Small-Variance Priors Can Prevent Detecting Important Misspecifications in Bayesian Confirmatory Factor Analysis

Terrence D. Jorgensen, Mauricio Garnier-Villarreal, Sunthud Pornprasermanit and Jaehoon Lee

Abstract We simulated Bayesian CFA models to investigate the power of PPP to detect model misspecification by manipulating sample size, strongly and weakly informative priors for nontarget parameters, degree of misspecification, and whether data were generated and analyzed as normal or ordinal. Rejection rates indicate that PPP lacks power to reject an inappropriate model unless priors are unrealistically restrictive (essentially equivalent to fixing nontarget parameters to zero) and both sample size and misspecification are quite large. We suggest researchers evaluate global fit without priors for nontarget parameters, then search for neglected parameters if PPP indicates poor fit.

Keywords Structural equation modeling · Confirmatory factor analysis · Bayesian inference · Model evaluation · Model modification

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 (MacCallum, Edwards, & Cai, 2012; Muthén & Asparouhov, 2012; Rindskopf, 2012), such as maximum likelihood (ML) estimation. Bayesian estimates of model parameters are based on a sampling plausible parameter values from the posterior distribution of the model parameters, which is estimated using Markov chain Monte Carlo (MCMC) estimation (see Muthén & Asparouhov, 2012, for details). Programs available for analyzing a BSEM include Amos (Arbuckle, 2012), M*plus* (Muthén & Muthén, 2012), and more recently the R (R Core Team, 2018) package blavaan (Merkle & Rosseel, 2018), which utilizes the

T. D. Jorgensen (✉)
University of Amsterdam, Nieuwe Achtergracht 127, 1018WS Amsterdam, The Netherlands
e-mail: T.D.Jorgensen@uva.nl

M. Garnier-Villarreal
Marquette University, 2340 N. Cramer St., Unit 515, Milwaukee, WI 53211, USA
e-mail: mauricio.garnier@marquette.edu

S. Pornprasermanit · J. Lee
Texas Tech University, 41071, Lubbock, TX 79409, USA
e-mail: jaehoon.lee@ttu.edu

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings in Mathematics & Statistics 265, https://doi.org/10.1007/978-3-030-01310-3_23

more general Bayesian modeling software JAGS (Plummer, 2003) and Stan (Carpenter et al., 2017).

An important step in fitting BSEMs (or models in general) is to investigate how well the hypothesized model can reproduce the observed data, for which posterior predictive model checking (PPMC; Gelman, Meng, & Stern, 1996) was developed (see Levy, 2011, for a review of PPMC specifically for BSEM). PPMC is not a test statistic per se, but it can be based on a test statistic, such as the traditional χ^2 test of exact fit derived from the ML discrepancy function (although any discrepancy measure of interest can be used, such as the SRMR fit index; Levy, 2011). PPMC uses the samples from the joint posterior distribution to check data–model correspondence by comparing observed data to expected data. For the sampled parameters at one iteration of a Markov chain (after the burn-in iterations), a random sample of N observations is simulated from the population with those parameters. The simulated sample's sufficient statistics (means and covariances) are calculated and compared (using the ML discrepancy function) with the sufficient statistics implied by the model parameters drawn from the posterior distribution at that iteration; likewise, the same discrepancy is calculated comparing the sufficient statistics of the observed data to the model-implied sufficient statistics at that iteration. This results in a discrepancy measure both for the observed data and for the simulated data. A score of 1 is assigned if the observed data have less discrepancy (i.e., fit better) than the simulated data; otherwise, a score of 0 is assigned—this score can be considered as a Bernoulli random variable. These binary numbers are averaged across all iterations sampled from the posterior distribution. The average is referred to as the posterior predictive p value (PPP).

Naturally, the model should fit well to the simulated data at every iteration because simulated data are drawn from those population parameters. If the hypothesized model is an appropriate approximation of the population process from which the real data was sampled, then the model should fit the real data often, too. On average, an appropriate model will fit the real data better than the simulated data about as often as the other way around, so the mean of this Bernoulli random variable is 50%, the expected value of PPP when the target model is approximately correct. The probability will decrease as the appropriateness of the hypothesized model decreases in its ability to explain the phenomena under investigation. That is, if the model is grossly inappropriate, the model will continue to fit well to the simulated data drawn from those population parameters, but it will rarely fit well to the real data, so the expected value of PPP will approach 0%.

There is no theoretical cutoff for how low PPP must be to indicate unignorable misfit, nor is there a consensus about how applied users should interpret PPP (e.g., treat it like a frequentist p value and compare it to an alpha level, or use as a fit index). Muthén and Asparouhov (2012) recently suggested after their initial simulations that the traditional approach of using "posterior predictive p values of 0.10, 0.05, or 0.01 appears reasonable" (p. 315), with the caveat that further investigations were needed to establish how these methods behave in practice with various models and data. Depaoli (2012) already began to address that gap in a recent Monte Carlo simulation study of PPP values in the context of growth mixture modeling, and found

that PPP and graphical PPMC were only likely to identify extremely misspecified growth mixture models. We use a Monte Carlo simulation study to investigate (a) the sensitivity of PPP to varying levels of misspecification in confirmatory factor analysis (CFA) models, as well as (b) how consistently Muthén and Asparouhov's guidelines would apply across varying samples sizes and different informative priors.

1 Method

1.1 Continuous (Standard Normal) Indicators

Using the MONTECARLO command in M*plus* (version 6.11 for Linux; Muthén & Muthén, 2002), 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 = 0.25, factor loadings = 0.70, indicator intercepts = 0.0, and indicator residual variances = 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. For ease of interpretation, we refer to $\lambda_{32} = 0.2$ as minor misspecification (using $\alpha = 0.05$, 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 drew 200 samples from each of 320 conditions (20 sample sizes, four levels of CL, and four prior variances) with normally distributed indicators.

Following Muthén and Asparouhov's (2012) advice, 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 in each condition was at least 98% except when sample size was small ($N \leq 100$) and CL was large ($\lambda_{32} \geq 0.5$). The smallest convergence rate was 82% ($N = 50$, $\lambda_{32} = 0.7$).

1.2 Categorical Indicators

Because behavioral data are so often measured using discrete scales rather than truly continuous data, we also simulated binary and ordinal data. Rhemtulla, Brosseau-Liard, and Savalei's (2012) simulation results suggest that when ordinal variables have at least five categories, robust estimation methods for nonnormal data provide similar conclusions as estimation for categorical data. For few categories, we were interested in how PPP and constrained nontarget parameters would be affected by the data distribution. Thus, we manipulated the same conditions as for normal indicators described above, but we additionally varied the number of categories (from two to five) and how the data were analyzed (appropriately as ordinal or inappropriately as normal). Thresholds were not manipulated, but were chosen to mimic a unimodal symmetric distribution: 0 for binary; $\pm$ 0.8 for three categories; -1, 0, and 1 for four categories; and -1.6, -0.8, 0.8, and 1.6 for five categories. Whereas 100% of the models converged when the indicators were analyzed as ordinal, when the indicators were analyzed as normal, convergence rates were 94.78, 97.34, 98.45, and 98.13% for 2, 3, 4, and 5 categories, respectively.

2 Results

For each model, we investigate the sampling variability of PPP across conditions, and calculate power and Type I error[1] rates using traditional cutoff values ($\alpha = 0.10$, 0.05, or 0.01) for PPP to identify "significant" misfit.

2.1 Sampling Variability of PPP

Table 1 shows the effect sizes for each model under investigation. Using Cohen's (1988) criteria for interpreting the size of η^2 (negligible $< 0.01 <$ small $< 0.06 <$ moderate $< 0.14 <$ large), N had a negligible effect on PPP when normal or ordinal data were analyzed assuming normality, but N explained 4% of variance in PPP when data were analyzed as ordinal. PPP values were largely influenced by the magnitude of the neglected cross-loading (CL), but much more so for normal data ($\eta^2 = 34.2\%$) than for categorical data analyzed as normal ($\eta^2 = 17.1\%$) or as ordinal ($\eta^2 = 18\%$). The magnitude of prior variance for estimating nontarget CLs had a large effect on PPP when normal ($\eta^2 = 22.1\%$) or categorical ($\eta^2 = 20.9\%$) data were analyzed as normal, but only a moderate effect when categorical data were analyzed as ordinal ($\eta^2 = 7\%$). When categorical data were analyzed as ordinal, the number of cate-

[1] We use the term "Type I error" when referring to any model that does not omit a substantial parameter, although in the categorical data conditions, the model contains another type of misspecification (incorrect likelihood) when analyzed as though the data were normally distributed.

Table 1 Proportions of variance (η^2) of PPP explained by Monte Carlo factors

	How data were generated (and analyzed)		
	Normal data (as normal)	Categorical data (as normal)	Categorical data (as ordinal)
N	0.002	0.001	0.040
Prior variance	**0.221**	**0.209**	**0.070**
Misfit	**0.343**	**0.171**	**0.180**
Number of categories (#CAT)		0.019	**0.072**
N × Prior	0.014	0.024	0.018
N × Misfit	0.002	0.003	0.021
Prior × Misfit	**0.103**	**0.096**	0.023
N × #CAT		0.000	0.001
Prior × #CAT		0.010	0.008
Misfit × #CAT		0.013	0.017
N × Prior × Misfit	0.009	0.010	0.012
N × Prior × #CAT		0.001	0.001
N × Misfit × #CAT		0.000	0.001
Prior × Misfit × #CAT		0.008	0.003
N × Prior × Misfit × #CAT		0.002	0.001

Note Medium and large effect sizes using Cohen's (1988) criteria (i.e., effect explains at least 6% of variance) are in bold font. Data generated as continuous did not have varying numbers of categories, so cells involving the #CAT effect are blank

gories also had a moderate effect of PPP ($\eta^2 = 7.2\%$). The only substantial two-way interactions were found between prior variance and magnitude of neglected CL, for normal data ($\eta^2 = 10.3\%$) and for categorical data analyzed as normal ($\eta^2 = 9.6\%$). All other interactions effects were negligible or small ($\eta^2 \leq 4\%$). The effect of *N* in most conditions appears approximately linear in Fig. 1 (and Figs. A1–A8 provided in the online[2] supplemental materials), so we treated *N* as a continuous[3] covariate to calculate η^2.

Figure 1 (online supplemental material) reveals the nature of the interaction between magnitude of prior variances and of the neglected parameter (λ_{32}) in normal-data conditions. When $\lambda_{32} = 0$ (no misspecification), the average PPP value is consistent with its expected value of 50%, regardless of the magnitude of prior variance. As the magnitude of the neglected population parameter increases, the average PPP decreases, but PPP shows more sensitivity to misspecification when prior variances are restrictive than when only weakly informative. Figure 1 plots PPP values only

[2]The online supplemental materials can be retrieved at the following URL: https://osf.io/buhvg/.

[3]Treating *N* as a categorical factor showed no substantial difference in the effect sizes.

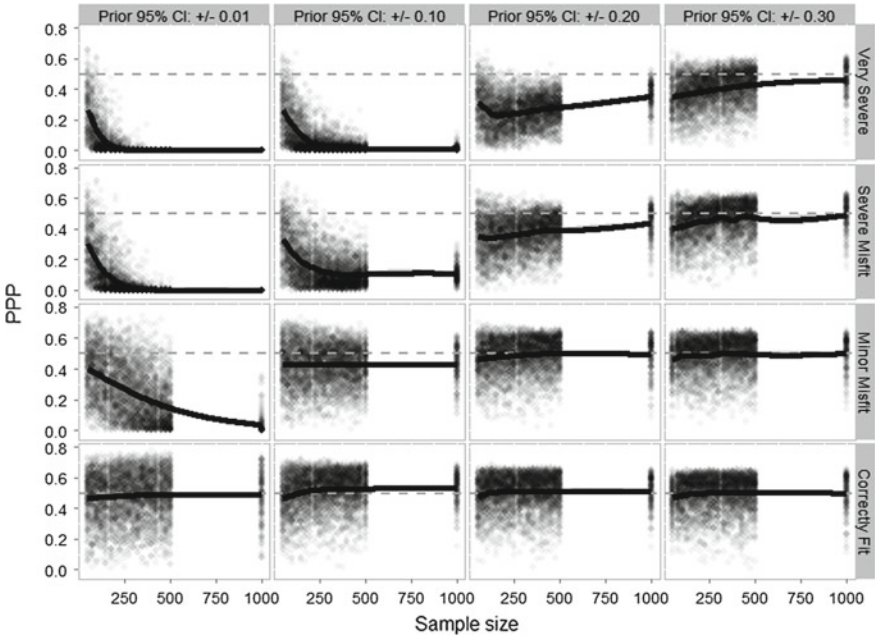

Fig. 1 Variability of PPP as a function of sample size, plotted separately for each condition of prior variance and magnitude of neglected cross-loading (λ_{32}). There is a smoothed regression line indicating how the mean PPP changes, and the horizontal dashed line in each cell at PPP = 50% refers to the expected value under the null hypothesis of no misfit

for the normal-data conditions, but this two-way interaction is characterized by similar patterns in all categorical-data conditions, with the exception that PPP appears less sensitive for two-category data, especially when analyze as ordinal (see Figures A1–A4 in the online appendix). PPP is also less variable when categorical data have fewer categories and are analyzed as ordinal, so PPP's distribution resembles even less the expected uniform [0, 1] distribution of traditional *p* values (Hoijtink & van de Schoot, 2018).

2.2 *Detecting Misfit*

Figure 2 (and online supplemental material) plots the rejection rates for normal-data models against *N*, with separate panels for each magnitude of prior variance and misspecification. We used three different criteria for rejecting a model due to lack of fit: PPP < 0.10, 0.05, or 0.01, to evaluate Muthén and Asparouhov's (2012, p. 315) suggestion. Rejection rates in the bottom row of Fig. 2 represent Type I error rates because the model is not misspecified (i.e., $\lambda_{32} = 0$), whereas rejection rates in the top two rows represent power (i.e., when $\lambda_{32} = 0.5$ or 0.7). Rejection rates in the

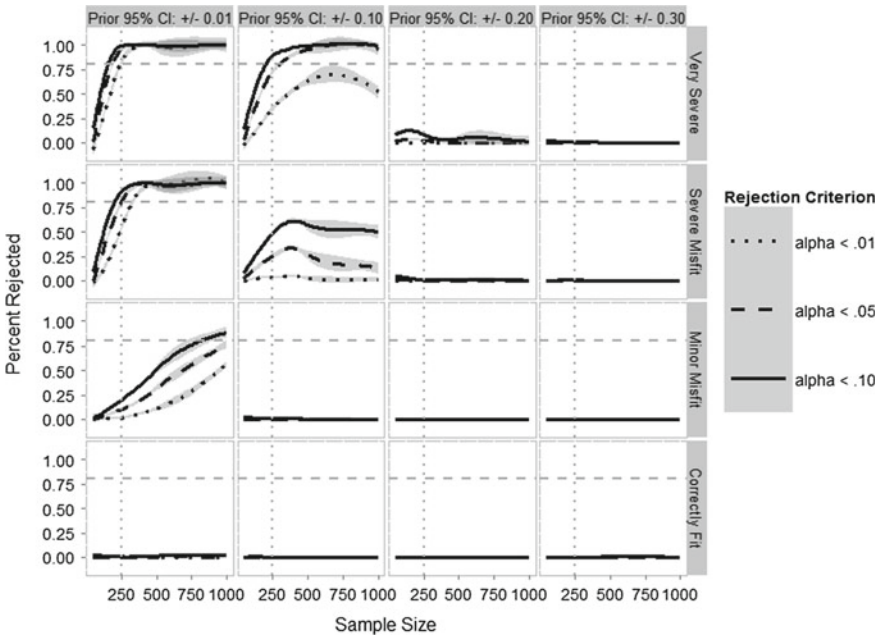

Fig. 2 Rejection rates as a function of sample size, plotted separately across conditions of varying priors and magnitude of neglected cross-loading (λ_{32}). The dashed horizontal line at 80% represents acceptable power, and the dotted vertical line at $N = 250$ is for convenience when judging sample sizes required for adequate power

row labeled "Minor Misfit" correspond to $\lambda_{32} = 0.2$ and could be classified as Type I errors or power, depending on the whether the analyst wishes to test exact fit or close fit (Browne & Cudeck, 1992). That is, a researcher might consider a neglected cross-loading of 0.2 to be of little substantive consequence, so the analysis model would be considered to correspond closely enough to the population model that it should not be rejected.

Consistent with prior research showing that PPP is more conservative than nominal error rates (Gelman et al., 1996; Levy, 2011), the Type I error rate is near zero in almost every condition, much lower than nominal levels using any of the three rejection criteria. However, power rarely exceeds 80% (a commonly preferred minimum) unless the neglected parameter is quite large or the prior variance is quite small (or both). Depending on the rejection criterion, power exceeds 80% when $N > 200$–300 when using the most restrictive priors. This suggests that with sufficient sample size, researchers could only be confident about detecting misfit by specifying priors so informative that their 95% confidence limits are approximately ± 0.01—so strongly informative that the model bears little practical distinction from one in which no informative priors are specified for cross-loadings. Using more realistic informative priors with 95% confidence limits approximately ± 0.1, power only exceeded 80% for the most severe level of misspecification ($\lambda_{32} = 0.7$). Perhaps most noteworthy,

power was close to zero to detect severe misspecification ($\lambda_{32} = 0.5$) at any N when using Muthén and Asparouhov's (2012) suggested priors (95% confidence limits approximately ± 0.2). Similar results were found for categorical data (see Figs. A9–A16), although power was even lower when data with fewer categories were analyzed as ordinal (e.g., power remained nearly zero in all binary conditions; see Fig. A9).

3 Discussion

The assessment of fit and detection of misspecification in SEM is no less important in a Bayesian context than in a traditional frequentist paradigm, but tools currently available in BSEM are few, and their behavior is largely unknown. In the conditions we investigated, PPP lacks power unless $N > 200$–300, misspecification is severe, and priors for nontarget parameters are highly (even unrealistically) restrictive. This implies that informative priors for nontarget parameters should be chosen very carefully. Asparouhov, Muthén, and Morin (2015) suggested a data-driven sensitivity analysis to choose priors that balanced detecting substantial misfit and allowing negligible misfit. More recently, Cain and Zhang (in press) found larger Type I error rates with simulated 3-factor models than we did with 2-factor models, and they recommended different PPP criteria for models with different numbers of indicators. This calls into question whether any uniform cutoffs can be expected to perform consistently across conditions with different data and model characteristics. Because power to detect substantial misspecification only appears adequate when priors are so restrictive that they are nearly equivalent to fixing the nontarget parameters to zero, we suggest researchers simply evaluate global fit without priors for nontarget parameters, then search for neglected parameters only if PPP indicates poor fit. But because even minor misspecification can be detected with great power in asymptotically large samples (Hoofs, van de Schoot, Jansen, & Kant, 2018), the development of complementary fit indices for evaluating BSEMs (similar to those used in SEM) would be a welcome and useful addition to the researcher's toolbox. The BRMSEA (Hoofs et al., 2018) is the first such attempt, and it appears promising, but further development is needed.

References

Arbuckle, J. L. (2012). *IBM SPSS Amos 21 user's guide*. Chicago, IL: IBM.

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

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21*(2), 230–258. https://doi.org/10.1177/0049124192021002005.

Cain, M. K., & Zhang, Z. (in press). Fit for a Bayesian: An evaluation of PPP and DIC for structural equation modeling. *Structural Equation Modeling*. https://doi.org/10.1080/10705511.2018.1490648.

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

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

Depaoli, S. (2012). The ability for posterior predictive checking to identify model misspecification in Bayesian growth mixture modeling. *Structural Equation Modeling, 19*(4), 534–560. https://doi.org/10.1080/10705511.2012.713251.

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.

Hoijtink, H., & van de Schoot, R. (2018). Testing small variance priors using prior-posterior predictive *p* values. *Psychological Methods, 23*(3), 561–569. https://doi.org/10.1037/met0000131.

Hoofs, H., van de Schoot, R., Jansen, N. W., & Kant, I. (2018). Evaluating model fit in Bayesian confirmatory factor analysis with large samples: Simulation study introducing the BRM-SEA. *Educational and Psychological Measurement, 78*(4), 537–568. https://doi.org/10.1177/0013164417709314.

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.

MacCallum, R. C., Edwards, M. C., & Cai, L. (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychological Methods, 17*(3), 340–345. https://doi.org/10.1037/a0027131.

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. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8.

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

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Retrieved from http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/.

R Core Team. (2018). R: A language and environment for statistical computing (version 3.5.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from the comprehensive R archive network (CRAN): https://www.R-project.org/.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354–373. https://doi.org/10.1037/a0029315.

Rindskopf, D. (2012). Next steps in Bayesian structural equation models: Comments on, variations of, and extensions to Muthen and Asparouhov (2012). *Psychological Methods, 17*(3), 336–339. https://doi.org/10.1037/a0027130.