Psychological Methods
2012, Vol. 17, No. 3, 313–335

1082-989X/12/$12.00 DOI: 10.1037/a0026802

Bayesian Structural Equation Modeling: A More Flexible Representation of Substantive Theory

Bengt Muthén and Tihomir Asparouhov
Muthén & Muthén, Los Angeles, California

This article proposes a new approach to factor analysis and structural equation modeling using Bayesian analysis. The new approach replaces parameter specifications of exact zeros with approximate zeros based on informative, small-variance priors. It is argued that this produces an analysis that better reflects substantive theories. The proposed Bayesian approach is particularly beneficial in applications where parameters are added to a conventional model such that a nonidentified model is obtained if maximum-likelihood estimation is applied. This approach is useful for measurement aspects of latent variable modeling, such as with confirmatory factor analysis, and the measurement part of structural equation modeling. Two application areas are studied, cross-loadings and residual correlations in confirmatory factor analysis. An example using a full structural equation model is also presented, showing an efficient way to find model misspecification. The approach encompasses 3 elements: model testing using posterior predictive checking, model estimation, and model modification. Monte Carlo simulations and real data are analyzed using Mplus. The real-data analyses use data from Holzinger and Swineford's (1939) classic mental abilities study, Big Five personality factor data from a British survey, and science achievement data from the National Educational Longitudinal Study of 1988.

Keywords: confirmatory factor analysis, informative priors, posterior predictive checking, Markov chain Monte Carlo method

This article proposes a new approach to factor analysis and structural equation modeling (SEM) using Bayesian analysis (i.e., Bayesian structural equation modeling [BSEM]). It is argued that current analyses using maximum-likelihood (ML) and likelihood-ratio chi-square testing apply unnecessarily strict models to represent hypotheses derived from substantive theory. This often leads to rejection of the model (see, e.g., Marsh et al., 2009) and a series of model modifications that may capitalize on chance (see, e.g., MacCallum, Roznowski, & Necowitz, 1992). The hypotheses are reflected in parameters fixed at zero. Examples include zero cross-loadings and zero residual correlations in factor analysis.

The new approach is intended to produce an analysis that better reflects substantive theories. It does so by replacing the parameter specification of exact zeros with approximate zeros. The new approach uses Bayesian analysis to specify informative priors for such parameters. In key applications, freeing these parameters in a conventional analysis, the model would not be identified. The Bayesian analysis, however, identifies the model by substantively driven small-variance priors. Model testing is carried out using posterior predictive checking, which is found to be less sensitive than likelihood-ratio chi-square testing to ignorable degrees of model misspecification. A side product of the proposed approach is information to modify the model in line with the use of modification indices in ML analysis. ML modification indices inform about model improvement when a single parameter is freed and can lead to a long series of modifications. In contrast, the proposed approach informs about model modification when all parameters are freed and does so in a single-step analysis.

The next section presents a brief overview of the Bayesian analysis framework that is used. Following this, two studies are presented that illustrate the new approach. Each study consists of a real-data example showing the problem, the proposed Bayesian solution for the real-data problem, and simulations showing how well the method works. Study 1 considers factor analysis where cross-loadings make simple structure confirmatory factor analysis (CFA) inadequate. As an example, we reanalyze Holzinger and Swineford's (1939) classic mental abilities data, where a simple structure does not fit well by ML CFA standards. Study 2 considers residual correlations in factor analysis, which make a factor model inadequate. As an example, the Big Five factor model is analyzed using an instrument administered in the British Household Panel Survey, where the hypothesized five-factor pattern is not well recovered by ML CFA or exploratory factor analysis (EFA) because of many minor factors. The two studies are followed by an SEM example that illustrates the use of priors for both structural and measurement parameters. All analyses are carried out by Bayesian analysis in Mplus (Muthén & Muthén, 1998–2010), and scripts are available at www.statmodel.com. The article ends with a discussion of related approaches, extensions, reflections on analysis strategies, and caveats.

Bengt Muthén and Tihomir Asparouhov, Muthén & Muthén, Los Angeles, California.

Bengt Muthén and Tihomir Asparouhov are both part of the Mplus development team.

Correspondence concerning this article should be addressed to Bengt Muthén, Muthén & Muthén, 3463 Stoner Avenue, Los Angeles, CA 90066. E-mail: bmuthen@ucla.edu

Bayesian Analysis

There are many books on Bayesian analysis, and most are quite technical. Gelman, Carlin, Stern, and Rubin (2004) provided a good general statistical description, whereas Kruschke (2010) and Lynch (2010) gave somewhat more introductory accounts. Press (2003) discussed Bayesian factor analysis. Lee (2007) gave a discussion from an SEM perspective. Schafer (1997) gave a statistical discussion from a missing data and multiple imputation perspective, whereas Enders (2010) gave an applied discussion of these same topics. Statistical overview articles include Gelfand, Hills, Racine-Poon, and Smith (1990) and Casella and George (1992). Overview articles of an applied nature and with a latent variable focus include Scheines, Hoijtink, and Boomsma (1999); Rupp, Dey, and Zumbo (2004); Yuan and MacKinnon (2009); and Kaplan and Depaoli (in press).

Bayesian analysis is firmly established in mainstream statistics, and its popularity is growing. Part of the reason for the increased use of Bayesian analysis is the success of new computational algorithms referred to as Markov chain Monte Carlo (MCMC) methods.

Outside of statistics, however, applications of Bayesian analysis lag behind. One possible reason is that Bayesian analysis is perceived as difficult to do, requiring complex statistical specifications, such as those used in the flexible but technically oriented general Bayes program WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003). These observations were the background for developing Bayesian analysis in Mplus (Muthén & Muthén, 1998–2010). In Mplus, simple analysis specifications with convenient defaults allow easy access to a rich set of analysis possibilities. For a technical description of the Mplus implementation, see Asparouhov and Muthén (2010b).

Four key points motivate taking an interest in Bayesian analysis:

1. More can be learned about parameter estimates and model fit.
2. Better small-sample performance can be obtained and large-sample theory is not needed.
3. Analyses can be made less computationally demanding.
4. New types of models can be analyzed.

Point 1 is illustrated by parameter estimates that do not have a normal distribution. An example is an indirect effect $a \times b$ in a mediation model (MacKinnon, 2008). ML gives a parameter estimate and its standard error and assumes that the distribution of the parameter estimate is normal on the basis of asymptotic (large-sample) theory. In contrast, Bayes does not rely on large-sample theory and provides the whole distribution, referred to as the posterior distribution, not assuming that it is normal. The ML confidence interval $Estimate \pm 1.96 \times SE$ assumes a symmetric distribution, whereas the Bayesian credibility interval based on the percentiles of the posterior allows for a strongly skewed distribution. Bayesian exploration of model fit can be done in a flexible way using posterior predictive checking (see, e.g., Gelman et al., 2004, chapter 6; Gelman, Meng, Stern, & Rubin, 1996; Lee, 2007, chapter 5; Scheines et al., 1999). Any suitable test statistics for the observed data can be compared with statistics based on simulated data obtained through draws of parameter values from the posterior distribution, avoiding statistical assumptions about the distribution of the test statistics.

Point 2 is illustrated by better Bayesian small-sample performance for factor analyses prone to Heywood cases and better performance when a small number of clusters are analyzed in multilevel models. This, however, requires a judicious choice of prior. For examples, see Asparouhov and Muthén (2010a).

Point 3 may be of interest for an analyst who is hesitant to move from ML estimation to Bayesian estimation. Many models are computationally cumbersome or impossible using ML, such as with categorical outcomes and many latent variables resulting in many dimensions of numerical integration. Such an analyst may view the Bayesian analysis simply as a computational tool for getting estimates that are analogous to what would have been obtained by ML had it been feasible. This is obtained with diffuse priors, in which case ML and Bayesian results are expected to be close in large samples (W. J. Browne & Draper, 2006, p. 505).

Point 4 is exemplified by models with a very large number of parameters or where ML does not provide a natural approach. Examples of the former include image analysis (see, e.g., Green, 1996), and examples of the latter include random change-point analysis (see, e.g., Dominicus, Ripatti, Pedersen, & Palmgren, 2008). The Bayesian SEM approach proposed in this article is a further example of the new type of models that can be analyzed.

Bayesian Estimation

Bayesian analysis is a large topic, and this article does not attempt to give a full, pedagogical introduction to the topic. Instead, the article gives a brief outline of the necessary main points, referring the reader to the literature just mentioned for further studies. The emphasis is instead on how Bayesian analysis can be used for SEM.

Frequentist analysis (e.g., ML) and Bayesian analysis differ by the former viewing parameters as constants and the latter viewing them as variables. Bayesian analysis uses the term *prior* to refer to the parameter distribution. Priors can be diffuse (noninformative) or informative. Information about priors can be built up from a sequence of formulating hypotheses from theory, carrying out pilot studies, and revising hypotheses. An example that is discussed in detail later on was drawn from Holzinger and Swineford's (1939) classic factor analysis study. Using two new samples of subjects, a set of well-known tests was used to measure factors that had been derived from several previous factor analyses. A factor loading matrix was hypothesized where each item had nonzero loadings on only the factor or factors it was hypothesized to measure and had zero loadings (cross-loadings) on other factors. Although Holzinger and Swineford did not invoke Bayesian analysis, their careful groundwork could have been used in the analysis of the two new samples to specify informative priors centered around zero for the cross-loadings.

ML finds estimates by maximizing a likelihood computed for the data. In Bayesian analysis, data inform about a parameter and modify the prior into a posterior that gives the Bayesian estimate. This is illustrated in Figure 1, which shows distributions for a prior and a posterior for a parameter, together with the likelihood. The likelihood can be thought of as the distribution of the data given a parameter value. In Figure 1, the major portion of the prior

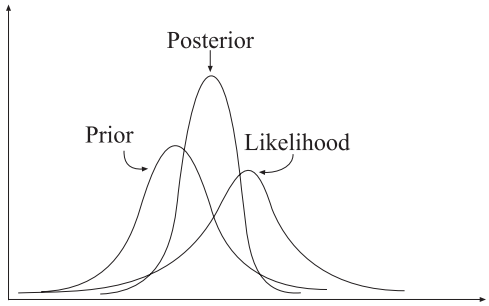

Figure 1. Prior, likelihood, and posterior for a parameter.

distribution has a lower parameter value than that at the peak of the likelihood. The posterior is obtained as a compromise between the prior and the likelihood.

Priors can be noninformative or informative. A noninformative prior, also called a diffuse prior, can, for example, have a uniform distribution or have a normal distribution with a large variance. A large variance reflects large uncertainty in the parameter value. With a large prior variance, the likelihood contributes relatively more information to the formation of the posterior, and the estimate is closer to an ML estimate.

Bayes's Theorem

Formally, the formation of a posterior draws on Bayes's theorem. Consider the probabilities of events *A* and *B, P*(*A*) and *P*(*B*). By probability theory, the joint event *A* and *B* can be expressed in terms of conditional and marginal probabilities:

$$P(A, B) = P(A|B)\,P(B) = P(B|A)\,P(A). \quad (1)$$

Dividing by *P*(*A*) it follows that

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}, \quad (2)$$

which is Bayes's theorem. Applied to modeling, let data take the role of *A* and the parameter values take the role of *B*. The posterior can then be expressed symbolically as

$$posterior = parameters|data \quad (3)$$

$$= \frac{data|parameters \times parameters}{data} \quad (4)$$

$$= \frac{likelihood \times prior}{data} \quad (5)$$

$$\propto likelihood \times prior, \quad (6)$$

where $\propto$ means proportional to, not including the data portion of Equation 5.

The prior distribution is the key element of Bayesian analysis. Priors reflect prior beliefs in likely parameter values before collecting new data. These beliefs may come from substantive theory and previous studies of similar populations. The priors modify the likelihood to obtain the posterior distribution. The Bayesian estimates are obtained as means, modes, or medians of their posterior distributions. The posterior distribution is obtained with MCMC algorithms. MCMC is briefly outlined in the Appendix and is not discussed here. The reader is instead referred to the Bayesian literature. The Appendix also discusses determination of convergence of the MCMC process. For a technical description of the Mplus Bayesian implementation, see Asparouhov and Muthén (2010b).

Model Fit

Model fit assessment is possible using posterior predictive checking proposed by Gelman et al. (1996). With continuous outcomes, posterior predictive checking as implemented in Mplus builds on the standard likelihood-ratio chi-square statistic in mean- and covariance-structure modeling. This posterior predictive checking procedure is described in Scheines et al. (1999) and Asparouhov and Muthén (2010a, 2010b) and is briefly reviewed here. Gelman et al. (2004) presented a more general discussion of posterior predictive checking, not tied to likelihood-ratio chi-square.

A posterior predictive *p* value of model fit can be obtained with a fit statistic, *f,* based on the usual likelihood-ratio chi-square test of an H_0 model against an unrestricted H_1 model. A low posterior predictive *p* value indicates poor fit. Let $f(Y, X, \pi_i)$ be computed for the data *Y, X* using the parameter values at MCMC iteration *i.* Here, *X* denotes covariates that are conditioned on in the analysis. At iteration *i,* generate a new data set Y_i^* of synthetic or replicated data of the same sample size as the original data. In this generation, the parameter values at iteration *i* are used. For these replicated data, the fit statistic $f(Y_i^*, X, \pi_i)$ is computed. This data generation and fit statistic computation is repeated over the *n* iterations, after which posterior predictive *p* value is approximated by the proportion of iterations where

$$f(Y, X, \pi_i) < f(Y_i^*, X, \pi_i). \quad (7)$$

In the Mplus implementation (Asparouhov & Muthén, 2010b) posterior predictive *p* value is computed using every 10th iteration among the iterations used to describe the posterior distribution of parameters. A 95% confidence interval is produced for the difference in the *f* statistic for the real and replicated data. A positive lower limit is in line with a low posterior predictive *p* value and indicates poor fit. An excellent-fitting model is expected to have a posterior predictive *p* value around .5 and an *f* statistic difference of zero falling close to the middle of the confidence interval.

It should be noted that the posterior predictive *p* value does not behave like a *p* value for a chi-square test of model fit (see also Hjort, Dahl, & Steinbakk, 2006). The Type I error is not 5% for a correct model. There is not a theory for how low the posterior predictive *p* value can be before the model is significantly ill fitting at a certain level. In this sense, posterior predictive *p* value is more akin to an SEM fit index rather than a chi-square test. Empirical experience with different models and data has to be established for posterior predictive *p* value, and some simulation studies are presented here. From these simulations and further ones in Asparouhov and Muthén (2010a), however, using posterior predictive *p* value values of .10, .05, or .01 appears reasonable. This warrants further investigations, however. In the simulations to follow, a posterior predictive *p* value of .05 is used.

BSEM: A More Flexible SEM Approach

A new approach to SEM based on Bayesian analysis, BSEM, is described next. It is intended to produce an analysis that better reflects the researcher's theories and prior beliefs. It does so by systematically using informative priors for parameters that should not be freely estimated according to the researcher's theories and prior beliefs. In a frequentist analysis, such parameters are typically fixed at zero or are constrained to be equal to other parameters. In key applications, freeing these parameters would, in fact, produce a nonidentified model. The Bayesian analysis, however, identifies the model by substantively driven small-variance priors. It should be recognized that BSEM refers to the specific Bayesian approach proposed here of using informative, small-variance priors to reflect the researcher's theories and prior beliefs. Typically, this would be combined with the use of noninformative priors for parameters that would not be restricted in a corresponding ML analysis. For example, major loadings would have a normal prior with a very large variance.

The BSEM approach of using informative priors is applicable to any constrained parameter in an SEM. This article focuses on parameters in the measurement part, but restrictions in the structural part are also considered. Two types of measurement model features are considered, cross-loadings in CFA and residual correlations in CFA. Further examples are considered in the Conclusion section.

Informative Priors for Cross-Loadings in CFA

An analyst who is used to frequentist methods, such as ML, may at first feel uncomfortable specifying informative priors. It is argued here, however, that a user of CFA is in a sense already engaged in specifying such priors. Consider the CFA model for an observed p-dimensional vector $\boldsymbol{y}_i$ of factor indicators for individual i,

$$\begin{aligned} \boldsymbol{y}_i &= \nu + \boldsymbol{\Lambda}\eta_i + \varepsilon_i, \\ E(\boldsymbol{y}_i) &= \nu + \boldsymbol{\Lambda}\alpha, \\ V(\boldsymbol{y}_i) &= \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \end{aligned} \tag{8}$$

where ν is an intercept vector, $\boldsymbol{\Lambda}$ is a loading matrix, η_i is an m-dimensional factor vector, ε_i is a residual vector, α is a factor mean vector, $\boldsymbol{\Psi}$ is a factor covariance matrix, and $\boldsymbol{\Theta}$ is a residual covariance matrix. Here, ε and η are assumed normally distributed and uncorrelated.

Drawing on substantive theory, zero cross-loadings in $\boldsymbol{\Lambda}$ are specified for the factor indicators that are hypothesized to not be influenced by certain factors. Table 1 shows one such example drawing on Holzinger and Swineford's (1939) study with 19 tests hypothesized to measure four domains. Here, an X denotes a free loading to be estimated, and 0 denotes a fixed, zero loading. This example is further described and analyzed in a later section.

An exact zero loading can be viewed as a prior distribution that has mean zero and variance zero. A prior that probably more accurately reflects substantive theory uses a mean of zero and a normal distribution with small variance. Figure 2 shows an example where a loading $\lambda \sim N(0, 0.01)$ so that 95% of the loading variation is between -0.2 and 0.2. Using standardized factor indicators and factors, a loading of 0.2 is considered a small loading, implying that this prior essentially says that the cross-loading is close to zero, but not exactly zero. The prior is strongly informative, but it is not assumed that the parameter is literally zero.

Table 1
Holzinger and Swineford's (1939) Hypothesized Four Domains Measured by 19 Tests: Factor Loading Pattern

Test	Spatial	Verbal	Speed	Memory
Visual perception	X	0	0	0
Cubes	X	0	0	0
Paper form board	X	0	0	0
Flags	X	0	0	0
General information	0	X	0	0
Paragraph comprehension	0	X	0	0
Sentence completion	0	X	0	0
Word classification	0	X	0	0
Word meaning	0	X	0	0
Addition	0	0	X	0
Code	0	0	X	0
Counting groups of dots	0	0	X	0
Straight and curved capitals	0	0	X	0
Word recognition	0	0	0	X
Number recognition	0	0	0	X
Figure recognition	0	0	0	X
Object–number	0	0	0	X
Number–figure	0	0	0	X
Figure–word	0	0	0	X

In frequentist analysis, freeing all cross-loadings in a CFA model such as Table 1 leads to a nonidentified model because the m^2 restrictions, where m is the number of factors, necessary to eliminate indeterminacies are not present (see, e.g., Hayashi & Marcoulides, 2006). Using small-variance priors for all cross-loadings, however, brings information into the analysis that avoids the nonidentification problem. The choice of variance for the prior should correspond to the researcher's theories and prior beliefs. As stated earlier, the variance of 0.01 produces a prior where 95% lies between -0.2 and 0.2. Other choices are shown in Table 2.

A smaller variance may not let cross-loadings escape sufficiently from their zero prior mean, producing a worse posterior predictive p value. A larger variance may let a cross-loading have too large a probability of having a substantial value. For example, a variance of 0.08 corresponds to 95% lying between -0.55 and 0.55, which on a standardized variable scale approaches a major loading size. When the variance is increased, the prior contributes less information so that the model gets closer to being nonidentified, which eventually causes nonconvergence of the MCMC algorithm. It should be noted that the prior variance should be determined in relation to the scale of the observed and latent variables. A prior variance of 0.01 corresponds to small loadings for variables with unit variance, but it corresponds to a smaller loading for an observed variable with variance larger than one. This means that, for convenience, observed variables may be brought to a common scale either by multiplying them by constants or by standardizing if the model is scale free.

BSEM has an additional advantage. It produces posterior distributions for cross-loadings that can be used in line with modification indices to free parameters for which the credibility interval

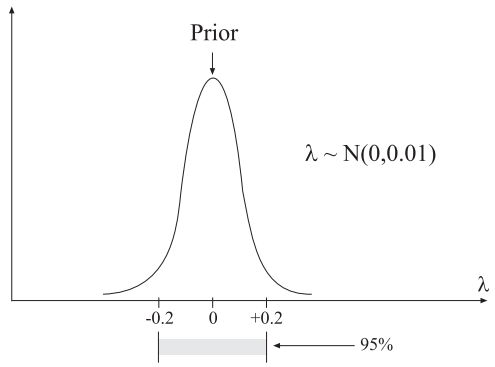

Figure 2. Informative prior for a factor loading parameter.

does not cover zero. Modification indices pertain to freeing only one parameter at a time, and a long sequence of model modification is often needed, running the risk of capitalizing on chance (see, e.g., MacCallum et al., 1992). In contrast, the small-variance prior approach provides information on model modification that considers the fixed parameters jointly in a single analysis. The relative benefits of the Bayesian approach to modifying the model compared with the use of modification indices with ML need further study, however.

Informative Priors for Residual Correlations in CFA

An analogous idea can also be used to study residual correlations among factor indicators. In Equation 8, the residual covariance matrix $\mathbf{\Theta}$ is commonly assumed to be diagonal. Some residuals may, however, be correlated because of the omission of several minor factors. It is difficult to foresee which residuals should be covaried, and freeing all of them leads to a nonidentified model in the conventional ML framework. BSEM provides a possible approach to this problem. The corresponding MCMC algorithms are, however, more complex and draw on Bayesian theories that cannot be extensively described here because of lack of space.

Instead of assuming a diagonal residual covariance matrix, a more realistic covariance structure model may be expressed as

$$V(\mathbf{y}_i) = \mathbf{\Lambda\Psi\Lambda}' + \Omega + \mathbf{\Theta}^*, \quad (9)$$

where $\mathbf{\Omega}$ is a covariance matrix for the minor factors, not assumed to be diagonal, and $\mathbf{\Theta}^*$ is a diagonal covariance matrix. Here, a freely estimated $\mathbf{\Omega}$ is not separately identified from $\mathbf{\Lambda}$, $\mathbf{\Psi}$, and $\mathbf{\Theta}^*$. In Bayesian analysis, however, $\mathbf{\Omega}$ can be given an informative prior using the inverse-Wishart distribution so that the posterior distribution can be obtained. The inverse-Wishart is a standard prior distribution for covariance matrices in Bayesian analysis. Although difficult to give an intuitive description, aspects of the inverse-Wishart are described in the Appendix. The reader is referred to the further readings suggested in the Appendix. In this way, the diagonal and off-diagonal elements of $\mathbf{\Omega}$ can be restricted to small values. This implies that the residual covariance matrix $\mathbf{\Omega} + \mathbf{\Theta}^*$ contains residual covariances that are allowed to deviate to a small extent from zero means. Sufficiently stringent priors for the off-diagonal elements are needed so that the essential correlations are channeled through $\mathbf{\Lambda}$, $\mathbf{\Psi}$, $\mathbf{\Lambda}'$. The sums on the diagonal of $\mathbf{\Omega} + \mathbf{\Theta}^*$ produce the residual variances.

The BSEM approach for residual covariances outlined in connection with Equation 9 is referred to as Method 1. A more direct method, Method 2, applies an inverse-Wishart prior directly on $\mathbf{\Theta}$ in Equation 8. This approach has been discussed in Press (2003, chapter 15). One advantage of Method 1 over Method 2 is that the prior for the total residual variance is not tied to the prior of the residual covariances because the residual covariance has two components that have different priors. Method 2, however, is simpler to carry out. A disadvantage of both Method 1 and Method 2 is that particular residual covariance elements cannot be given their own priors. For example, an analysis may show that some residual covariances should be freely estimated with noninformative priors because they have 95% credibility intervals that do not cover zero. To this aim, Method 3 makes it possible to specify element-specific normal priors for the residual covariances. Mplus allows two different algorithms for Method 3, a random-walk algorithm (Chib & Greenberg, 1988) and a proposal prior algorithm (Asparouhov & Muthén, 2010b). The difference in performance between the three methods is studied in simulations described in a later section. In these simulations, Method 2 appears to be preferable.

The choice of inverse-Wishart prior should be made to reflect prior beliefs in the potential magnitude of residual covariances. This is accomplished by using a sufficiently large choice for the degrees of freedom (*df*) of the inverse-Wishart distribution. To obtain a proper posterior where the marginal mean and variance is defined, $df \geq p + 4$ needs to be chosen, where p is the number of variables *y*. The prior means for the residual covariances can be chosen as zero, and the degree of informativeness can be specified using the *df,* which affects the marginal prior variance through $df - p$. For example, Equation A17 of the Appendix shows that using the inverse-Wishart prior $IW(\mathbf{I}, df)$ with $df = p + 6$ gives a prior standard deviation of 0.1, so that two standard deviations below and above the zero mean correspond to the residual covariance range of −0.2 to 0.2. The effect of priors is relative to the variances of the *y*s. For scale-free models, the variables may be standardized before analysis. For larger sample sizes, the prior needs to use a larger *df* to give the same effect.

Methods 1 and 2 both use conjugate priors, that is, the posterior distribution of the covariance matrices is also of the Wishart

Table 2
Choice of Variance for a Normal Prior With Mean Zero

Variance	90% limits	95% limits
0.001	±0.05	±0.06
0.005	±0.12	±0.14
0.01	±0.16	±0.20
0.02	±0.23	±0.28
0.03	±0.28	±0.34
0.04	±0.33	±0.39
0.05	±0.37	±0.44
0.06	±0.40	±0.48
0.07	±0.44	±0.52
0.08	±0.47	±0.55
0.09	±0.49	±0.59
0.10	±0.52	±0.62

family of distributions. This generally produces good convergence of the MCMC chain. Both versions of Method 3, random walk and proposal prior algorithm, are instead based on the Metropolis–Hastings algorithm, and that generally yields somewhat worse convergence performance. The random walk algorithm has difficulty converging or converges very slowly when the variance–covariance matrix has a large number of parameters. However, when a large number of parameters have small prior variance, the convergence is fast. The proposal prior algorithm generally works well but not when the prior variance is very small.

When estimating these models all the algorithms are typically applied under difficult conditions with a nearly unidentified model and near zero prior variance. The models should be estimated with a large number of MCMC iterations, for example, $I = 50{,}000$ or $I = 100{,}000$. Convergence of the MCMC sequence should be carefully evaluated, and automated convergence criteria, such as PSR, are not always reliable. The stability of the parameter values across the iterations should be studied. This can be done by comparing estimates from, say, I, $2I$, and $4I$ MCMC iterations.

Study 1: Cross-Loadings in CFA

Holzinger and Swineford's (1939) Mental Abilities Example: ML Analysis

The first example uses data from the classic 1939 factor analysis study by Holzinger and Swineford (1939). Twenty-six tests intended to measure a general factor and five specific factors were administered to seventh- and eighth-grade students in two schools, the Grant–White school ($n = 145$) and the Pasteur school ($n = 156$). Students from the Grant–White school came from homes where the parents were mostly American born, whereas students from the Pasteur school came largely from working-class parents of whom many were foreign born and used their native language at home.

Factor analyses of these data have been described, for example, by Harman (1976, pp. 123–132) and Gustafsson (2002). Of the 26 tests, 19 were intended to measure four domains, five measured general deduction, and two were revisions/new test versions. Typically, the last two tests are not analyzed. Excluding the five general deduction tests, 19 tests measuring four domains are considered here, where the four domains are spatial ability, verbal ability, speed, and memory. The design of the measurement of the four domains by the 19 tests is shown in the factor loading pattern matrix of Table 1. Here, an X denotes a free loading to be estimated, and 0 a fixed, zero loading. This corresponds to a simple structure CFA model with a variable complexity of one, that is, each variable loads on only one factor.

Using ML estimation, the model fit using both CFA and EFA is reported in Table 3 for both the Grant–White and Pasteur schools. It is seen that the CFA model is rejected by the likelihood-ratio chi-square test in both samples. Given the rather small sample sizes, one cannot attribute the poor fit to the chi-square test being overly sensitive to small misspecifications due to a large sample size as is often done. For completeness, the common fit indices root-mean-square error of approximation (RMSEA) and comparative fit index (CFI) are also shown. In contrast to the CFA, Table 3 shows that the EFA model fits the data well in both schools.

Table 3
Maximum Likelihood Model Testing Results for Holzinger and Swineford's (1939) Data for the Grant–White (N = 145) and Pasteur (N = 156) Schools

School and model	χ^2	*df*	*p*	RMSEA	CFI
Grant–White					
CFA	216	146	.000	0.057	0.930
EFA	110	101	.248	0.025	0.991
Pasteur					
CFA	261	146	.000	0.071	0.882
EFA	128	101	.036	0.041	0.972

Note. RMSEA = root-mean-square error of approximation; CFI = comparative fit index; CFA = confirmatory factor analysis; EFA = exploratory factor analysis.

Table 4 shows the EFA factor solution for both schools using the Geomin rotation. The Quartimin rotation gives similar results. For a description of these rotations, see, for example, Asparouhov and Muthén (2009) and M. W. Browne (2001). The table shows that the major loadings of the EFA correspond to the hypothesized four-factor loading pattern (bolded entries). Several of the tests, however, also have significant cross-loadings on other factors (significance on the 5% level marked with asterisks). There are six significant cross-loadings for the Grant–White solution and nine for the Pasteur solution. This explains the poor fit of the CFA model.

The question arises of how to go beyond postulating only the number of factors as in EFA and maintain the essence of the hypothesized factor loading pattern without resorting to an exploratory rotation. Cross-loadings need to be allowed to some degree, but a model with freely estimated cross-loadings is not identified. The proposed Bayesian solution to this problem is presented next. As an intermediate step, however, it is instructive to consider the EFA alternative of target rotation (Asparouhov & Muthén, 2009; M. W. Browne, 2001). Here, a rotation is chosen to match certain zero target loadings using a least-squares fitting function. Target rotation is similar to BSEM in that it replaces mechanical rotation with rotation guided by the researcher's judgment, in this case using zero targets for cross-loadings. Target rotation is also similar in that the fitting function can result in nonzero values for the targets. It is different from BSEM by not allowing user-specified stringency of closeness to zero by varying the prior variance, replacing that with least-squares fitting. It is also different from BSEM because specifying $m - 1$ zeros for each of the m factors gives the same model fit as specifying more zeros. For Holzinger and Swineford's (1939) two data sets, applying target rotation with zero targets for all cross-loadings gives results similar to EFA using Geomin or Quartimin rotation. Four more significant cross-loadings are obtained for Grant–White, adding to 10, whereas Pasteur obtains six more cross-loadings, adding to 15. The increase in number of significant loadings is due to smaller standard errors for target rotation as compared with Geomin rotation. The smaller standard errors are a function of the more fully specified rotation criterion. As is shown next, BSEM using small-variance cross-loading priors gives far simpler loading patterns by shrinking the cross-loadings toward their zero prior means.

Table 4
Holzinger and Swineford's (1939) Maximum Likelihood Exploratory Factor Analysis Using 19 Variables and Geomin Rotation: Four-Factor Solution

	Grant–White school				Pasteur school			
Test	Spatial	Verbal	Speed	Memory	Spatial	Verbal	Speed	Memory
				Loadings				
Visual perception	**0.628***	0.065	0.091	0.085	**0.580***	0.307*	−0.001	0.053
Cubes	**0.485***	0.050	0.007	−0.003	**0.521***	0.027	−0.078	−0.059
Paper form board	**0.406***	0.107	0.084	0.083	**0.484***	0.101	−0.016	−0.229*
Flags	**0.579***	0.160	0.013	0.026	**0.687***	−0.051	0.067	0.101
General information	0.042	**0.752***	0.126	−0.051	−0.043	**0.838***	0.042	−0.118
Paragraph comprehension	0.021	**0.804***	−0.056	0.098	0.026	**0.800***	−0.006	0.069
Sentence completion	−0.039	**0.844***	0.085	−0.057	−0.045	**0.911***	−0.054	−0.029
Word classification	0.094	**0.556***	0.197*	0.019	0.098	**0.695***	0.008	0.083
Word meaning	0.004	**0.852***	−0.074	0.069	0.143*	**0.793***	0.029	−0.023
Addition	−0.302*	0.029	**0.824***	0.078	−0.247*	0.067	**0.664***	0.026
Code	0.012	0.050	**0.479***	0.279*	0.004	0.262*	**0.552***	0.082
Counting groups of dots	0.045	−0.159	**0.826***	−0.014	0.073	−0.034	**0.656***	−0.166
Straight and curved capitals	0.346*	0.043	**0.570***	−0.055	0.266*	−0.034	**0.526***	−0.056
Word recognition	−0.024	0.117	−0.020	**0.523***	−0.005	0.020	−0.039	**0.726***
Number recognition	0.069	0.021	−0.026	**0.515***	−0.026	−0.057	−0.057	**0.604***
Figure recognition	0.354*	−0.033	−0.077	**0.515***	0.329*	0.042	0.168	**0.403***
Object–number	−0.195	0.045	0.154	**0.685***	−0.123	−0.005	0.333*	**0.469***
Number–figure	0.225	−0.127	0.246*	**0.450***	−0.014	0.092	0.092	**0.427***
Figure–word	0.069	0.099	0.058	**0.365***	0.139	0.013	0.237*	**0.291***
				Factor correlations				
Spatial	—				—			
Verbal	0.378*	—			0.186*	—		
Speed	0.372*	0.386*	—		0.214	0.326*	—	
Memory	0.307*	0.380*	0.375*	—	0.190*	0.100	0.242*	—

Note. Values in bold indicate hypothesized major loadings.
* $p < .05$.

Holzinger and Swineford's (1939) Mental Abilities Example: Bayesian Analysis

This section uses data from the Grant–White and Pasteur schools in Holzinger and Swineford's (1939) study to illustrate the BSEM approach with informative cross-loading priors. The factor loading pattern of the four-factor model of Table 1 is used. Table 5 repeats the fit statistics for the ML CFA and EFA and adds the fit statistics for Bayesian analysis using both the original CFA model and the proposed CFA model with informative, small-variance priors for cross-loadings. The cross-loading priors use variances 0.01. All other parameters have noninformative priors. Standardized variables are analyzed, and the factor variances are fixed at one in order for the priors to correspond to standardized loadings. In all analyses, the reported estimates are the median values of the parameter posterior (this is the Mplus default).

Table 5 shows that the Bayesian analysis of the CFA model with exact zero cross-loadings gives almost zero posterior predictive *p* values in line with the ML CFA. In contrast, for the proposed Bayesian CFA with cross-loadings, model fit is acceptable in that the posterior predictive *p* value is .361 for Grant–White and .162 for Pasteur. Also given are the 2.5% and 97.5% posterior predictive limits for the difference in the fit statistic for the real and replicated data described earlier in the Model Fit section. A positive lower limit is in line with a low posterior predictive *p* value and indicates poor fit, whereas an interval with a zero fit statistic difference falling close to the middle of the interval indicates good fit.

As an aside, the Bayesian estimates can be used as fixed parameters in an ML analysis to get the likelihood-ratio test (LRT) value for the Bayes solution. They can be viewed as a descriptive measure of fit that can be compared with the ML likelihood-ratio chi-square values. It is seen in Table 5 that the Bayesian LRT values for the CFA model are close to those of ML chi-square values. In contrast, the Bayesian LRT values for the model with cross-loadings show a great improvement, falling in between the ML CFA and EFA chi-square values although closer to the EFA values.

The Bayes solutions for the two schools are shown in Table 6. It is interesting to compare the Bayes solution with the ML EFA solution of Table 4. The Bayes factor loadings are, on the whole, somewhat larger than those for ML, and there are far fewer statistically significant cross-loadings (marked with asterisks). In the Bayes context, the frequentist term *significant* should be taken to mean that the 95% credibility interval does not contain zero. For ML EFA, there are six significant cross-loadings for Grant–White and nine for Pasteur, whereof only three appear for both schools. Because they appear for both schools, a researcher may be tempted to free these three cross-loadings. For Bayes, the Grant–White sample has only two cross-loadings that are significant (have a 95% credibility interval that does not cover zero), and Pasteur has one. These three cross-loadings are also significant in the ML

Table 5
Maximum Likelihood Versus Bayes Model Testing Results for Holzinger and Swineford's (1939) Data for Grant–White (N = 145) and Pasteur (N = 156) Schools

	Maximum likelihood analysis				
Model	χ^2	*df*	*p*	RMSEA	CFI
Grant–White					
CFA	216	146	0.000	0.057	0.930
EFA	110	101	0.248	0.025	0.991
Pasteur					
CFA	261	146	0.000	0.071	0.882
EFA	128	101	0.036	0.041	0.972

	Bayesian analysis			
Model	Sample LRT	2.5% PP limit	97.5% PP limit	PP *p*
Grant–White				
CFA	219	12	112	0.006
CFA with cross-loadings	142	−39	61	0.361
Pasteur				
CFA	264	56	156	0.000
CFA with cross-loadings	156	−28	76	0.162

Note. RMSEA = root-mean-square error of approximation; CFI = comparative fit index; CFA = confirmatory factor analysis; EFA = exploratory factor analysis; LRT = likelihood-ratio test; PP = posterior predictive.

EFA. In the Bayes analysis, the significant cross-loadings are different in the two schools. Because of the lack of agreement, freeing these three cross-loadings could be capitalizing on chance and is also not necessary on behalf of model fit. Bayes clearly gives a simpler pattern than ML EFA for these data. This is achieved by shrinking the cross-loadings toward their zero prior means. The degree of shrinking that is possible while still obtaining reasonable model fit is gauged by the posterior predictive *p* value.

It should be emphasized that cross-loadings that are found to be important in BSEM—in the sense that the 95% credibility interval does not cover zero and the cross-loading has strong substantive backing—can be freely estimated with noninformative priors while keeping small-variance priors for other cross-loadings. This should improve the results because the small-variance prior gives a too small estimate for such a cross-loading. Monte Carlo simulations show that this gives better estimation.

The ML EFA factor correlations are smaller than the Bayesian factor correlations as is seen when comparing Table 4 with Table 6. The greater extent of cross-loadings in the EFA may contribute to the lower factor correlations in that less correlation among variables needs to go through the factors. The Bayesian factor correlations are not excessively high, however, because the factors are expected to correlate to a substantial degree according to theory. Holzinger and Swineford (1939) hypothesized that the variables are all influenced by a deductive factor, which in the current model is not partialed out of the four factors.

The choice of cross-loading prior variance should be linked to the researcher's prior beliefs. It could be argued, however, that the choice of a variance of 0.01 resulting in 95% cross-loading limits of ±0.20 is not substantially different from a variance of, say, 0.02 resulting in 95% limits of ±0.28; see Table 2. It is therefore of interest to vary the prior variance to study sensitivity in the results. Increasing the prior variance tends to affect the posterior predictive *p* value and also increase the variability of the estimates. At a certain point of increasing the prior variance, the model is not sufficiently identified, and the MCMC algorithm tends to give nonconvergence. Table 7 shows the effects of varying the prior variance for cross-loadings from 0.01 to 0.10 for the data from both the Grant–White and the Pasteur schools. The table gives the absolute value of the 95% limit of the prior distribution, the posterior predictive *p* value, the largest cross-loading with its posterior standard deviation, and the range of the factor correlations. For Grant–White, the largest cross-loading is observed for the straight test loading on the spatial factor, whereas for Pasteur the largest loading is observed for the figure recognition test loading on the spatial factor. The change in prior variance does not affect the hypothesized pattern of major loadings, and this is not reported. The range of factor correlations is included given that larger prior variance may lead to larger cross-loadings, which in turn may have the effect of lowering the factor correlations because correlations among the tests have to be channeled through the factors to a lesser degree.

Table 7 suggests that the prior variance of 0.01 may be on the low side in the sense that for both schools, the posterior predictive *p* value peaks at the prior variance 0.03 (95% cross-loading limit of 0.34). The change in prior variance, however, does not affect the results in important ways for these two data sets. For all prior variance choices, the largest cross-loading for Grant–White and for Pasteur is detected, in the sense that it has its 95% credibility interval not covering zero. For Pasteur, the three highest prior variances result in the figurer cross-loading not being detected (getting a 95% credibility interval that does cover zero; entries shown as dashes in the table). This is because of the higher posterior variability at the higher prior variances. In hindsight, perhaps a prior variance of 0.02 or 0.03 would have been a slightly better choice, but this may not be true for other examples. On the other hand, when presenting results, it is useful to give information on how a range of prior choices affects the results. Although the factor correlations show smaller values with increasing prior variance, the decrease is small and of little substantive importance. All in all, these results are reassuring in that the exact degree of informativeness does not seem critical. Also, with larger sample sizes, the choice is less important in that the data provide relatively more information than the priors.

In summary, BSEM provides a simpler model and a model that fits the researcher's prior beliefs better than ML. BSEM provides an approach that is a compromise between that of EFA and CFA. The ML CFA rejects the hypothesized model, presumably because it is too strict. ML EFA does not reject the model, but the model does not match the researcher's prior beliefs because it postulates only the number of factors, not where the large and small loadings should appear. Furthermore, ML EFA provides a solution through a mechanical rotation algorithm, whereas BSEM uses priors to represent the researcher's beliefs.

Cross-Loading Simulations

This section discusses Monte Carlo simulations of BSEM applied to factor modeling with cross-loadings. The aim is to dem-

Table 6
Bayes for Holzinger and Swineford's (1939) Example: Four-Factor Solution Using Informative Priors for Cross-Loadings

	Grant–White school				Pasteur school			
Test	Spatial	Verbal	Speed	Memory	Spatial	Verbal	Speed	Memory
				Loadings				
Visual perception	**0.640***	0.012	0.050	0.047	**0.633***	0.145	0.027	0.039
Cubes	**0.521***	−0.008	−0.010	−0.012	**0.504***	−0.027	−0.041	−0.030
Paper form board	**0.456***	0.040	0.041	0.047	**0.515***	0.018	−0.024	−0.118
Flags	**0.672***	0.046	−0.020	0.005	**0.677***	−0.095	0.026	0.093
General information	0.037	**0.788***	0.049	−0.040	−0.056	**0.856***	0.027	−0.084
Paragraph comprehension	−0.001	**0.837***	−0.053	0.030	0.015	**0.801***	−0.011	0.050
Sentence completion	−0.045	**0.885***	0.021	−0.055	−0.063	**0.925***	−0.032	−0.036
Word classification	0.053	**0.612***	0.096	0.029	0.055	**0.694***	0.013	0.063
Word meaning	−0.012	**0.886***	−0.086	0.020	0.092	**0.803***	0.001	0.012
Addition	−0.172*	0.030	**0.795***	0.004	−0.147	−0.004	**0.655***	0.010
Code	−0.002	0.054	**0.560***	0.130	−0.004	0.111	**0.655***	0.049
Counting groups of dots	0.013	−0.092	**0.828***	−0.049	0.025	−0.058	**0.616***	−0.057
Straight and curved capitals	0.189*	0.043	**0.633***	−0.035	0.132	−0.067	**0.558***	0.001
Word recognition	−0.040	0.044	−0.031	**0.556***	−0.058	0.006	−0.090	**0.731***
Number recognition	0.003	−0.004	−0.038	**0.552***	0.006	−0.098	−0.106	**0.634***
Figure recognition	0.132	−0.024	−0.049	**0.573***	0.156*	0.027	0.064	**0.517***
Object–number	−0.139	0.014	0.029	**0.724***	−0.097	0.007	0.122	**0.545***
Number–figure	0.099	−0.071	0.095	**0.564***	−0.029	0.041	0.003	**0.474***
Figure–word	0.012	0.045	0.007	**0.445***	0.049	0.018	0.085	**0.397***
				Factor correlations				
Spatial	—				—			
Verbal	0.535*	—			0.348*	—		
Speed	0.471*	0.443*	—		0.307	0.457*	—	
Memory	0.526*	0.515*	0.557*	—	0.324*	0.179	0.405*	—

Note. Values in bold indicate hypothesized major loadings. Statistically significant cross-loadings (marked with asterisks) have a 95% credibility interval that does not cover zero.

onstrate that the proposed approach provides good results for data with known characteristics.

The factor loading pattern of Table 8 is considered where X denotes a major loading and x denotes cross-loadings. The major loadings are all 0.8. The sizes of the three cross-loadings are varied as 0.0, 0.1, 0.2, and 0.3 in different simulations. The observed and latent variables have unit variances so the loadings are on a standardized scale. A cross-loading of 0.1 is considered to be of little importance, a cross-loading of 0.2 is considered to be of some importance, and a cross-loading of 0.3 is considered to be of importance (see also Cudeck & O'Dell, 1994). The correlations among the three factors are all 0.5. The factor metric is determined by fixing the first loading for each factor at 0.8. Noninformative priors are used for all parameters except for cross-loadings when those are included as parameters in the analysis. Informative priors are used for all cross-loadings, not just the three that have population values different from zero. For cross-loading priors, a variance of 0.01 is chosen. As a first step, cross-loadings are not included in the analysis, although the data are generated with three cross-loadings, to compare regular ML CFA with Bayesian CFA (not using BSEM). Sample sizes of $n = 100$, $n = 200$, and $n = 500$ are studied.

A total of 500 replications are used. The reported parameter estimate is the median in the posterior distribution for each parameter. The key result is what frequentists would refer to as the 95% coverage, that is, the proportion of the replications for which the 95% Bayesian credibility interval covers the true parameter value used to generate the data (credibility intervals obtained through percentiles of the posterior). For cross-loadings, it is also of interest to study what corresponds to power in a frequentist setting. This is computed as the proportion of the replications for which the 95% Bayesian credibility interval does not cover zero. Results are reported only for a representative set of parameters or functions of parameters: the major loading of $y2$, the cross-loading for $y6$, the variance for the first factor, and the correlation between the first and second factor. The model rejection rate is reported using the proportion of replications with a posterior predictive p value of at least .05.

The results are divided into three categories: Bayesian analysis using noninformative priors, model fit comparisons between ML and Bayes with noninformative priors, and Bayesian analysis with informative priors. Not all results are presented here, but some tables are instead available on the Web page www.statmodel.com/examples/penn.shtml#baysem.

Bayes, Noninformative Priors

As a check of the Bayesian analysis procedure, a first analysis is carried out with noninformative priors and ignoring cross-loadings. Results can be found in Web Table 1 at www.statmodel.com/examples/penn.shtml#baysem. Data are generated both with zero and nonzero cross-loadings. For zero cross-loadings, the analysis is correctly specified and close to 95% coverage is obtained for all free parameters. Posterior predictive p value rejection rates for the model fit assessment are 0.036, 0.032, and 0.024, respectively, for the

Table 7
Effects of Using Different Variances for the Informative Priors of the Cross-Loadings for Holzinger and Swineford's (1939) Data

Prior variance	95% cross-loading limit	PP *p*	Cross-loading (posterior *SD*)	Factor correlation range
Grant–White school				
0.01	0.20	0.361	0.189 (0.078)	0.443–0.557
0.02	0.28	0.441	0.248 (0.096)	0.439–0.542
0.03	0.34	0.457	0.275 (0.109)	0.423–0.530
0.04	0.39	0.455	0.292 (0.120)	0.413–0.521
0.05	0.44	0.453	0.303 (0.130)	0.404–0.513
0.06	0.48	0.447	0.309 (0.139)	0.400–0.510
0.07	0.52	0.439	0.315 (0.148)	0.395–0.508
0.08	0.55	0.439	0.319 (0.156)	0.387–0.508
0.09	0.59	0.435	0.323 (0.163)	0.378–0.506
1.00	0.62	0.427	0.327 (0.171)	0.369–0.504
Pasteur school				
0.01	0.20	0.162	0.132 (0.076)	0.179–0.457
0.02	0.28	0.205	0.201 (0.088)	0.184–0.441
0.03	0.34	0.219	0.223 (0.098)	0.188–0.431
0.04	0.39	0.218	0.237 (0.106)	0.189–0.424
0.05	0.44	0.205	0.247 (0.115)	0.175–0.408
0.06	0.48	0.196	0.255 (0.122)	0.175–0.402
0.07	0.52	0.195	0.261 (0.128)	0.176–0.397
0.08	0.55	0.192	—	0.176–0.394
0.09	0.59	0.187	—	0.177–0.391
0.10	0.62	0.185	—	0.177–0.388

Note. Dashes indicate that for the Pasteur school, the three highest prior variances result in the figurer cross-loading not being detected (getting a 95% credibility interval that does cover zero). PP *p* = posterior predictive probability.

three sample sizes of $n = 100$, $n = 200$, and $n = 500$, that is, reasonably close to the nominal 5% level. Bayesian analysis with noninformative priors works well in this situation.

With cross-loadings of 0.1, the effects of model misspecification show up in that the coverage is less good. Posterior predictive *p* value rejection rates for the model fit assessment are 0.056, 0.080, and 0.262, respectively, for the three sample sizes. This shows limited power to reject the incorrect model. On the other hand, the misspecification is deemed of little importance given the small size of the cross-loadings.

Table 8
Factor Loading Pattern for Simulation Study of Cross-Loadings: Factor Loading Pattern

Variable	Factor 1	Factor 2	Factor 3
*y*1	X	0	x
*y*2	X	0	0
*y*3	X	0	0
*y*4	X	0	0
*y*5	X	0	0
*y*6	x	X	0
*y*7	0	X	0
*y*8	0	X	0
*y*9	0	X	0
*y*10	0	X	0
*y*11	0	x	X
*y*12	0	0	X
*y*13	0	0	X
*y*14	0	0	X
*y*15	0	0	X

Note. X denotes a major loading, and x denotes cross-loadings.

With cross-loadings of 0.2 (not shown) the posterior predictive *p* value rejection rate is .196 for $n = 100$, .474 for $n = 200$, and .984 for $n = 500$, showing excellent power at higher sample sizes. With cross-loadings of 0.3 the posterior predictive *p* value is .544 for $n = 100$, .944 for $n = 200$, and 1.000 for $n = 500$, showing that the power is excellent when the cross-loading is of an important magnitude.

Comparing Model Fit for ML Versus Bayes With Noninformative Priors

Model fit assessment comparing ML to Bayesian analysis with noninformative priors is shown in Table 9. The correctly specified model with zero cross-loadings shows an inflated ML 5% rejection rate of 0.172 at $n = 100$. This small-sample bias is well-known for ML chi-square testing (see, e.g., Scheines et al., 1999). The posterior predictive *p* value rejection rate of .036 based on the Bayesian analysis does not show the same problem. For the 0.1 size of the cross-loadings, which is deemed of little substantive importance, ML rejects the model 46% of the time at $n = 500$. This reflects the common notion that the ML LRT chi-square can be oversensitive to small degrees of model misspecification. For the important degree of misspecification with cross-loading 0.3, the ML test is more powerful than Bayes, but the Bayes power is sufficient for sample sizes of at least $n = 200$.

Web Table 2 shows ML model estimation results as a comparison with the Bayesian analysis with noninformative priors presented earlier. For both the correctly specified model with zero cross-loadings and for the misspecified model with cross-loadings 0.1, the ML coverage is close to that of Bayes. The mean square error (*MSE*) is also similar for Bayes and ML. On the basis of this, there is no reason to prefer one method over the other.

Table 9
Rejection Rates for ML Confirmatory Factor Analysis and Bayes's Confirmatory Factor Analysis With Noninformative Priors

Cross-loading and sample size	ML LRT rejection rate	Bayes PP p rejection rate
0.0		
100	0.172	0.036
200	0.090	0.032
500	0.060	0.024
0.1		
100	0.226	0.056
200	0.228	0.080
500	0.460	0.262
0.2		
100	0.488	0.196
200	0.726	0.474
500	0.998	0.984
0.3		
100	0.830	0.544
200	0.996	0.944
500	1.000	1.000

Note. ML = maximum likelihood; LRT = likelihood-ratio test; PP p = posterior predictive probability.

Bayes, Informative Priors

As the next step, the proposed BSEM approach of using Bayesian analysis with informative, small-variance priors for the cross-loadings is applied. The informative priors are applied not only to the three cross-loadings used to generate the data but to all cross-loadings to reflect a real-data analysis situation. The prior variance is chosen as 0.01. All other parameters are given noninformative priors. Table 10 shows good coverage, and for the top part of the table corresponding to the correctly specified analysis with zero cross-loadings, the coverage remains largely the same as in Web Table 1. For cross-loadings of 0.1, however, the bottom part of the table shows that coverage has improved by the introduction of informative, small-variance priors for the cross-loadings. The coverage is acceptable also for the cross-loading. The power to detect the cross-loading is, however, small at this low cross-loading magnitude, 0.038, 0.098, and 0.176, respectively, for the three sample sizes. The posterior predictive p value is on the low side in all four cases.

It is interesting to compare the coverage results for the four parameters in the case of cross-loadings 0.1 given in Table 10 for BSEM with the results from the ML approach in Web Table 2. ML is outperformed by BSEM by its use of informative priors.

Table 11 shows the results of BSEM where data have been generated with larger cross-loadings of 0.2 and 0.3. Here, the coverage is also good with the exception of the cross-loadings. For the cross-loadings, however, the focus is on power as shown in the last columns. For a cross-loading of 0.2, a sample size of $n = 500$ is needed to obtain power above 0.8. For a cross-loading of 0.3, a sample size of $n = 200$ is sufficient to obtain power above 0.8. This shows that the approach of using informative, small-variance priors for cross-loadings leads to a successful way to modify the model, allowing free estimation of the indicated cross-loadings. When freed and estimated using noninformative priors, these cross-loadings are well estimated.

The point estimates indicate that the key parameter of factor correlation is overestimated. Note, however, that given the power to detect cross-loadings, estimating the cross-loadings freely results in good point estimates for factor correlations.

For the case of 0.3 cross-loadings in Table 11, the alternative prior variance of 0.02 was also tried, yielding improved results. The average estimates for the four entries were 0.8060, 0.2117, 1.0965, and 0.5620, whereas the coverage was 0.956, 0.886, 0.954, and 0.982.

In summary, the cross-loading simulation study shows that the Bayesian analysis performs well. It also shows that in terms of parameter coverage and for the case of small cross-loadings, ML is inferior to BSEM. In terms of model testing, BSEM avoids the small-sample inflation of the ML chi-square and also avoids the ML chi-square sensitivity to rejecting a model with an ignorable degree of misspecification.

Study 2: Residual Correlations in CFA

British Household Panel Study Big Five Personality Example: ML Analysis

A second example uses data from the British Household Panel Study of 2005 and 2006. A 15-item, five-factor instrument uses three items to measure each of the Big Five personality factors: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness. Each item uses the statement, "I see myself as someone who . . . ," followed by a statement. There are seven response categories ranging from 1 (*does not apply*) to 7 (*applies perfectly*). A total of 14,021 subjects are included. The Big Five factors are expected a priori to have low correlations and are known to be related to gender and age (see, e.g., Marsh, Nagengast, & Morin, 2010). For simplicity, the current analyses hold age constant by considering the subgroup of ages 50 to 55. This produces a sample of $n = 691$ female and $n = 589$ male participants.

The item wording and hypothesized loading pattern are shown in Table 12. For all factors except openness, there are two positively worded items and one negatively worded item. Marsh et al. (2010) suggested that the four negatively worded items may a priori have correlated residuals (correlated uniquenesses) when applying factor analysis.

Using ML estimation, Table 13 reports model fit using CFA, CFA with correlated uniquenesses among the negatively worded items, and EFA. It is seen that the fit is not acceptable for either of the two CFA models as judged by chi-square or the two model fit indices. The EFA model is also rejected by chi-square and is only marginally acceptable for male participants when judged by RMSEA or CFI.

An interesting finding is that the EFA solutions for female and male participants do not fully capture the hypothesized factors. This is the case using the Geomin rotation as well as using Quartimin and Varimax. The Geomin rotation for each gender is shown in Table 14. The bolded entries are loadings that are the largest for the item. When comparing Table 14 with Table 12, it is seen that only the factors Extraversion, Neuroticism, and Openness are found, not the Agreeableness and Conscientiousness factors. A possible reason for this is the existence of correlated residuals among the items. As the CFA with correlated uniquenesses model showed, however, allowing residual correlations among the

Table 10
Bayesian Analysis Using Informative, Small-Variance Priors for Cross-Loadings 0.0 and 0.1

Parameter	Population	Estimates average	*SD*	*SE* average	*MSE*	95% cover	% sig coeff
		Cross-loading = 0.0, n = 100, 5% reject proportion for the PP p = .006					
Major loading	0.800	0.8472	0.1212	0.1310	0.0169	0.950	1.000
Cross-loading	0.000	0.0141	0.0455	0.0732	0.0023	0.998	0.002
Factor variance	1.000	0.9864	0.2595	0.2624	0.0674	0.930	1.000
Factor correlation	0.500	0.4967	0.0869	0.1076	0.0075	0.980	0.982
		Cross-loading = 0.0, n = 200, 5% reject proportion for the PP p = .002					
Major loading	0.800	0.8311	0.0893	0.0894	0.0089	0.942	1.000
Cross-loading	0.000	0.0079	0.0421	0.0633	0.0018	1.000	0.000
Factor variance	1.000	0.9662	0.1799	0.1840	0.0335	0.948	1.000
Factor correlation	0.500	0.4962	0.0605	0.0860	0.0037	0.990	1.000
		Cross-loading = 0.0, n = 500, 5% reject proportion for the PP p = .010					
Major loading	0.800	0.8161	0.0520	0.0552	0.0030	0.962	1.000
Cross-loading	0.000	0.0033	0.0313	0.0530	0.0010	1.000	0.000
Factor variance	1.000	0.9741	0.1096	0.1293	0.0127	0.958	1.000
Factor correlation	0.500	0.4960	0.0406	0.0708	0.0017	1.000	1.000
		Cross-loading = 0.1, n = 100, 5% reject proportion for the PP p = .006					
Major loading	0.800	0.8218	0.1177	0.1263	0.0143	0.950	1.000
Cross-loading	0.100	0.0594	0.0449	0.0728	0.0037	0.982	0.038
Factor variance	1.000	1.0600	0.2738	0.2808	0.0784	0.934	1.000
Factor correlation	0.500	0.5206	0.0850	0.1047	0.0076	0.976	0.992
		Cross-loading = 0.1, n = 200, 5% reject proportion for the PP p = .006					
Major loading	0.800	0.8151	0.0860	0.0882	0.0076	0.950	1.000
Cross-loading	0.000	0.0666	0.0424	0.0636	0.0029	0.978	0.098
Factor variance	1.000	1.0217	0.1895	0.1978	0.0363	0.942	1.000
Factor correlation	0.500	0.5204	0.0602	0.0843	0.0040	0.984	1.000
		Cross-loading = 0.1, n = 500, 5% reject proportion for the PP p = .008					
Major loading	0.800	0.8089	0.0517	0.0551	0.0027	0.964	1.000
Cross-loading	0.100	0.0732	0.0316	0.0532	0.0017	0.990	0.176
Factor variance	1.000	1.0169	0.1160	0.1371	0.0137	0.972	1.000
Factor correlation	0.500	0.5229	0.0404	0.0688	0.0022	0.998	1.000

Note. sig = significant; coeff = coefficient; PP p = posterior predictive probability.

reverse-coded items is not sufficient. It is likely that, in addition to the Big Five factors, the personality instrument measures many minor factors.

The question arises of how correlated residuals can be accounted for while maintaining the hypothesized factor loading pattern. A model with all residual correlations freely estimated is not identified. The proposed BSEM solution to this problem is presented next.

British Household Panel Study Big Five Personality Example: Bayesian Analysis

This section uses the Big Five personality data in the British Household Panel Study to illustrate the BSEM approach with an informative prior for the residual covariance matrix. Because of its relative simplicity, Method 2 is used. The simulation studies to be presented also favor Method 2. An inverse-Wishart prior, $IW(\boldsymbol{I}, df)$ with $df = p + 6 = 21$, is used, corresponding to prior means and standard deviations for residual covariances of zero and 0.1, respectively (see Appendix, Equation A17). Standardized variables are analyzed. Because of high auto-correlation among the MCMC iterations, only every 10th iteration is used with a total of 100,000 iterations to describe the posterior distribution. Informative cross-loading priors are also used with prior distributions $N(0, 0.01)$.

The posterior predictive p values are .534 and .518, respectively, for female and male samples, indicating a good match between the model and the data. For the two samples, 17 and 37 residual covariances, respectively, were found significant in the sense of the 95% Bayesian credibility interval not covering zero. The average absolute residual correlation (range) is 0.329 (−0.462 to 0.647) for the female sample and 0.285 (−0.484 to 0.590) for the male sample. For both genders, only one residual correlation exceeds 0.5 in absolute value. This suggests that many small residual correlations need to be included in the factor model, as was expected. The fact that these residual correlations are left out in the ML analyses may contribute to the poor ML fit and the poor ML EFA loading pattern recovery.

Table 15 gives the results for the female and male samples. Standardized loadings are presented so that the results can be compared with the ML EFA of Table 14. The hypothesized major loadings are all recovered at substantial values with no significant cross-loadings. The factor correlations are all small to moderate as

Table 11
Bayesian Analysis Using Informative, Small-Variance Priors for Cross-Loadings 0.2 and 0.3

Parameter	Population	Estimates average	*SD*	*SE* average	*MSE*	95% cover	% sig coeff
Cross-loading = 0.2, n = 100, 5% reject proportion for the PP p = 0.010							
Major loading	0.800	0.7952	0.1152	0.1217	0.0133	0.952	1.000
Cross-loading	0.200	0.1024	0.0453	0.0731	0.0116	0.840	0.188
Factor variance	1.000	1.1522	0.2990	0.3018	0.1124	0.908	1.000
Factor correlation	0.500	0.5439	0.0819	0.1023	0.0086	0.966	0.996
Cross-loading = 0.2, n = 200, 5% reject proportion for the PP p = .004							
Major loading	0.800	0.7979	0.0835	0.0859	0.0070	0.940	1.000
Cross-loading	0.200	0.1239	0.0418	0.0638	0.0075	0.856	0.492
Factor variance	1.000	1.0850	0.1978	0.2109	0.0463	0.942	1.000
Factor correlation	0.500	0.5424	0.0581	0.0823	0.0052	0.974	1.000
Cross-loading = 0.2, n = 500, 5% reject proportion for the PP p = .006							
Major loading	0.800	0.8010	0.0514	0.0554	0.0026	0.964	1.000
Cross-loading	0.200	0.1427	0.0327	0.0541	0.0044	0.922	0.854
Factor variance	1.000	1.0595	0.1222	0.1473	0.0184	0.966	1.000
Factor correlation	0.500	0.5445	0.0382	0.0669	0.0034	0.986	1.000
Cross-loading = 0.3, n = 100, 5% reject proportion for the PP p = .012							
Major loading	0.800	0.7671	0.1104	0.1166	0.0139	0.924	1.000
Cross-loading	0.300	0.1428	0.0460	0.0734	0.0268	0.364	0.470
Factor variance	1.000	1.2532	0.3158	0.3281	0.1636	0.860	1.000
Factor correlation	0.500	0.5650	0.0810	0.0999	0.0108	0.952	0.996
Cross-loading = 0.3, n = 200, 5% reject proportion for the PP p = .012							
Major loading	0.800	0.7790	0.0807	0.0836	0.0069	0.948	1.000
Cross-loading	0.300	0.1755	0.0419	0.0640	0.0173	0.518	0.856
Factor variance	1.000	1.1623	0.2134	0.2257	0.0718	0.890	1.000
Factor correlation	0.500	0.5642	0.0577	0.0804	0.0075	0.950	1.000
Cross-loading = 0.3, n = 500, 5% reject proportion for the PP p = .006							
Major loading	0.800	0.7891	0.0493	0.0553	0.0025	0.958	1.000
Cross-loading	0.300	0.2077	0.0318	0.0545	0.0095	0.640	1.000
Factor variance	1.000	1.1116	0.1252	0.1573	0.0281	0.930	1.000
Factor correlation	0.500	0.5636	0.0368	0.0661	0.0054	0.960	1.000

Note. sig = significant; coeff = coefficient; PP p = posterior predictive probability.

was expected. The Extraversion, Neuroticism, and Openness factors that were recovered in the ML EFA of Table 14 have lower correlations in the Bayesian solution than in the ML EFA.

In summary, BSEM provides a solution that fits the researcher's prior beliefs better than ML. The ML CFA rejects the hypothesized model, presumably because it is too strict in terms of requiring exactly zero residual covariances. ML EFA does not recover the researcher's hypothesized Big Five factor pattern.

Residual Correlations Simulations

This section discusses Monte Carlo simulations of the BSEM approach to factor modeling with residual correlations. A factor model with 10 variables and two factors is used, where the first five variables load on only the first factor, and the second five variables load only on the second factor. The loadings are all 0.8, the factor variances are 1, and the factor correlation is 0.5. The residual variances are 0.36 so that observed variables all have variances of 1. Two residual covariances (correlations) are included, one for the first and sixth variables and one for the second and seventh variables. In this way, ignoring the residual covariances in the modeling tends to inflate the factor correlation. An example would be an instrument administered at two time points, where some variables have residuals that are correlated over time. Residual correlations of 0.0, 0.1, and 0.3 are considered together with sample sizes n = 200 and n = 500. A total of 500 replications are used, and the results are presented in the format used for the cross-loading simulations. The simulations present results for all three methods presented earlier. For Method 1, both a more informative prior with df = 30 and a less informative prior with df = 14 (= p + 4) is studied. For Method 2, df = 30 is used. For Method 3, the normal prior variance is set at 0.001. Tables with results are available at the Web page www.statmodel.com/examples/penn.shtml#baysem.

Comparing ML With Bayes

As a first step, model testing using ML and Bayes with noninformative priors is compared for both correctly and misspecified models. With residual correlations of 0.0, both ML and Bayes give acceptable rejection rates with the correctly specified model (results presented in Web Table 3). Both ML and Bayes reject the

Table 12
Wording and Hypothesized Factor Loading Pattern for the 15 Items Used to Measure the Big Five Personality Factors in the British Household Panel Data ("I See Myself As Someone Who . . .")

Item	A	C	E	N	O
*y*1: Is sometimes rude to others (reverse scored)	X	0	0	0	0
*y*2: Has a forgiving nature	X	0	0	0	0
*y*3: Is considerate and kind to almost everyone	X	0	0	0	0
*y*4: Does a thorough job	0	X	0	0	0
*y*5: Tends to be lazy (reverse scored)	0	X	0	0	0
*y*6: Does things efficiently	0	X	0	0	0
*y*7: Is talkative	0	0	X	0	0
*y*8: Is outgoing, sociable	0	0	X	0	0
*y*9: Is reserved (reverse scored)	0	0	X	0	0
*y*10: Worries a lot	0	0	0	X	0
*y*11: Gets nervous easily	0	0	0	X	0
*y*12: Is relaxed, handles stress well (reverse scored)	0	0	0	X	0
*y*13: Is original, comes up with new ideas	0	0	0	0	X
*y*14: Values artistic, aesthetic experiences	0	0	0	0	X
*y*15: Has an active imagination	0	0	0	0	X

Note. A = Agreeableness; C = Conscientiousness; E = Extraversion; N = Neuroticism; O = Openness.

model with ignorable residual correlations of 0.1, although ML is more sensitive to this misspecification. For the larger misspecification with residual correlations of 0.3, both ML and Bayes show sufficient power to reject the model at both $n = 200$ and $n = 500$.

With BSEM Method 1 and residual correlations 0.1, good coverage is found for all parameters except the residual covariance (Web Table 4). There is sufficient power to detect the residual covariance of 0.1 at $n = 500$. There is no important difference between using $df = 30$ and $df = 14$, except that the point estimate and the power for the residual covariance is slightly better for the less informative prior with $df = 14$.

With BSEM Method 1 and residual correlations 0.3, acceptable coverage is found when using the less informative prior with $df = 14$, except for the residual correlation (Web Table 5). The power to detect the residual correlations is, however, excellent in all cases. For the more informative prior with $df = 30$, the coverage is less good. The key parameter of the factor correlation shows an important overestimation, which is also seen with $df = 14$.

Table 13
Maximum Likelihood Model Testing Results for Big Five Personality Factors Using British Household Panel Data for Female (N = 691) and Male (N = 589) Participants

Model	χ^2	*df*	*p*	RMSEA	CFI
Female					
CFA	552	80	.000	0.092	0.795
CFA + CUs	432	74	.000	0.084	0.845
EFA	183	40	.000	0.072	0.938
Male					
CFA	516	80	.000	0.096	0.795
CFA + CUs	442	74	.000	0.092	0.826
EFA	113	40	.000	0.056	0.965

Note. RMSEA = root-mean-square error of approximation; CFI = comparative fit index; CFA = confirmatory factor analysis; CU = correlated uniqueness; EFA = exploratory factor analysis.

The simulation results for BSEM Methods 2 and 3 using residual correlations of 0.3 are studied next (Web Table 6). The 5% reject proportion for the posterior predictive *p* value is 0 in all cases. For Method 2, the results are very good except for the residual covariance being underestimated and having poor coverage. The power to detect it is, however, excellent. The factor correlation is also somewhat underestimated. Method 2 performs considerably better than Method 1. The Method 3 results are somewhat worse than those of Method 1 for $df = 14$, with poorer performance for the residual covariance and the factor correlation. The power to detect the residual covariance is, however, excellent also for Method 3. It should be noted that Method 3 is the only one of the three methods that can let such a residual covariance be freely estimated, that is, using a noninformative prior. Using a less informative Method 3 prior with a larger variance of 0.01 did not alter the results very much. In summary, Method 2 performs the best of the three methods, and Method 3 performs the worst for this simulation setting.

Method 3 works well when the two residual covariances are freely estimated, that is, using noninformative priors (Web Table 7). The remaining residual covariances are using the same informative priors as before. Results are shown for $n = 200$ and $n = 500$.

An Example of Small-Variance Priors for Both Structural and Measurement Parameters

The BSEM approach is not limited to measurement modeling but is also applicable to restrictions on structural coefficients in SEM. Also, although the two application areas studied in this article show the particular advantage of BSEM when ML estimation produces a nonidentified model, this nonidentification aspect is not a requirement for BSEM. Although the topic of Bayesian informative priors for structural coefficients in SEM is not the primary target of this article, the following section gives a brief discussion in the context of an example. The analyses also illustrate how the use of informative priors in the measurement modeling is combined with the use of informative priors in the structural modeling.

Table 14
Maximum Likelihood Exploratory Factor Analysis of the Big Five Personality Factors Using British Household Panel Data

	Female sample					Male sample				
Item	F1	F2	E	N	0	F1	F2	E	N	0
					Loadings					
*y*1	**0.827***	0.000	0.014	−0.011	−0.005	**0.389***	0.010	−0.016	−0.083	−0.294*
*y*2	0.147*	0.215*	**0.323***	0.033	0.020	0.188	**0.447***	0.123*	−0.030	−0.011
*y*3	0.103*	**0.569***	0.280*	0.046	−0.095*	**0.506***	0.469*	0.026	0.042	−0.030
*y*4	0.018	**0.455***	−0.003	−0.025	0.270*	**0.406***	−0.011	0.119*	0.010	0.272*
*y*5	**0.365***	0.220*	−0.039	−0.068	0.009	**0.654***	−0.449*	−0.037	0.009	0.004
*y*6	−0.016	**0.852***	0.001	−0.052	0.087	**0.656***	0.077	0.020	−0.090	0.141*
*y*7	−0.154*	0.053	**0.541***	0.015	0.129*	0.047	0.015	**0.629***	0.045	0.123
*y*8	−0.041	−0.024	**0.748***	−0.049	−0.002	0.012	0.024	**0.795***	−0.051	0.032
*y*9	0.064	**−0.416***	0.346*	−0.116*	0.031	−0.156	−0.380*	**0.396***	−0.010	−0.049
*y*10	−0.045	0.061	0.063	**0.727***	0.036	0.023	0.020	−0.029	**0.698***	0.294*
*y*11	0.021	0.001	−0.039	**0.670***	0.013	−0.075	0.279*	−0.052	**0.519***	0.021
*y*12	0.022	−0.250*	−0.061	**0.547***	−0.063	−0.004	−0.311*	0.051	**0.648***	−0.109
*y*13	0.024	0.011	−0.036	−0.107	**0.764***	0.069	−0.007	−0.001	−0.023	**0.734***
*y*14	0.011	−0.088*	0.038	0.125*	**0.659***	−0.073	0.057	0.035	0.036	**0.486***
*y*15	−0.054	0.140*	0.087	−0.014	**0.541***	0.002	0.006	0.038	−0.146*	**0.671***
					Factor correlations					
F1	—					—				
F2	0.151*	—				0.399*	—			
E	−0.024	0.362*	—			0.268*	0.200*	—		
N	−0.085	0.100*	−0.108*	—		−0.311*	0.065	−0.252*	—	
O	−0.142*	0.229*	0.473*	−0.175*	—	0.344*	0.397*	0.454*	−0.180*	—

Note. The bolded entries are loadings that are the largest for the item. F1 = Factor 1; F2 = Factor 2; E = Extraversion; N = Neuroticism; O = Openness.
* $p < .05$.

Figure 3 shows a structural equation model for science achievement proposed in Kaplan (2009, Figure 4.1). Drawing on the Rand input–process–output model (Shavelson, McDonnell, & Oakes, 1989), Kaplan (2009) stated that

> it is hypothesized that the background student characteristics of previous science grades (scigra6) and socioeconomic status (ses) influence science achievement indirectly through 10th grade science grades. The role of teacher certification in science is hypothesized to predict the extent of hands-on science involvement. This in turn is hypothesized to predict student perceptions of a challenging classroom environment, which in turn should predict science achievement through science grades. (p. 54)

Kaplan (2009) analyzed a sample of 6,677 students in 10th grade in public schools from the National Educational Longitudinal Study of 1988. ML estimation of this model results in a chi-square test of model fit of 1,731 with 39 degrees of freedom (RMSEA = 0.081, CFI = 0.844) so that the model is clearly rejected. The model obtains a multitude of large modification indices, which could lead to a long model respecification search. The corresponding Bayesian analysis with noninformative priors also points to a poorly fitting model with a posterior predictive *p* value of zero. This is Model 1 in Table 16, showing the lower and upper limits of the 95% interval for the difference in the fit statistic for the real and replicated data described earlier in the Model Fit section. A positive lower limit indicates poor fit, whereas an interval with a zero fit statistic difference falling close to the middle of the interval indicates good fit. As shown later, it is useful to study the change in the 95% posterior predictive *p* value intervals when comparing alternative models. The Bayesian structural parameter estimates for Model 1 are shown in Table 17 to provide a comparison with later models. The ML estimates are very close to the Bayesian values.

A series of Bayesian analyses with informative priors is reported in Table 16 as Model 2 through Model 8. For the priors to have the intended effect, it is important to put the variables on a scale that relates to the prior choices. To this aim, the observed variables are all standardized apart from the dichotomized covariate teacher certification. Standardization is innocuous in this case because the model is scale free. Also, the metric of the two latent variables is set by fixing their residual variance at one. Because the R^2 is rather low, this metric setting creates latent variables that have variances close to one.

Model 2 in Table 16 shows the 95% posterior predictive *p* value interval for the Bayesian analysis where informative, low-variance priors are applied to the 11 zero restrictions in the structural part of the model in Figure 3. Normal priors with mean zero and variance 0.01 are used so that the 95% limits of the prior distribution are ±0.20. Because of the variable standardization, this means that the prior distributions largely contain values that are of ignorable effect size. In this sense, the essence of the original model hypothesis is maintained by a priori allowing only minor deviations from the 11 zero restrictions. Although still not well-fitting, Model 2 represents a large improvement over Model 1 in the posterior predictive *p* value interval, moving its limits to the left on the real line. The Model 2 structural parameter estimates are shown in Table 18, where bolded rows correspond to parameters included in the initial model of Figure 3, and asterisked estimates correspond to new parameters that are significant

Table 15
Bayesian Analysis Using Informative, Small-Variance Priors for Residual Correlations Using Data for British Household Panel Study Female and Male Samples, Method 2

	Female sample					Male sample				
Item	A	C	E	N	0	A	C	E	N	0
					Loadings					
*y*1	**0.772***	−0.006	−0.026	0.000	−0.012	**0.842***	−0.013	−0.011	−0.010	−0.018
*y*2	**0.575**	−0.014	0.021	−0.013	0.028	**0.394**	−0.006	0.024	−0.006	0.018
*y*3	**0.503***	0.034	0.023	0.012	−0.010	**0.479***	0.040	0.005	0.021	0.013
*y*4	−0.029	**0.704***	0.014	−0.003	0.024	−0.040	**0.683***	0.027	0.019	0.017
*y*5	0.017	**0.657***	−0.001	0.006	−0.028	0.014	**0.708***	−0.020	0.002	−0.018
*y*6	0.032	**0.548***	−0.015	−0.007	0.015	0.043	**0.579***	0.000	−0.036	0.007
*y*7	−0.008	0.014	**0.685***	0.024	0.006	−0.005	0.005	**0.748***	0.016	−0.005
*y*8	0.023	0.002	**0.702***	−0.017	0.003	0.024	0.011	**0.754***	−0.023	0.013
*y*9	−0.016	−0.021	**0.622***	−0.008	0.002	−0.025	−0.015	**0.575***	0.005	−0.005
*y*10	−0.003	0.025	0.022	**0.791***	0.023	0.001	0.009	0.013	**0.801***	0.044
*y*11	0.016	−0.007	−0.024	**0.736***	−0.008	0.012	−0.010	−0.022	**0.708***	−0.024
*y*12	−0.012	−0.022	−0.004	**0.695***	−0.027	−0.017	−0.006	0.003	**0.613***	−0.034
*y*13	0.006	0.020	0.006	−0.047	**0.780***	0.004	0.023	−0.008	0.007	**0.732***
*y*14	−0.008	−0.010	−0.007	0.046	**0.738***	0.004	−0.021	−0.013	0.023	**0.672***
*y*15	0.006	−0.006	0.011	−0.003	**0.660***	−0.011	0.001	0.031	−0.035	**0.651***
					Factor correlations					
A	—					—				
C	0.366*	—				0.319*	—			
E	0.081	0.119	—			0.025	0.197	—		
N	−0.059	−0.093	−0.163	—		−0.133	−0.238*	−0.160	—	
O	0.041	0.201	0.321*	−0.158	—	0.040	0.250	0.297*	−0.091	—

Note. The bolded entries are loadings that are the largest for the item. A = Agreeableness; C = Conscientiousness; E = Extraversion; N = Neuroticism; O = Openness.
* $p < .05$.

in the sense of their 95% posterior distribution credibility intervals not including zero. As seen in Table 18, six new parameters have significant values, four of them for the key science achievement regression.

An interesting aspect of the informative priors for Model 2 is that when the four significant parameters for the science achievement regression are estimated with noninformative priors (i.e., as completely free parameters) the posterior predictive *p* value interval and the four estimates change very little. This indicates that model fit benefits from letting the remaining 11 − 4 = 7 structural zeros have small-variance priors even when the resulting estimates are quite small. This is in line with the case of cross-loadings in the earlier factor analyses. Also, overlooking the need to include a significant and substantial effect, such as that of socioeconomic status influencing science achievement, is not important in the informative priors analysis of Model 2 because the estimate is almost the same as if the parameter had been included in the original model.

Because of the 39 degrees of freedom, the Figure 3 model implies many more restrictions beyond the structural part. The factor indicators may have direct influence from the three covariates, the factors may have cross-loadings, there may be residual correlations among the factor indicators, and the factor indicators may have direct relationships to the two dependent variables science Grade 10 and science achievement. Using informative priors, the first three sets of these restrictions are relaxed in Models 3 through 8 of Table 16, adding to the informative priors for the structural part of Model 2.

Model 3 adds small-variance (0.01) normal priors to the 18 direct effects from the three covariates to the six factor indicators. Direct effects imply differential item functioning, so that the factor indicator measurement intercepts differ for different covariate values, for example, for certified and not certified teachers. Including all direct effects in an ML analysis gives a nonidentified model, but one that is rendered identified by the Bayesian approach of small-variance priors. None of the direct effects are significant in the sense of the 95% Bayesian credibility interval containing zero, but small, nonzero estimates are obtained. This model further reduces the lower and upper limits of the posterior predictive *p* value interval relative to Model 2 but does not produce a well-fitting model. Model 4 instead adds 12 small-variance (0.01) normal priors for the cross-loadings of the factors. Again, none of the cross-loadings are significant. This does not improve posterior predictive *p* value fit as much as in Model 3. Model 5 adds small-variance priors to the residual correlations among the factor indicators. Drawing on the earlier residual correlation simulation study, Method 2 is chosen with the setting *IW*(I, 30). (Note that replacing $df = 30$ with $df = 15$ gave essentially the same results.) Out of the 15 residual correlations, as many as 11 are significant. This indicates a misspecified measurement model, although only three residual correlations obtain estimates larger than ±0.2, and the largest is 0.34. Model 5 gives a relatively large improvement of the posterior predictive *p* value interval. Model 6 and Model 7 use residual correlations and either direct effects or cross-loadings, thereby improving the posterior predictive *p* value intervals but not sufficiently for good fit. Model 7 is the first model in the sequence to obtain a positive *p* value, although it is a small value of .032. Model 8 uses all three features,

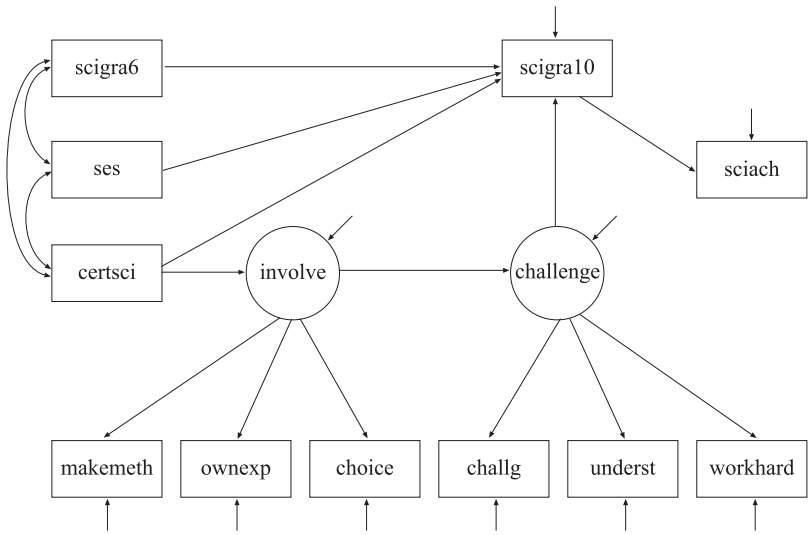

Figure 3. Structural equation model for science achievement (Kaplan, 2009). scigra6 = science grades in Grade 6; scigra10 = science grades in Grade 10; ses = socioeconomic status; sciach = science achievement; certsci = teacher certification in science; involve = perceptions of hands on involvement; challenge = students felt challenged in the classroom; makemeth = make up your own problems and work out your own methods to investigate problems; ownexp = design and conduct experiments or projects on your own; choice = make your own choice of science topic or problem to study; challg = students feel challenged in the science classroom; underst = students feel challenged to show understanding; workhard = students work hard in science class. From *Structural Equation Modeling: Foundations and Extensions* (2nd ed.), by D. Kaplan, 2009, Newbury Park, CA: Sage. Copyright 2009 by Sage Publications. Reprinted with permission.

resulting in a well-fitting model with a posterior predictive *p* value interval that includes zero and a *p* value of .276. The results for the structural parameters of Model 8 are very similar to those of Model 2 in Table 18 and are not shown.

As a technical point, the Model 2 analysis converges very quickly, needing only approximately 1,000 iterations and showing low autocorrelations. This is in contrast to Model 3 through Model 8. For these models, many more iterations are needed to ensure convergence, and the autocorrelation is high for many of the parameters. To lower the autocorrelations, thinning is used so that only every 100th iteration is included in the posterior distribution used to report estimates. As a further check on convergence, such thinning is used with 2,000, 5,000, and 10,000 iterations, yielding essentially the same results. Similar parameter estimates are obtained with and without thinning, suggesting that high autocorrelation is not necessarily a detriment to good estimation.

Table 16
Bayesian Analyses of the Science Achievement Model

Model	Informative priors	95% PP *p* interval
1	None	[1,664, 1,720]
2	Structural	[448, 505]
3	Structural, direct effects	[369, 428]
4	Structural, cross-loadings	[380, 445]
5	Structural, residual correlations	[88, 155]
6	Structural, direct effects, residual correlations	[28, 90]
7	Structural, cross-loadings, residual correlations	[−2, 64]
8	Structural, direct effects, cross-loadings, residual correlations	[−24, 44]

Note. PP *p* = posterior predictive probability.

In concluding this SEM example, one observation is that Model 8 could have been the first model in the analysis sequence. It captures the spirit of the Figure 3 model in that the zero parameters not shown in the figure are given priors that imply only small deviations from these hypothesized zeros. The hypothesis could instead be the BSEM hypothesis of almost zeros. It is also interesting to compare the use of priors for the measurement part of a model with the use of priors for the structural part. The cross-loadings and residual correlations of the two studies in this article refer to nuisance parameters and therefore appear different than the priors for structural parameters, which are of prime interest in the modeling. Specifying almost zero cross-loadings, however, does imply a hypothesis of where the major loadings are thought to appear and is, in this sense, akin to specifying almost zero structural parameters. Finally, it should be noted that the use of small-variance priors does not imply that a model is always going to be correct. The Figure 3 model has an important structural flaw in the sense that the Model 8 estimate for science achievement regressed on socioeconomic status has not only a significant estimate but an estimate that is substantial enough to be of practical interest. A somewhat less important flaw of the Figure 3 model is that the measurement model for the two factors is misspecified in the sense of needing many residual correlations. This misspecification, however, does not seem to impact the structural part of the model.

Table 17
Bayes Results for the Structural Parameters of Model 1

Relationship (slope)	Estimate	Posterior *SD*	One-tailed *p*	95% CI	
				Lower 2.5%	Upper 2.5%
Science achievement regressed on					
Science grades in Grade 10	0.383	0.011	.000	0.361	0.405
Science grades in Grade 6					
Socioeconomic status					
Teacher certification					
Students felt involved					
Students felt challenged					
Science grades in Grade 10 regressed on					
Students felt challenged	0.130	0.013	.000	0.105	0.157
Science grades in Grade 6	0.414	0.011	.000	0.392	0.435
Socioeconomic status	0.098	0.011	.000	0.076	0.119
Teacher certification	0.017	0.037	.322	−0.054	0.091
Students felt involved					
Students challenged regressed on					
Students felt involved	0.168	0.018	.000	0.133	0.202
Science grades in Grade 6					
Socioeconomic status					
Teacher certification					
Students involved regressed on					
Teacher certification	0.026	0.052	.318	−0.077	0.126
Science grades in Grade 6					
Socioeconomic status					

Note. Lines with empty entries correspond to parameters that are added in later model modifications. CI = confidence interval.

Conclusions

This article proposes a new approach to factor analysis and SEM using Bayesian analysis. The new approach represents hypotheses in a new way, replacing parameter specifications of exact zeros with approximate zeros based on informative, small-variance priors. It is argued that this produces an analysis that better reflects substantive theories. The proposed Bayesian approach with informative priors for hypothesized parameter restrictions—BSEM—is particularly beneficial in applications where if those parameters are added to a conventional model, a nonidentified model is obtained using ML. The extra model parameters can be viewed as nuisance parameters that on the basis of substantive theory and previous studies are hypothesized to be close to zero, although perhaps not exactly zero. This approach is useful for measurement aspects of latent variable modeling, such as with CFA and the measurement part of SEM. Two application areas are studied: cross-loadings in CFA and residual correlations in CFA. BSEM is also useful for the structural part of an SEM as shown in a real-data illustration. The approach encompasses three elements: model testing, model estimation, and model modification. The first two are evaluated by Monte Carlo simulation studies, whereas the third warrants further studies. The Monte Carlo and real-data results can be summarized as follows.

Summary of Findings

Model testing uses a posterior predictive probability approach that has not previously been investigated this extensively. It is found that posterior predictive probability works well both for models with only noninformative priors and for the proposed BSEM approach where some parameters have informative priors. Posterior predictive probability is found to perform better than the ML likelihood-ratio chi-square test at small sample sizes where ML typically inflates chi-square and is found to be less sensitive than ML to ignorable deviations from the correct model. Posterior predictive probability is found to have sufficient power to detect important model misspecifications.

Bayesian model estimation is shown to perform well with both noninformative and informative priors. Using BSEM with both ignorable and nonignorable degrees of model misspecification, key parameters are in most cases reasonably well estimated in terms of their coverage. BSEM outperforms ML estimation with misspecified models.

BSEM also provides a counterpart to ML-based model modification. ML modification indices inform about model improvement when a single parameter is freed and can lead to a long series of modifications. In contrast, BSEM informs about model modification when all parameters are freed and does so in a single step. The simulations show sufficient power to detect model misspecification in terms of 95% Bayesian credibility intervals not covering zero. As with ML model modification, BSEM model modification should be supported by substantive interpretability.

An example for each of the two application areas shows the promise of BSEM. For Holzinger and Swineford's (1939) example, a well-fitting factor model is found that is superior to ML-based models. Instead of choosing between an ill-fitting ML CFA model and a well-fitting but unnecessarily weakly specified ML EFA model, BSEM maintains the spirit of CFA while allowing small cross-loadings. A comparison is also made with target rotation (Asparouhov & Muthén, 2009; M. W. Browne, 2001). Target rotation is similar to BSEM in that it replaces mechanical rotation with rotation guided by the researcher's judgment, in this case using zero targets for cross-loadings. It is different from BSEM by not allowing user-specified stringency of closeness to zero by

Table 18
Bayes Results for the Structural Parameters of Model 2

Relationship (slope)	Estimate	Posterior *SD*	One-tailed *p*	95% CI	
				Lower 2.5%	Upper 2.5%
Science achievement regressed on					
Science grades in Grade 10	**0.244**	**0.012**	**.000**	**0.222**	**0.267**
Science grades in Grade 6	0.196*	0.012	.000	0.172	0.219
Socioeconomic status	0.270*	0.011	.000	0.249	0.291
Teacher certification	−0.126*	0.034	.000	−0.193	−0.059
Students felt involved	−0.102*	0.013	.000	−0.128	−0.077
Students felt challenged	−0.009	0.013	.236	−0.035	0.015
Science grades in Grade 10 regressed on					
Students felt challenged	**0.127**	**0.014**	**.000**	**0.100**	**0.153**
Science grades in Grade 6	**0.410**	**0.011**	**.000**	**0.388**	**0.433**
Socioeconomic status	**0.098**	**0.011**	**.000**	**0.077**	**0.120**
Teacher certification	**0.018**	**0.038**	**.322**	**−0.056**	**0.091**
Students felt involved	0.019	0.013	.074	−0.007	0.046
Students challenged regressed on					
Students felt involved	**0.173**	**0.018**	**.000**	**0.137**	**0.209**
Science grades in Grade 6	0.118*	0.015	.000	0.089	0.147
Socioeconomic status	0.006	0.015	.343	−0.024	0.036
Teacher certification	0.001	0.045	.495	−0.085	0.092
Students involved regressed on					
Teacher certification	**0.026**	**0.052**	**.301**	**−0.076**	**0.125**
Science grades in Grade 6	−0.012	0.015	.208	−0.042	0.018
Socioeconomic status	−0.049*	0.015	.001	−0.079	−0.018

Note. Bolded rows correspond to parameters included in the initial model of Figure 3, and asterisked estimates correspond to new parameters that are significant in the sense of their 95% posterior distribution credibility intervals not including zero. CI = confidence interval.

varying the prior variance, replacing that with least-square fitting. For Holzinger and Swineford's example, applying target rotation with zero targets for all cross-loadings gives results similar to EFA using Geomin or Quartimin rotation, except yielding more significant cross-loadings. BSEM using small-variance cross-loading priors gives far simpler loading patterns, shrinking the cross-loadings toward the prior mean. A check of the degree of shrinkage that matches the data is provided by the BSEM posterior predictive *p* value approach. Table 7 shows that for these data, the prior variance choice does not have an important impact on the results.

For the Big Five personality example, a well-fitting factor model is found that recovers the hypothesized factor loading pattern by allowing for a large number of small residual correlations. In contrast, ML CFA is ill fitting even when allowing for a priori residual correlations, and ML EFA does not recover the hypothesized factor loading pattern.

Applying BSEM is easy and fast for analyses of cross-loadings. Analysis with residual covariances leads to heavier computations because of slow MCMC convergence. A further benefit of the Bayesian analysis is that estimation works well also for models that are large relative to the sample size (see also Asparouhov & Muthén, 2010a).

Related Approaches

BSEM with its adjoining posterior predictive *p* value model test is similar in spirit to the frequentist conceptualization of *close fit* (M. W. Browne & Cudeck, 1993). ML model testing of close fit rather than conventional exact chi-square fit is expressed by the RMSEA fit index. In assessing differences between models, MacCallum, Browne, and Cai (2006) also argued against exact fit as being of limited empirical interest given that it is never true in practice. RMSEA uses an overall approximate fit level deemed sufficient. In contrast, BSEM allows informative priors to reflect notions of closeness for each parameter.

Press (2003, chapter 15) discusses a Bayesian factor analysis approach that has some similarities to the one proposed in this article. The MCMC algorithm is not used, but instead estimates are obtained as expected values in the posterior distributions. Press specified a prior for the loading matrix with a mean that uses a specific target pattern of large and zero loadings. All loadings have the same prior variances. In the example (Press, 2003, pp. 368–372), the variances are chosen to give weakly informative priors. In contrast, the current approach has zero prior means for all loadings, with small prior variances for nontarget loadings and large prior variances for target loadings so that target loadings are solely determined by the data. In this sense, Press's approach is closer to EFA, and the current approach is closer to CFA.

In BSEM, the ability to free all loadings in a measurement model can be viewed as the ability to form an EFA with the rotation guided by the priors. BSEM is, however, more general than EFA and essentially has the flexibility of exploratory SEM (ESEM; Asparouhov & Muthén, 2009) because it can accommodate correlated residuals in an EFA model, it can accommodate covariates in an EFA model, and it can accommodate an EFA model as part of a larger model. In terms of the measurement model, ESEM is exploratory in nature, while BSEM has more of a confirmatory flavor. BSEM also generalizes ESEM in the following way. In ESEM, the optimal rotation is determined only on the basis of the unrotated loadings as in EFA, that is, the optimal rotation does not consider residual covariances or covariate direct

effects in the optimal rotations. In contrast, in BSEM the optimal rotation is determined by all parts of the model.

BSEM Extensions

The BSEM ideas presented here can be extended in several ways, both for the measurement part of an SEM and for the structural part. BSEM can be extended to include equality constraints. A typical SEM example is multiple-group analysis with measurement invariance. It is common to find small deviations from exact invariance that cause rejection by the ML LRT. Group differences in measurement intercept vectors and loading matrices can be given zero-mean, small-variance priors. The special case of intercept noninvariance can be handled by letting the grouping variables be covariates that influence the factors, also referred to as MIMIC (multiple-indicators, multiple-causes) modeling (see, e.g., Muthén, 1989). Here, noninvariance is defined as direct effects from covariates to the factor indicators. With ML, including all direct effects results in a nonidentified model, whereas BSEM solves the problem using zero-mean, small-variance priors for the direct effects. This is illustrated in the reanalysis of Kaplan's (2009) science achievement model. For a study of this extension, see www.statmodel.com/examples/penn.shtml#baysem.

An extension of the residual correlations approach to categorical indicators in latent class and latent trait analysis is given in Asparouhov and Muthén (2011).

Reflections on Analysis Strategies

This article discusses several factor analysis alternatives: EFA using both mechanical and target rotation, confirmatory (ML and Bayes) factor analysis, BSEM with informative cross-loading priors, and BSEM with informative residual covariances. It is worthwhile to consider the different choices made with these different types of analyses to gain further understanding of the epistemological implications of BSEM.

One key aspect of factor analysis is the resulting factor correlations. The analysis of Holzinger and Swineford's (1939) data provides an illustration of the different factor correlation findings obtained by the different analysis alternatives. In EFA using oblique rotation, the nonzero correlations among the factors typically reduce the size of cross-loadings relative to orthogonal rotation because correlations among the factor indicators on different factors can be channeled through the factors. From the point of view of BSEM with informative cross-loading priors, the factor correlations from EFA with oblique rotation may be too low because too many nonzero cross-loadings are allowed. BSEM shrinks the cross-loadings toward their prior means of zero, and the BSEM posterior predictive p value gauges whether a certain degree of shrinking, corresponding to a certain prior variance, is compatible with the data. In Holzinger and Swineford's data the EFA factor correlations were lower than the BSEM factor correlations, and this is expected to generally be the case. EFA with target rotation did not change this picture.

ML CFA with correlated factors fixes many cross-loadings to zero so that the rotation of EFA is avoided. Because of the many cross-loadings fixed at zero, CFA tends to require higher factor correlations than EFA with oblique rotation in order to represent the correlations among the factor indicators (see also Asparouhov & Muthén, 2009; Marsh et al., 2009, 2010). From the point of view of BSEM with informative cross-loading priors, these CFA factor correlations are too high. This is because BSEM postulates cross-loadings that are not exactly zero, which in turn leads to lower BSEM factor correlations. In this sense, factor correlations from BSEM with informative priors for cross-loadings are expected to be a compromise between EFA and CFA factor correlations. This is the case for Holzinger and Swineford's (1939) data.

In the case of informative priors for residual covariances, BSEM is expected to result in smaller factor correlations than CFA with zero residual covariances given that less of the correlation among factor indicators needs to be channeled through the factors.

Given these observations, a possible strategy is to use EFA with mechanical rotation in early pilot studies of a measurement instrument until a body of knowledge about the factor indicators and the factors has been built up. Although EFA was here carried out by ML, it could also be carried out by Bayesian analysis with noninformative priors. A switch can then be made from EFA to BSEM with informative priors, where the informative priors can be chosen with smaller and smaller variances. In this sense, the Bayesian approach provides a continuum of analyses to be carried out in a series of studies, choosing priors to reflect increasing knowledge about the measurement situation. Here, ML CFA is the frequentist counterpart to the far end of the continuum. The Bayesian approach avoids the big increase in model parsimony going from an ML EFA to an ML CFA, which often leads to an ill-fitting CFA model. Similarly, it avoids the big jump in going directly to an ML CFA without preceding EFA steps as is currently often advocated, also typically leading to an ill-fitting CFA model.

A devil's advocate may argue that the BSEM approach adds *junk parameters* to permit model fit. A first response in the context of cross-loadings is that EFA potentially adds more such parameters and, unlike BSEM, does not test statistically whether they are needed. A more important response is that unlike CFA, BSEM allows the researcher to specify the degree of precision with which he or she wants to portray prior beliefs. For CFA, the only choice is what corresponds to a prior variance of exactly zero, whereas with BSEM an exact zero is not required. For models and data where the choice of prior variance makes a difference to the interpretation of the results, this informs the researcher that the data does not carry enough information on the model. A more comforting situation is illustrated in Table 7, showing ignorable dependence on the prior variance choice.

Testing model fit assumes a somewhat different form in BSEM relative to ML. First, the SEM example of science achievement shows that the BSEM approach provides for a relatively easy way to isolate the parts of a model that contribute to model misfit by applying small-variance priors to only that part. Second, even though the BSEM use of small-variance priors for key model parts often leads to well-fitting models as judged by the posterior predictive p value test of model fit, model flaws can be found by the significance of parameters that a priori were hypothesized to be ignorable.

Caveats

Several warnings are important for using BSEM. This is especially the case regarding the use of BSEM with informative priors and residual covariances. First, it may be difficult to balance the need for small residual covariances against small cross-loadings in that both

aid in representing correlations among factor indicators. Second, allowing for small residual covariances may obscure the need to add minor but still important factors. Third, medium-sized residual covariances may obscure that the postulated factor pattern is misspecified.

Furthermore, this article presents only a beginning of the study of BSEM. Much more experience is needed. For example, the posterior predictive *p* value approach to model checking needs further study. How much are the *p* values influenced by the number of variables, the number of observations, and other model features? Preliminary investigations of moderate departures from the assumed multivariate normality of the observed variables does not seem to have a critical impact, but this needs to be studied further. Another question is to which extent the maximum posterior predictive *p* value should guide which prior the results should be reported for. Although priors should be decided on before the data are analyzed, often a range of priors are equally motivated. Also, it would be worthwhile to offer several posterior predictive tests, extending posterior predictive probability beyond merely using the LRT statistic for the overall model and also focusing on a particularly important part of the model implications. Furthermore, the idea of the BSEM-derived counterpart to modification indices needs to be evaluated. It is of interest to see if this is more likely to lead to the correct model when the initial model needs several modifications. More needs to be learned about the performance of BSEM parameter posterior estimation using different informative priors for different types of models, sample sizes, and variable distributions. Hopefully, this article will stimulate such further research.

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16,* 397–438. doi:10.1080/10705510903008204

Asparouhov, T., & Muthén, B. (2010a). *Bayesian analysis of latent variable models using Mplus* (Technical report). Los Angeles, CA: Muthén & Muthén.

Asparouhov, T., & Muthén, B. (2010b). *Bayesian analysis using Mplus: Technical implementation* (Technical appendix). Los Angeles, CA: Muthén & Muthén.

Asparouhov, T., & Muthén, B. (2011). *Using Bayesian priors for more flexible latent class and latent trait analysis* (Technical report). Los Angeles, CA: Muthén & Muthén.

Barnard, J., McCulloch, R. E., & Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica, 10,* 1281–1311.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36,* 111–150. doi:10.1207/S15327906MBR3601_05

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 3,* 473–514.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *American Statistician, 46,* 167–174. doi:10.2307/2685208

Chib, S., & Greenberg, E. (1998). Bayesian analysis of multivariate probit models. *Biometrika, 85,* 347–361. doi:10.1093/biomet/85.2.347

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,* 475–487. doi:10.1037/0033-2909.115.3.475

Dominicus, A., Ripatti, S., Pedersen, N. L., & Palmgren, J. (2008). A random change point model for assessing the variability in repeated measures of cognitive function. *Statistics in Medicine, 27,* 5786–5798. doi:10.1002/sim.3380

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: Guilford Press.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association, 85,* 972–985. doi:10.2307/2289594

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Gelman, A., Meng, X. L., Stern, H. S., & Rubin, D. B. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733–807.

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

Green, P. (1996). MCMC in image analysis. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 381–399). London, United Kingdom: Chapman & Hall.

Gustafsson, J.-E. (2002). Measurement from a hierarchical point of view. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 73–95). London, United Kingdom: Erlbaum.

Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago, IL: University of Chicago Press.

Hayashi, K., & Marcoulides, G. A. (2006). Examining identification issues in factor analysis. *Structural Equation Modeling, 13,* 631–645. doi:10.1207/s15328007sem1304_7

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

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bifactor solution. Supplementary educational monographs.* Chicago, IL: University of Chicago.

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Newbury Park, CA: Sage.

Kaplan, D., & Depaoli, S. (in press). Bayesian structural equation modeling; In R. Hoyle (Ed.), *Handbook on structural equation modeling.* New York, NY: Guilford Press.

Kruschke, J. K. (2010). *Doing Bayesian data analysis: A tutorial with R and BUGS.* New York, NY: Elsevier.

Lee, S. Y. (2007). *Structural equation modeling: A Bayesian approach.* Chichester, United Kindgom: Wiley. doi:10.1002/9780470024737

Lynch, S. M. (2010). *Introduction to applied Bayesian statistics and estimation for social scientists.* New York, NY: Springer.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods, 11,* 19–35. doi:10.1037/1082-989X.11.1.19

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalizing on chance. *Psychological Bulletin, 111,* 490–504. doi:10.1037/0033-2909.111.3.490

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis.* New York, NY: Erlbaum.

Marsh, H. W., Muthén, B., Asparouhov, A., Ldtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16,* 439–476. doi:10.1080/10705510903008220

Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2012). Measurement invariance of Big-Five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology*. Advance online publication. doi:10.1037/a0026913

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557–585. doi:10.1007/BF02296397

Muthén, B., & Muthén, L. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Authors.

Press, S. J. (2003). *Subjective and objective Bayesian statistics: Principles, models, and applications* (2nd ed.). New York, NY: Wiley.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling, 11,* 424–451. doi:10.1207/s15328007sem1103_7

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London, United Kingdom: Chapman & Hall.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika, 64,* 37–52. doi:10.1007/BF02294318

Shavelson, R. J., McDonnell, L. M., & Oakes, J. (Eds.). (1989). *Indicators for monitoring mathematics and science: A sourcebook.* Santa Monica, CA: Rand Corporation.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS user manual, Version 1.4*. Cambridge, United Kingdom: MRC Biostatistics Unit.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods, 14,* 301–322. doi:10.1037/a0016972

Appendix

Obtaining the Posterior Distribution

Bayesian estimation uses Markov chain Monte Carlo (MCMC) algorithms. The idea behind MCMC is that the conditional distribution of one set of parameters given other sets can be used to make random draws of parameter values, ultimately resulting in an approximation of the joint distribution of all the parameters. For a technical discussion, see, for example, Gelman et al. (2004). For the technical implementation in Mplus, see Asparohov and Muthén (2010a). Denote by π_i a vector of unknowns consisting of parameters, latent variables, and missing observations at iteration i. The vector is divided into several sets, $\pi = (\pi_{1i}, \pi_{2i}, \ldots, \pi_{Si})'$. For example, in an application without latent variables and missing data, the parameters may be divided into means, intercepts, and slopes in one set and divided into variance and residual variances in another set. Normal priors are commonly used for the first set, whereas inverse-gamma and inverse-Wishart priors are commonly used for the second set. The conditional distribution for the first set is normal and is inverse-gamma or inverse-Wishart for the second set.

The MCMC sequence of random draws can be described as follows. Using a set of parameter starting values, new π values are obtained by the following steps over $i = 1, 2, \ldots, n$ iterations, in each step drawing from a conditional posterior parameter distribution:

$$\textit{Step } 1\text{: } \pi_{1,i}|\pi_{2,i-1}, \ldots, \pi_{S,i-1}, \textit{data}, \textit{priors} \quad \text{(A1)}$$

$$\textit{Step } 2\text{: } \pi_{2,i}|\pi_{1,i}, \pi_{3,i-1}, \ldots, \pi_{S,i-1}, \textit{data}, \textit{priors} \quad \text{(A2)}$$

$$\ldots \quad \text{(A3)}$$

$$\textit{Step } S\text{: } \pi_{S,i}|\pi_{1,i}, \ldots, \pi_{S-1,i-1}, \textit{data}, \textit{priors}. \quad \text{(A4)}$$

For Step 1 Iteration 1, the parameter values for iteration $i - 1 = 0$ are starting values. Step 1 produces values for the parameters of π_1. In Step 2 Iteration 1, those values and the starting values for the other parameters produce values for the parameters of π_2 and so on up to Step S Iteration 1. Iterations $2, \ldots, n$ go through the same steps in the same fashion. Typically, several MCMC chains are used, starting from different starting values and using different random seeds when making the random draws. The chains form independent sequences of iterations and give an opportunity to monitor convergence.

In certain cases, it is not possible to draw from the conditional posterior distributions described earlier because they do not exist in explicit form. In such cases, the Metropolis–Hastings algorithm (Gelman et al., 2004) is used instead. Suppose that in Step 1, $\pi_{1,i}$ cannot be explicitly drawn. Then $\pi^*_{1,i}$ is drawn from a distribution, J, usually referred to as the jumping distribution. The distribution J is chosen to be similar to the conditional distribution in Step 1 and to allow for explicit draws. The new draw is accepted as $\pi_{1,i}$ with probability

$$R = \frac{J(\pi_{1,i-1})}{J(\pi^*_{1,i})} \frac{P(\pi^*_{1,i}|*)}{P(\pi_{1,i-1}|*)}.$$

Otherwise $\pi_{1,i-1}$ is used as the next draw, $\pi_{1,i}$.

Assessing Convergence

In the analyses in this article, convergence is investigated in the following way. Consider n iterations in m chains, where π_{ij} is the value of parameter π in iteration i of chain j.

Define the within- and between-chain variation as

$$\bar{\pi}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} \pi_{ij}, \quad \text{(A5)}$$

$$\bar{\pi}_{\cdot\cdot} = \frac{1}{m} \sum_{j=1}^{m} \bar{\pi}_{\cdot j}, \quad \text{(A6)}$$

$$W = \frac{1}{m} \sum_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} (\pi_{ij} - \bar{\pi}_{\cdot j})^2, \quad \text{(A7)}$$

$$B = \frac{1}{m-1} \sum_{j=1}^{m} (\bar{\pi}_{\cdot j} - \bar{\pi}_{\cdot\cdot})^2. \quad \text{(A8)}$$

(*Appendix continues*)

Convergence is determined using the Gelman–Rubin convergence diagnostic (Gelman & Rubin, 1992; Gelman et al., 2004). This considers the potential scale reduction factor (PSR),

$$PSR = \sqrt{\frac{W + B}{W}}, \tag{A9}$$

where a PSR value not much larger than 1 is considered evidence of convergence. Gelman et al. (2004) suggested values of 1.1 or smaller for all parameters. This means that convergence is achieved when the between-chain variation is small relative to the within-chain variation. Gelman et al. used a slightly different definition of their potential scale reduction $\hat{R}$, but the difference relative to Equation A9 is a negligible factor of $n/(n - 1)$. It may be the case, however, that PSR convergence observed after n iterations may be negated when using more iterations. Because of this, a longer chain should be run to check that PSR values are close to 1 in a long sequence of iterations.

Priors

The density of the inverse-Wishart distribution $IW(\boldsymbol{S},\ d)$ with d degrees of freedom is given by

$$\frac{|\boldsymbol{S}|^{d/2}|\boldsymbol{X}|^{-(d+p+1)/2}\, Exp(-Tr(\boldsymbol{S}\boldsymbol{X}^{-1})/2)}{2^{dp/2}\Gamma_p(d/2)}, \tag{A10}$$

where the argument X of the density is a positive definite matrix, p is the number of variables, and Γ_p is the multivariate gamma function. To use an informative prior with a certain expected value, one can use the fact that the mean of the distribution is

$$\frac{\boldsymbol{S}}{d - p - 1}. \tag{A11}$$

The mean exists and is finite only if $d > p + 1$. If $d \leq p + 1$, then one can use the fact that the mode of the distribution is

$$\frac{\boldsymbol{S}}{d + p + 1}. \tag{A12}$$

The variance (i.e., the level of informativeness) is controlled exclusively by the parameter d. The larger the value of d, the more informative the prior is.

To evaluate the informativeness of the prior, one can consider the marginal distribution of the diagonal elements. The marginal distribution of the jth diagonal entry is the inverse-gamma distribution (Lee, 2007),

$$IG((d - p + 1)/2,\ \boldsymbol{S}_{jj}/2). \tag{A13}$$

Thus the marginal mean is

$$\frac{\boldsymbol{S}_{jj}}{d - p - 1} \tag{A14}$$

if $d > p + 1$, and the marginal variance is

$$\frac{2\boldsymbol{S}_{jj}^2}{(d - p - 1)^2\,(d - p - 3)} \tag{A15}$$

if $d > p + 3$. To use an informative prior with a certain variance, one can multiply the desired expected value by $(d - p - 1)$ to get $\boldsymbol{S}$.

The marginal distribution of the off-diagonal elements cannot be expressed in closed form, but the marginal mean for the (i, j) off-diagonal element is

$$\frac{\boldsymbol{S}_{ij}}{d - p - 1} \tag{A16}$$

if $d > p + 1$, and the marginal variance is

$$\frac{(d - p + 1)\boldsymbol{S}_{ij}^2 + (d - p - 1)\boldsymbol{S}_{ii}\boldsymbol{S}_{jj}}{(d - p)\,(d - p - 1)^2\,(d - p - 3)} \tag{A17}$$

if $d > p + 3$. As an example, using an identity matrix $\boldsymbol{S} = \boldsymbol{I}$ and $d = p + 6$ for $IW(\boldsymbol{S},\ d)$ gives a mean of zero and a variance of 0.0111 ($SD = 0.1054$).

It is clear that stating the level of informativeness using inverse-Wishart priors is rigid as the informativeness of one parameter in the matrix determines the informativeness of all other parameters. A special case is of particular interest. Setting the prior to $IW(\boldsymbol{D},\ p + 1)$, where $\boldsymbol{D}$ is a diagonal matrix, the marginal distribution for all correlations is uniform on the interval $(-1, 1)$, and the marginal distribution of the variance is $IG(1,\ d_{jj}/2)$. The values of the diagonal elements d_{jj} can be set to match the mode of the desired prior with the mode of $IG(1,\ d_{jj}/2)$, which is $d_{jj}/4$. Note, however, that the mean cannot be used for this purpose because the mean of $IG(1,\ d_{jj}/2)$ is infinity. Only the mode is defined for this distribution. In this case, the marginal distribution of the diagonal elements has infinite mean and variance. The marginal for the covariance elements has a mean of zero by symmetry but also has an infinite variance. The marginal mean for the correlation parameter is zero, and the marginal variance for the correlation parameter is 1/3.

More generally, setting the prior to $IW(\boldsymbol{D},\ d)$, where $\boldsymbol{D}$ is a diagonal matrix, the marginal distribution for all correlations is the beta distribution $B[(d - p + 1)/2,\ (d - p + 1)/2]$ on the interval $(-1, 1)$, if $d \geq p$ with a mean of 0 and a variance of

$$\frac{1}{d - p + 2}. \tag{A18}$$

Note also that the posterior distribution in the MCMC generation for the variance covariance parameter with prior $IW(\boldsymbol{S},\ d)$ is a weighted average of $\boldsymbol{S}/d$ and the sample variance where the weights are $d/(n + d)$ and $n/(n + d)$, respectively, where n is the sample size. Thus, one can interpret the degrees of freedom parameter d as the number of observations added to the sample with the prior variance covariance matrix. Naturally, as the sample size increases the weight $d/(n + d)$ will converge to 0, and the effect of the prior matrix $\boldsymbol{S}$ will diminish. To maintain the same effect of the prior on the estimation for larger sample sizes, the degrees of freedom parameter should be chosen proportionally larger.

More information on the inverse-Wishart distribution and the marginal distributions of all the entries in the matrix can be found in Barnard, McCulloch, and Meng (2000). See also Gelman et al. (2004).

Received September 28, 2010
Revision received August 15, 2011
Accepted August 16, 2011 ■