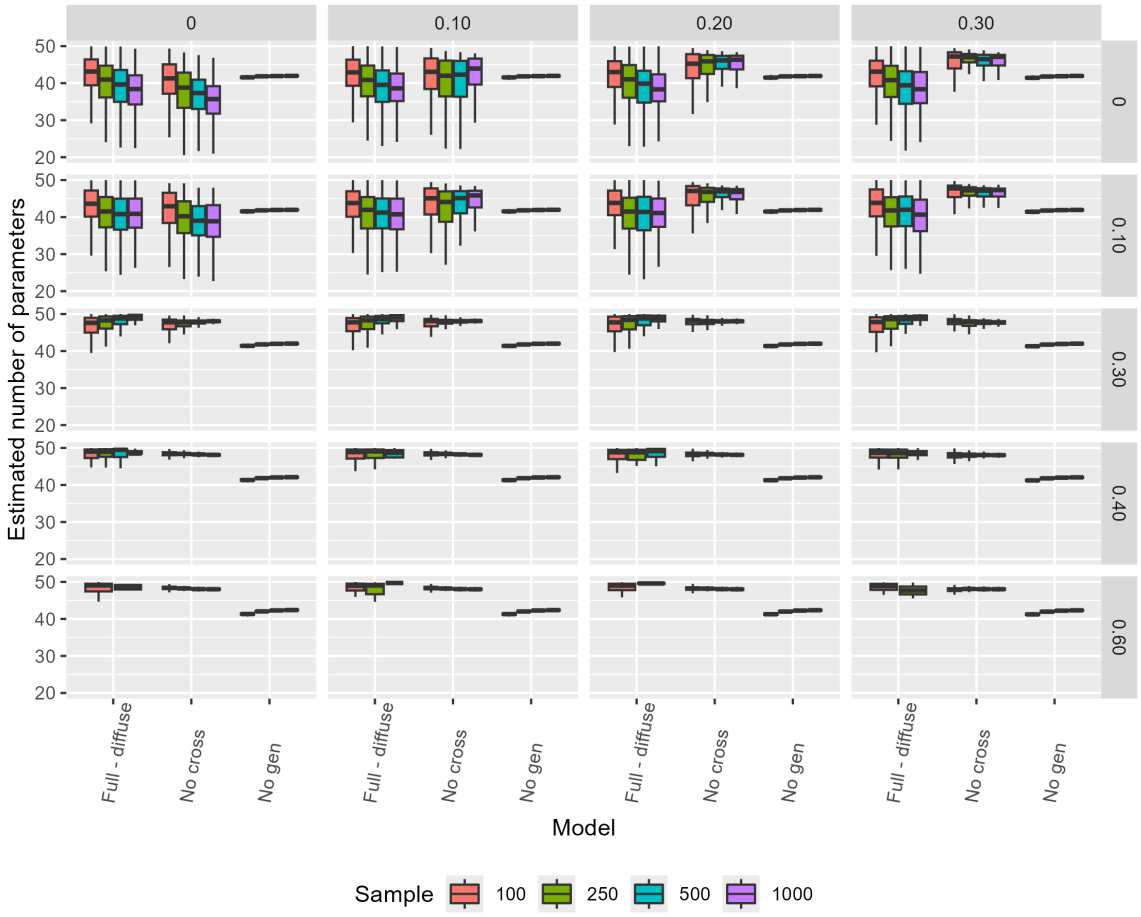
CI-based BTLI selection rates. The y-axis shows the proportion of replications. The x-axis represents fitted models. The colors of bars represent the decision about the fit: good, bad or inconclusive. Rows and first-order columns represent population general factor loadings and cross-loadings, respectively. Second-order columns represent different sample sizes.



;

Figure 29

Prior sensitivity of BTLI. The y-axis shows the percentage of replications where BTLI was above the 0.95 cutoff, indicating good fit. The x-axis represents variance of priors placed on general factor loadings. The last prior is centered at the true value, other priors are centered at zero. The rows and columns represent populations general factor loadings and cross-loadings, respectively. Line type and line color represent sample sizes.



;

Figure 30

Estimated number of parameters pD. The x-axis represents three fitted models. The rows and columns represent populations general factor loadings and cross-loadings, respectively. The box color represent sample size.