Springer Proceedings in Mathematics & Statistics

Marie Wiberg · Dylan Molenaar ·
Jorge González · Jee-Seon Kim ·
Heungsun Hwang *Editors*

Quantitative Psychology

The 87th Annual Meeting of the Psychometric Society, Bologna, Italy, 2022

Springer

Marie Wiberg • Dylan Molenaar • Jorge González •
Jee-Seon Kim • Heungsun Hwang
Editors

Quantitative Psychology

The 87th Annual Meeting of the Psychometric Society, Bologna, Italy, 2022

Springer

Editors
Marie Wiberg
Department of Statistics, Umeå School of Business, Economics & Statistics
Umeå University
Umeå, Sweden

Dylan Molenaar
Department of Psychology
University of Amsterdam
Amsterdam, The Netherlands

Jorge González
Facultad de Matemáticas, and Millennium Nucleus on Intergenerational Mobility: From Modelling to Policy (MOVI)
Pontificia Universidad Católica
Santiago, Chile

Jee-Seon Kim
Department of Educational Psychology
University of Wisconsin-Madison
Madison, WI, USA

Heungsun Hwang
Department of Psychology
McGill University
Montreal, QC, Canada

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-031-27780-1 ISBN 978-3-031-27781-8 (eBook)
https://doi.org/10.1007/978-3-031-27781-8

Mathematics Subject Classification: 62-06, 62P15

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Limited Utility of Small-Variance Priors to Detect Local Misspecification in Bayesian Structural Equation Models

Terrence D. Jorgensen and Mauricio Garnier-Villarreal

Abstract In a highly influential paper on current practice in Bayesian structural equation modeling (BSEM), Muthén and Asparouhov (Psychol Methods 17:313–335, 2012) proposed using small-variance priors to constrain non-target parameters to be close to (rather than exactly) zero, with the "side product" (p. 313) that the posterior distributions of such nontarget parameters could be used analogous to modification indices. This chapter presents 2 simulation studies of their utility, in the context of (a) constraining cross-loadings to be nearly zero and (b) constraining factor loadings and intercepts to be equivalent across groups or occasions. The first study reinforced earlier findings that small-variance priors can prevent detecting important misspecifications (i.e., global-fit indices indicate better fit as priors become less restrictive). In contrast, these local indicators have greater power to detect invalid constraints when priors are less restrictive. Study 2 revealed similar patterns in the context of detecting invalid equality constraints and showed limited utility of small-variance priors over modification indices under maximum-likelihood estimation. Our advice is to evaluate global fit in BSEM without small-variance priors, and only when hypothesized models are rejected, utilize small-variance priors to search for clues about possible respecification. We recommend exploring other tools for local-fit evaluation in BSEM, which might detect misspecifications without introducing additional complications of small-variance priors (e.g., propagation of bias).

Keywords Bayesian · Structural equation modeling · Measurement invariance · Differential item functioning · Modification indices

T. D. Jorgensen (✉)
Universiteit van Amsterdam, Amsterdam, the Netherlands
e-mail: t.d.jorgensen@uva.nl

M. Garnier-Villarreal
Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 422, https://doi.org/10.1007/978-3-031-27781-8_8

1 Introduction

Bayesian structural equation modeling (BSEM) has recently received substantial attention within psychology and the social sciences as an increasingly viable alternative to traditional frequentist SEM techniques, such as maximum likelihood (ML) estimation. Several tools are available to evaluate global (mis)fit of a BSEM, such as posterior predictive model checking (PPMC; Gelman et al., 1996), for which a posterior predictive *p* value (PPP) can be calculated that is analogous to the *p* value of a SEM's χ^2 statistic, which tests the null hypothesis (H_0) that a SEM perfectly represents the true data-generating process. Approximate global fit of a BSEM can be evaluated using SRMR (Levy, 2011) or χ^2-based fit indices analogous to those under maximum likelihood estimation (MLE), on the condition that the BSEM uses uninformative priors during Markov chain Monte Carlo (MCMC) estimation (Garnier-Villarreal & Jorgensen, 2020). Although PPP or fit indices may indicate poor model fit, they cannot provide clues about the specific source(s) of misspecification.

In a highly influential paper, Muthén and Asparouhov (2012) proposed using small-variance priors to constrain non-target parameters to be close to zero, as a less-restrictive alternative to fixing such parameters to exactly zero. The Bayesian credible intervals (BCI; interval estimates analogous to confidence intervals of frequentist estimators) for nontarget parameter estimates (constrained to be small) can be used to indicate local sources of misspecification. They suggested that "[the sensitivity of nontarget parameters] be used in line with modification indices [in MLE] to free parameters for which the credibility interval does not cover zero" (Muthén & Asparouhov, 2012, pp. 316–317), noting the advantage over modification indices in that BCIs for all parameters can be obtained simultaneously, preventing the problem of sequentially modifying one parameter at a time under ML estimation. The goal of this paper is to evaluate their proposal in the context of (a) cross-loadings in single-group SEM and (b) equality constraints on loadings (i.e., measurement equivalence) using Monte Carlo simulations.

2 Study 1: Priors for Approximately-Zero Constraints

This study was part of an investigation of PPP's frequency properties, so the Method details correspond to those published by Jorgensen et al. (2019). We focus only on normal-data conditions here because patterns of results for ordinal data were largely similar, although power decreased with fewer categories.

2.1 Method

Using the MONTECARLO command in M*plus* (version 6.11 for Linux; Muthén & Muthén, 2012), we simulated a two-factor CFA with three indicators per factor. In each of the four population models, factors were standard normal ($\mu = 0$, $\sigma = 1$), with a factor correlation $\psi_{21} = 0.25$, factor loadings $\lambda = 0.7$, indicator intercepts $= 0$, and indicator residual variances $\theta = 0.51$; thus, indicators had unit variance. To vary levels of misspecification of the analysis model, the third indicator of the first factor was specified to have a cross-loading on the second factor (λ_{32}) in the population. The magnitude of λ_{32} was 0.0, 0.2, 0.5, or 0.7 in the population, but was constrained to be close to zero in the analysis model using informative priors (see next paragraph). For ease of interpretation, we refer to $\lambda_{32} = 0.2$ as minor misspecification (using $\alpha = .05$, the ML χ^2 test has 80% power when $N > 500$, RMSEA = 0.06, SRMR = 0.03, CFI = 0.98), $\lambda_{32} = 0.5$ as severe misspecification (80% power when $N > 150$, RMSEA = 0.12, SRMR = 0.07, CFI = 0.92), and $\lambda_{32} = 0.7$ as very severe misspecification (80% power when $N > 100$, RMSEA = 0.14, SRMR = 0.07, CFI = 0.89).

In the analysis model, we specified noninformative priors for all target parameters (primary loadings, residual variances, and the factor covariance) using M*plus* defaults—for example, factor loadings ~ $N(\mu = 0, \sigma^2 =$ "infinity"). For all cross-loadings, we specified normally distributed priors with four levels of informative variance, chosen to correspond approximately with the prior belief in a 95% probability that the cross-loadings are within approximately ±0.01, ±0.10, ±0.20, or ± 0.30 of zero (i.e., $\sigma = 0.005$, 0.05, 0.10, and 0.15, or equivalently $\sigma^2 = 0.000025$, 0.0025, 0.01, and 0.0225). In each condition, sample sizes of $N = 50$–500 were drawn in increments of 25, along with an asymptotic condition of $N = 1000$. We generated 200 samples from each of 320 conditions (20 sample sizes, four levels of CL, and four prior variances) with normally distributed indicators.

We kept 100,000 iterations from the MCMC chains after thinning every 100th iteration. Over 99% of models converged on a proper solution, yielding 63,480 (out of 64,000) PPP values for analysis. Convergence was evaluated using Gelman and Rubin's (1992) potential scale reduction factor ("R-hat" < 1.1). Convergence in each condition was at least 98% except when sample size was small ($N < 100$) and CL was large ($\lambda_{32} > 0.5$). The smallest convergence rate was 82% ($N = 50$, $\lambda_{32} = 0.7$). Nonconverged solutions were omitted from Results. Nontarget cross-loadings were considered significantly different from 0 when their 95% BCI excluded 0.

2.2 Results

Whereas the power to reject an inappropriate model increased as prior variance decreased (negative association) when using PPP as an indicator of global misfit (see Jorgensen et al., 2019, for details), the power to detect local sources of misfit

(here, the neglected parameter λ_{32}) increased as prior variance increased (a positive association). Figure 1 depicts how often λ_{32} was detected as significantly different from 0. As would be expected, λ_{32} was never estimated to be significantly greater than 0 when it was in fact 0 in the population, and was very seldom estimated to be significant when it was only 0.2 in the population. As might also be expected, using the most restrictive priors—which yielded the greatest power of PPP to detect misspecification—λ_{32} was never estimated to be significantly greater than 0. Power was only adequate when the neglected parameter was severe ($\lambda_{32} = 0.5$ or 0.7). When the prior variance was reasonably informative (95% CI within ± 0.10 of 0), adequate power ($\geq 80\%$) to detect the neglected cross-loading (λ_{32}) was found for $N > 400$, and for $N > 300$ when priors were less informative (95% CI within ± 0.20 or $\pm$ 0.30 of 0).

We were also interested in the degree to which the neglected cross-loading would affect other parameters estimates in the model. Related cross-loadings (first and second indicators of the first factor, which did not cross-load onto the second factor in the population) were sometimes detected to be significantly different from 0 (although less frequently than the actual neglected cross-loading), and the factor correlation grew increasingly biased. Investigating the average parameter estimates for normal-data conditions in Table 1 (collapsed across sample size, which had no

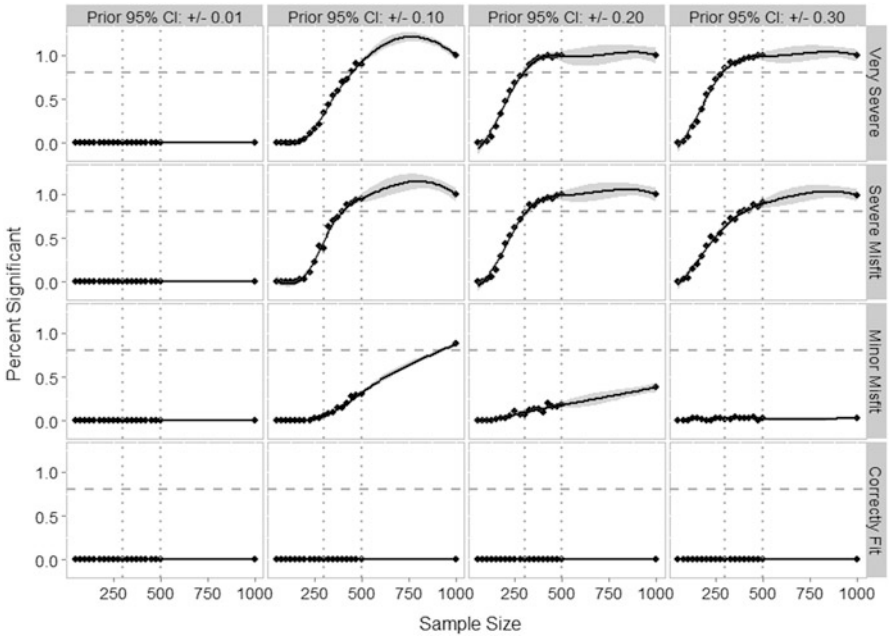

Fig. 1 Rejection rates for neglected cross-loading (λ_{32}) as a function of sample size, plotted separately across conditions of varying priors and magnitude of neglected cross-loading (λ_{32}). Dashed horizontal line provided for reference at 80% power, and dotted vertical lines at $N = 300$ and 500 provided for reference when judging sample sizes necessary for adequate power

Table 1 Effect of neglected cross-loading (λ_{32}) on estimates of related cross-loadings and factor correlation

Prior 95% CI	Population λ_{32}	$\hat{\lambda}_{32}$	$\hat{\lambda}_{12}$	$\hat{\lambda}_{22}$	$\hat{\psi}_{21}$
±0.01	0.0	0.000	0.000	0.000	0.247
	0.2	0.001	−0.001	−0.001	0.332
	0.5	0.001	−0.001	−0.001	0.448
	0.7	0.001	−0.001	−0.001	0.518
±0.10	0.0	0.000	0.001	0.001	0.244
	0.2	0.050	−0.027	−0.026	0.330
	0.5	0.088	−0.058	−0.057	0.438
	0.7	0.087	−0.068	−0.068	0.499
±0.20	0.0	0.002	0.002	0.001	0.243
	0.2	0.091	−0.045	−0.046	0.327
	0.5	0.184	−0.111	−0.110	0.433
	0.7	0.211	−0.145	−0.145	0.492
±0.30	0.0	0.002	0.002	0.001	0.242
	0.2	0.109	−0.055	−0.053	0.329
	0.5	0.234	−0.135	−0.135	0.433
	0.7	0.285	−0.187	−0.186	0.491

effect on the point estimates) reveals that as λ_{32} increased, (a) the average estimates of related cross-loadings decreased, although with less magnitude than the neglected λ_{32}, and (b) the average estimate of the factor correlation became greater than its true value (0.25). Note that although there would seldom be any indication (i.e., low power) that the pattern is significant when the neglected cross-loading is only minor ($\lambda_{32} = 0.2$), such a small neglected parameter estimate still results in a unacceptably biased factor correlation (relative bias = [0.33–0.25] / 0.25 = 0.32), according to Hoogland and Boomsma's (1998) criterion (< 0.05).

To verify that such bias would also occur using MLE, we simulated a single large sample (N = 10,000) from the population with $\lambda_{32} = 0.7$, and fit a model to that data in which all cross-loadings were fixed to 0. This yielded the same negative bias in the related cross-loadings and the same positive bias in the factor correlation. Modification indices indicated that fit would be significantly improved by freeing not only the true omitted cross-loading, but also by freeing other cross-loadings and residual correlations. Freeing only the true omitted cross-loading eliminated bias in any estimates.

3 Study 2: Priors for Approximate Equality Constraints

This is a subset of unpublished results from a dissertation project (Jorgensen, 2015). When evaluating measurement equivalence across contexts (e.g., different populations or occasions), small-variance priors can be specified for parameters that

represent differential item/indicator functioning (DIF), allowing for approximate rather than exact invariance. Although priors can now be easily specified for functions of parameters in M*plus* (Muthén & Muthén, 2012) and `blavaan`[1] (Merkle & Rosseel, 2018), this study was manually programmed in 2014 using Stan (Carpenter et al., 2017).

3.1 Method

Figure 2 represents the data-generating 1-factor SEMs for Study 2. In addition to type of invariance (groups vs. occasions), we manipulated total $N = 200$, 300, 400, 600, or 800 (balanced group sizes) and priors for DIF parameters ~ $N(\mu = 0$, $\sigma = 0.05$ or 0.10; i.e., 95% probability that $\Delta\lambda$ or $\Delta\tau$ fell within ± 0.10 or within ± 0.20, respectively). For longitudinal SEM, the autocorrelation for common and unique factors are indicated by the dashed line representing the factor correlation.

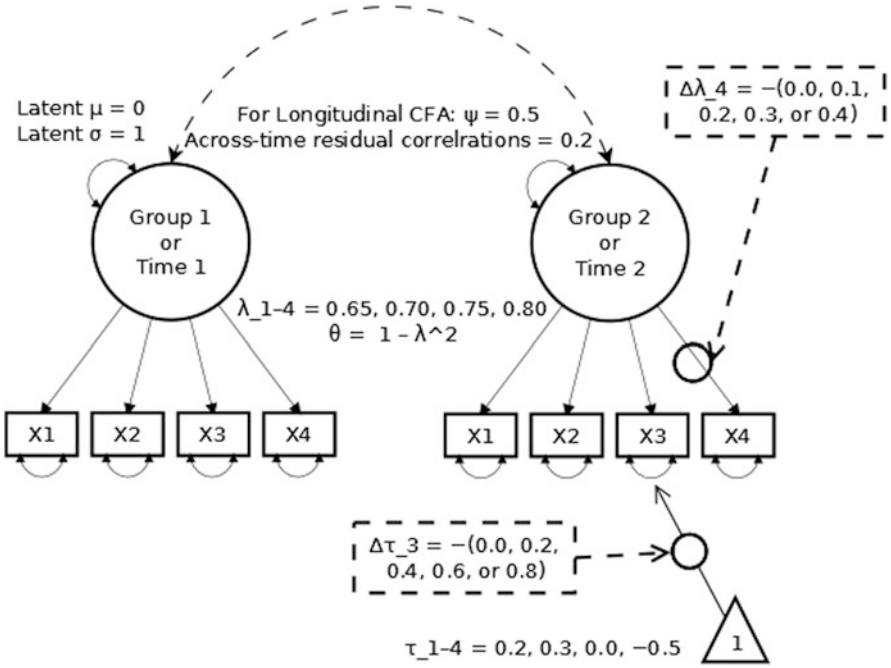

Fig. 2 Population model(s) for data generation in Study 2. Solid lines represent population characteristics that are constant across all conditions, whereas dashed lines represent varying conditions described in the dashed textboxes

[1] See https://ecmerkle.github.io/blavaan/articles/invariance.html for example syntax.

One of 5 effect sizes for DIF (see dashed boxes) were simultaneously added to Item 4's loadings and Item 3's intercept, yielding 5 DIF conditions. We generated 500 samples from each population.

Whereas M*plus* (Muthén & Muthén, 2012) uses Gibbs sampling, Stan (Carpenter et al., 2017) uses a modified Hamiltonian Monte Carlo algorithm called the no U-turn sampler (NUTS), which has efficiency advantages over Gibbs sampling. After 1000 burn-in iterations on each of three chains, we saved 1000 post-burn-in samples per chain. We fit models representing approximate metric invariance (Model 1), approximate full scalar invariance (Model 2b), and approximate partial scalar invariance (Model 2f). Model 2b represents a "backward" specification search, in which DIF is tested by releasing constraints from a fully restricted model. Model 2f represents a "forward" specification search, which proceeds from the least constrained configural model and applies more restrictive constraints. This is not strictly necessary in BSEM because all DIF parameters can be evaluated simultaneously, but it allows comparison of Muthén and Asparouhov's (2012) proposed approach to the traditional use of modification indices in ML estimation (using `lavaan`; Rosseel, 2012).

3.2 Results

Convergence was nearly 100% for Models 2b and 2f, but nonconvergence of Model 1 increased with N, particularly with less informative priors. When the prior $\sigma = 0.05$, convergence dropped from 100% when $N = 200$ to 50% when $N = 800$. When the prior $\sigma = 0.10$, convergence dropped from 100% when $N = 200$ to 25% when $N = 800$. In all conditions, there were > 100 converged results, and collapsing across conditions with little impact (e.g., no substantial differences between multigroup and longitudinal models) increased the Monte Carlo sample sizes used to draw conclusions.

Similar to Study 1, using small-variance priors on substantially nonzero parameters induced bias in other DIF parameters (which were truly zero in the population). Estimated DIF for DIF-free items appeared to counterbalance the invalidly constrained (truly nonzero) DIF parameter, and the effect was stronger in larger samples (see Fig. 3 for estimated DIF in intercepts). Furthermore, estimated parameters (posterior means) of latent variables were systematically biased by using small-variance priors on substantially nonzero parameters. In this case, latent means were biased more negatively as $\Delta\tau_4$ increased, more so in Model 2f (which correctly allowed for DIF in λ_4) than Model 2b (which invalidly constrained λ_4). Surprisingly, less restrictive priors exacerbated the situation: allowing the true DIF to be more negative did not alleviate the truly DIF-free estimates, which were also more positive. Similar results were found for DIF in factor loadings and how that biases estimated latent variance in the second group/occasion (see Jorgensen, 2015, Part III). Patterns were similar but more extreme using ML estimation, when Models

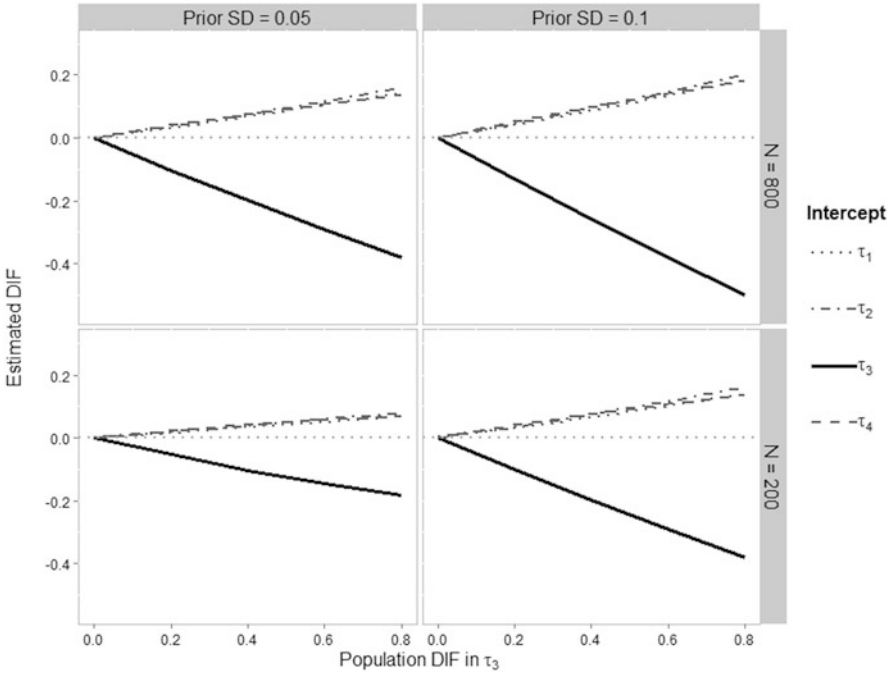

Fig. 3 Average posterior mean of $\Delta\tau$s by DIF, prior σ, and *N*, with separate lines per $\Delta\tau$

1 and 2(b and f) represented exact rather than approximate metric and (full or partial) scalar invariance.

The practical impact of these biased estimates can be reflected by rates at which the H_0 of invariance was rejected. Figure 4 compares Type I error rates (averaged across non-DIF parameters) between ML modification indices (grey lines) and 95% BCIs (black lines). While Type I error rates fluctuated around the nominal 5% for modification indices, the BCIs had near-zero error rates across conditions. As typically happens when Type I error rates are higher, power for modification indices was also somewhat higher in some conditions (see Fig. 5).

4 Discussion

The use of parameter estimates constrained by small-variance priors as a Bayesian analog to ML modification indices (Muthén & Asparouhov, 2012) seems to have some limited potential. Their power to detect DIF is often similar to (sometimes lower than) modification indices, but they have lower Type I error rates. However, small-variance priors continue to propagate bias throughout the model, just as invalid exactly-zero constraints do in ML estimation. So it may not be advisable

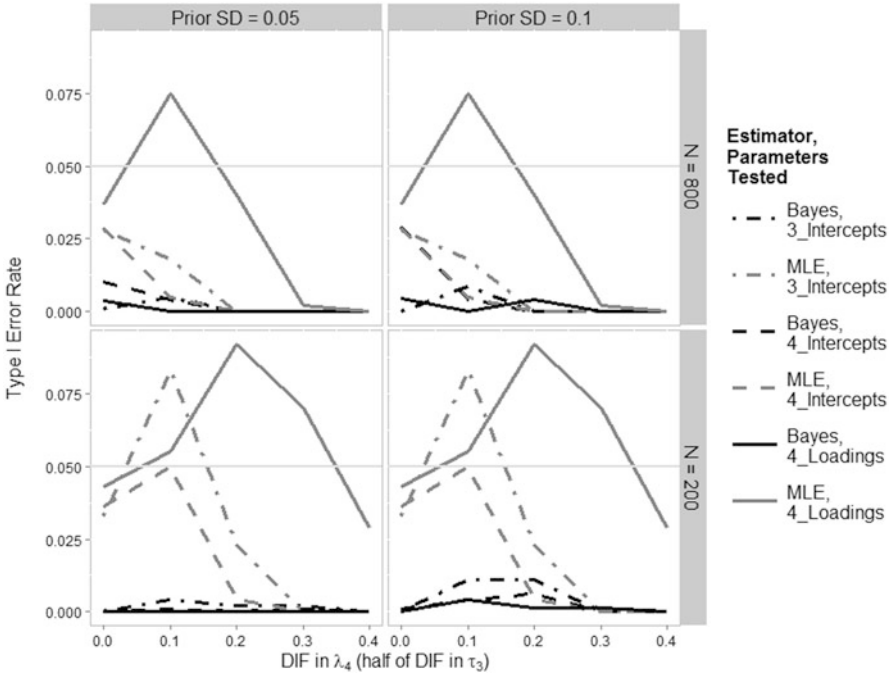

Fig. 4 Type I error rates by DIF, prior σ, and *N*, with separate lines per estimator and model

to use small-variance priors, even following a sensitivity analysis to choose their precision (e.g., Asparouhov et al., 2015). At the very least, relying on parameter estimates constrained by small-variance priors for clues about necessary model modifications (which Muthén & Asparouhov, 2012, indicated was a "side product of the proposed approach", p. 313) does not imply that models with small-variance priors should be used for inference.

As Muthén and Asparouhov (2012) assert in their subtitle, small-variance priors for nontarget parameters are intended to provide researchers with a more flexible representation of [their] substantive theory. But because PPP appears insensitive to minor misspecification (Jorgensen et al., 2019), nontarget parameters could potentially be fixed to zero without PPP indicating poor model fit. When misspecification is too severe to be ignorable, PPP would have even greater power to reject the model if priors for nontarget parameters were excluded altogether (i.e., nontarget parameters fixed to zero). When a SEM without small-variance priors indicates poor (exact or even approximate) fit, small-variance priors for nontarget parameters could then be added to help detect the local source of misfit; however, the priors should be only weakly informative to increase the probability that they indicate a neglected parameter should be "freed" and (contrary to Muthén and Asparouhov's advice) freed one parameter at a time rather than considering all parameters simultaneously. Future research should explore the possibility of

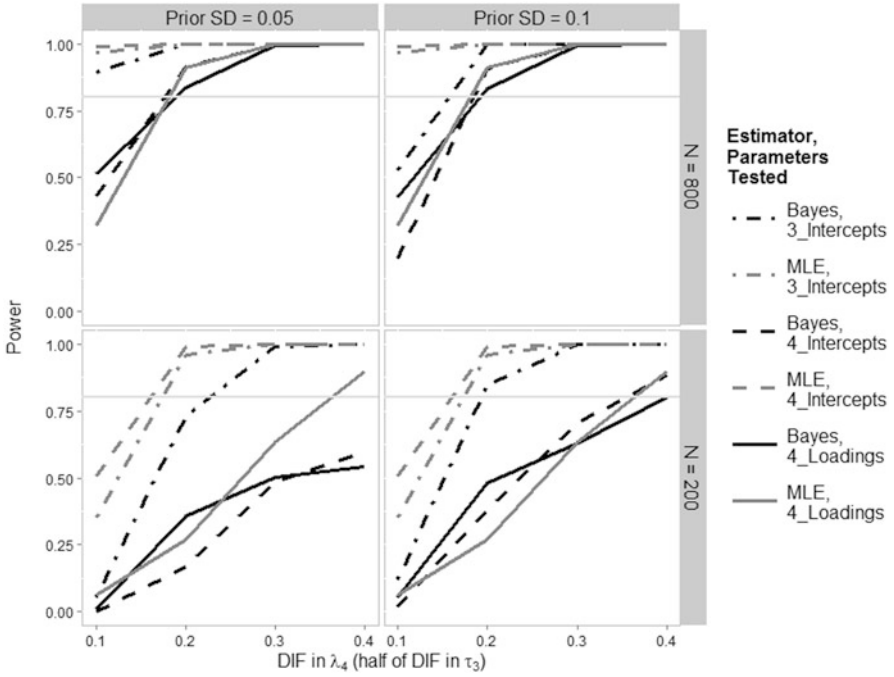

Fig. 5 Power by DIF, prior σ, and *N*, with separate lines per estimator and model

developing more reliable tools to detect local sources of misspecification in BSEM (i.e., sensitive to misspecification without propagating errors throughout the model), perhaps using a PPMC framework to investigate score-based statistics, analogous to actual modification indices in ML estimation.

References

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. https://doi.org/10.1177/0149206315591075

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods, 25*(1), 46–70. https://doi.org/10.1037/met0000224

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733–807. https://doi.org/10.1.1.142.9951

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods & Research, 26*, 329–367. https://doi.org/10.1177/0049124198026003003

Jorgensen, T. D. (2015). *Selecting an optimal measurement model and detecting differential item functioning using Bayesian confirmatory factor analysis* [Doctoral dissertation, University of Kansas]. https://doi.org/10.13140/RG.2.2.14104.03841.

Jorgensen, T. D., Garnier-Villarreal, M., Pornprasertmanit, S., & Lee, J. (2019). Small-variance priors can prevent detecting important misspecifications in Bayesian confirmatory factor analysis. In M. Wiberg, S. A. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology: The 83rd annual meeting of the psychometric society, New York, 2018* (pp. 255–263). Springer. https://doi.org/10.1007/978-3-030-01310-3_23

Levy, R. (2011). Bayesian data–model fit assessment for structural equation modeling. *Structural Equation Modeling, 18*(4), 663–685. https://doi.org/10.1080/10705511.2011.607723

Merkle, E. C., & Rosseel, Y. (2018). `Blavaan`: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software, 85*(4), 1–30. https://doi.org/10.18637/jss.v085.i04

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335. https://doi.org/10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Author.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02