A note on identification constraints and information criteria in Bayesian latent variable
models

Benjamin Graves & Edgar C. Merkle

University of Missouri

Author Note

This is a preprint of an article published in Behavior Research Methods. The final
authenticated version is available online at:

<https://doi.org/10.3758/s13428-021-01649-8>. Correspondence to Benjamin Graves,
Department of Psychological Sciences, 210 McAlester Hall, 320 S 6th Street, Columbia,
MO 65211. Email: blgpp5@mail.missouri.edu.

Abstract

It is well known that, in traditional SEM applications, a scale must be set for each latent variable: typically, either the latent variance or a factor loading is fixed to one. While this has no impact on the fit metrics in ML estimation, it can potentially lead to varying Bayesian model comparison metrics due to the use of different prior distributions under each parameterization. This is a problem, because a researcher could artificially improve one's preferred model simply by changing the identification constraint. Using a single-factor CFA as motivation for study, we first show that Bayesian model comparison metrics can systematically change depending on constraints used. We then study principled methods for setting the scale of the latent variable that stabilize the model comparison metrics. These methods involve (i) the placement of priors on ratios of factor loadings, as opposed to individual loadings; and (ii) use of effect coding. We illustrate the methods via simulation and application.

Keywords: Bayesian SEM, structural equation models, model comparison, identification, *blavaan*

A note on identification constraints and information criteria in Bayesian latent variable models

Introduction

Bayesian psychometric models have gained popularity due to their accessibility, their application to small data sets, and their ability to handle complex models that would prove difficult to estimate using maximum likelihood methods (e.g., Fox, 2010; Kaplan, 2014; Lee, 2007; Lee & Song, 2004; Levy & Mislevy, 2016; Muthén & Asparouhov, 2012; Rupp, Dey, & Zumbo, 2004; Song & Lee, 2012; van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017). These models generally involve latent variables representing person traits or attributes, requiring a constraint to be placed on the latent variable scale in order to identify other model parameters. This is typically done in one of two ways. The first constrains a single factor loading per latent variable to a constant, usually 1. This is sometimes called the fixed marker method or unit loading identification (ULI; Kline, 2015). The second constrains the latent variance to be a constant, which is again usually 1. This is known as the fixed factor method or unit variance identification (UVI). Further detail on traditional factor model identification appears in Jöreskog (1969, 1979) and in Peeters (2012a, 2012b).

When using traditional estimation methods for SEM, including maximum likelihood and least squares estimation, the specific identification constraint typically does not influence model fit or model comparison metrics (though see, e.g., Klößner & Klopp, 2019; O'Brien & Reilly, 1995; Steiger, 2002, for exceptions). This is not necessarily true for Bayesian models. For Bayesian models, the prior distributions often differ across identification constraints, which can in turn influence the fit metrics. For example, fixing the first factor loading to one may lead to a better-fitting model, as compared to fixing the second factor loading to one. Further, if priors are centered around zero, they suggest that free loadings are likely to be smaller than the loading that is fixed to one. Unlike in the frequentist setting, this change in priors across identification methods implies that the

models are distinct from each other even though they have the same structure. This is potentially problematic because it provides a route by which researchers can artificially improve or worsen a particular model’s metrics. For example, one can imagine a hypothetical situation where a researcher’s preferred model is not selected as best by an information criterion (say, by DIC). The researcher may then change the identification constraint, find that their preferred model is selected, and reports results under the new identification constraints. This issue is unique to Bayesian latent variable models and, to our knowledge, has not received attention in previous literature.

The purpose of this paper is to study how model identification constraints influence Bayesian model comparison criteria. We consider the popular Bayesian metrics of DIC (Spiegelhalter, Best, Carlin, & van der Linde, 2002), WAIC (Watanabe, 2010), and the Pareto-smoothed leave-one-out cross-validation metric (LOOIC; e.g. Vehtari, Gelman, & Gabry, 2017) that is implemented in the *loo* package for R (Vehtari et al., 2020). These metrics are defined as as

$$\text{DIC} = -2 \log L(\hat{\boldsymbol{\vartheta}}|\mathbf{Y}) + 2 \left[\log L(\hat{\boldsymbol{\vartheta}}|\mathbf{Y}) - \overline{\log L(\hat{\boldsymbol{\vartheta}}|\mathbf{Y})} \right] \quad (1)$$

$$\text{WAIC} = -2 \sum_{i=1}^n \log \left(\frac{1}{S} \sum_{s=1}^S f(\mathbf{y}_i|\boldsymbol{\vartheta}^s) \right) + 2 \sum_{i=1}^n \text{var}_s(\log f(\mathbf{y}_i|\boldsymbol{\vartheta})) \quad (2)$$

$$\text{LOOIC} = -2 \sum_{i=1}^n \log \left(\frac{\sum_{s=1}^S w_i^s f(\mathbf{y}_i|\boldsymbol{\vartheta}^s)}{\sum_{s=1}^S w_i^s} \right) \quad (3)$$

where $\hat{\boldsymbol{\vartheta}}$ is a vector of posterior parameter means, $L(\hat{\boldsymbol{\vartheta}}|\mathbf{Y})$ is the likelihood at those means, S is the number of posterior draws, $f(\mathbf{y}_i|\boldsymbol{\vartheta}^s)$ is the density of observation i w.r.t. sampled parameters at iteration s , and the w_i^s are importance sampling weights. We use a likelihood that is marginal over the latent variables in the model; see Merkle, Furr, and Rabe-Hesketh (2019) for further detail on this issue.

In the sections below, we first illustrate the fact that the information criteria mentioned above can exhibit different values, depending on the identification constraint used. In essence, different identification constraints can lead to different prior distributions, which in turn lead to different Bayesian models. We then explore the issue in more detail

via simulation and describe alternative identification constraints that can help resolve the issue. Finally, we describe an application and discuss general implications.

Illustration

We first illustrate the problem in the context of Bayesian CFA. We use a subset of the Holzinger and Swineford (1939) data available in the *MBESS* package (Kelley, 2018), focusing on the three spatial ability tests. We fit a one-factor model (see Figure 1) with four different identification constraints: ULI on each of the three loadings as well as UVI. Under traditional estimation methods, these four models exhibit exactly the same model fit metrics (χ^2 statistics, AICs, BICs, RMSEAs, etc). Thus, if we find that the models systematically differ on Bayesian model comparison metrics, it is notable because it means that a model could be artificially improved or worsened by choice of identification constraint. Note that we are not comparing the four models in order to select the model with the “best” identification constraint; we are instead illustrating that the four models differ in the values of their information criteria.

We estimated the one-factor model via three MCMC chains, with chain lengths being automatically determined by the Raftery and Lewis (1995) procedure that is implemented in R package *runjags* (Denwood, 2016). After burning in and sampling the model for a pre-specified length, the procedure determines the number of additional samples necessary to estimate each parameter’s 2.5th quantile to an accuracy of .005 with probability .95. Models were estimated and summarized via *blavaan* (Merkle & Rosseel, 2018), which provides an interface between *lavaan* model specification (Rosseel, 2012) and the *JAGS* software for MCMC (Plummer et al., 2003). Factor loading priors were set to normal with mean 0 and standard deviation of 10, and the latent precision prior was $\text{Gamma}(1, .5)$. Under the UVI constraints, it is possible for the sampled factor loadings to flip signs from positive to negative and back again. To address this issue, a sign constraint was placed on a single loading per factor, as discussed by Peeters (2012b). This is generally preferable to

half-normal or log-normal priors on all loadings, because those priors do not allow for reverse-coded variables or for loadings whose posterior distribution is symmetric around 0. Other priors were taken to be *blavaan* defaults, which are intended to be weakly informative. Manifest variable intercepts (ν) were set to a Normal with a mean of 0 and standard deviation of 32 and latent variable intercepts (α) are set to be a Normal with a mean of 0 and standard deviation of 10.

Results are shown in Table 1. Choice of identification constraint is presented in the first column, followed by the values of the marginal log-likelihood (LogL), WAIC, DIC, and LOOIC. We see up to 6-point differences in the values of these metrics, depending on the identification constraint employed. It is apparent that the fit metrics can differ depending on how the model is identified. If we were to explicitly compute uncertainty in these metrics, then these differences may not be large (and see the General Discussion for further remarks here). But uncertainty is not commonly considered here (at least, in Bayesian SEM), and the differing value would not be expected by researchers who have a background in frequentist SEM.

Simulation

The above results were obtained with a single data set and could therefore be due to noise. This section expands the illustration, to examine the generality of the results. While we focus on Bayesian information criteria, the results have implications for other types of fit metrics. The simulation compared a one-factor model with the latent variance fixed to 1 (UVI model) to a one-factor model with the first loading fixed to 1 (ULI model), and it specifically focused on how the identification constraint influences the information criteria.

Procedure

Data were generated via the model from the illustration section, with parameter values fixed to the ML estimates from the original Holzinger and Swineford (1939) data. The observed variables x_1 (“Visual”), x_2 (“Cubes”), and x_3 (“Lozenges”) have variances of

26.77, 13.01, 30.93, factor loadings of 4.92, 2.96, and 5.96, and intercepts of 29.32, 24.70, 14.84, respectively, with the latent variance being fixed to 1. To examine systematic differences in model comparison metrics, we fit two models to each simulated data set: one with the first factor loading fixed to one and the other with the latent variance fixed to 1 (as in the previous section). These models' metrics would not differ under traditional ML estimation methods.

We used two sets of priors that varied in informativeness. The less-informative set of priors used the default settings in *blavaan* in which the factor loading prior (λ) was set to be Normal(0, 10), where the mean is 0 and standard deviation is 10. The latent variable precision (ψ^{-1}) was set to a Gamma(1, .5), which has a mean of 2 and standard deviation of 2. The more informative set of priors involved Normal(0, 1) distributions on factor loadings and a Gamma(.5, 2) distribution on the latent precision. This gamma distribution has a mean of 0.25 and standard deviation of 0.354. Figure 2 shows the distributions of the priors and reference points for ML parameter estimates. We also manipulated sample size, generating data sets with 500 and 1,000 observations.

Each possible combination of priors and sample sizes was considered, leading to a total of 8 possible conditions (2 loading priors $\times$ 2 latent precision priors $\times$ 2 sample sizes). For each simulation condition, we generated 500 data sets, estimated the UVI and ULI models via MCMC, and recorded the model assessment metrics. Primary metrics of interest were each model's DIC, WAIC, LOOIC, and marginal log-likelihood (LogL). For each metric, we also recorded model preference by computing the proportion of time the UVI model had a better value than the corresponding ULI model. If the fit metrics are the same across identification constraints, then this proportion should be close to .5. Values above .5 suggest the UVI model is preferred, while values below .5 suggest the ULI model is preferred.

Results

Primary results are shown in Figure 3, which are split by sample size. The priors used are on the x-axis, model preference (proportion of replications favoring UVI) is on the y-axis, and the bars represent the various fit metrics. In the figures, we can see that there is a systematic preference for one model identification method over the other, depending on the prior specification used and on the sample size. When uninformative priors are used for the factor loadings ($\text{Normal}(0, 10)$), the UVI method was preferred across all of the fit metrics. Use of informative priors for the factor loadings ($\text{Normal}(0, 1)$) changed model preference to favor the ULI method across all fit metrics. The lack of appearance of bars in the figure is due to an overwhelming preference for the ULI models, where nearly 100% of the replications were favoring this identification method. Changes in the latent variance prior had little effect on model preference, but preference values did get closer to .5 when sample size was increased from 500 to 1000. This is because the influence of the prior weakens as sample size increases. Of the fit metrics, marginal LogL preferred one model over the other nearly 100% of the time depending on how the priors were specified.

Mean differences between the fit measures were also calculated for each simulation condition. The results are presented in Table 2. The conditions that showed the least preference also had the smallest mean differences, and these differences decreased as sample size was increased. The LogL exhibited smaller mean differences than the other metrics, which can be explained by the fact that the other metrics involve the LogL multiplied by -2 . The informative $\text{Normal}(0,1)$ prior on loadings also led to increased differences in metrics and favoring of the ULI model, whereas the UVI model was favored under the noninformative priors.

The results observed here can be explained by the extent to which the priors agree with the likelihood. To illustrate this, the asymptotic maximum likelihood estimates for our models (from *lavaan*) appear in Table 3. These estimates come from a simulated data set of 30,000 observations in order to approximate a “practically infinite” number of

observations. The rows show which identification method is used for the model, either UVI or ULI, and the columns show the maximum likelihood estimates for the factor loadings and latent variance. In this example, it is worth noting that changing the constrained parameter changes the values of the parameter estimates, but the relationships between the factor loadings are preserved (i.e. $\lambda_2 < \lambda_1 < \lambda_3$). This table shows that the estimated loadings under UVI are inconsistent with our informative prior distribution (Normal(0,1)), which explains why the UVI model was not preferred under that prior distribution.

To further illustrate the influence of the priors on factor loadings, we fit another model with UVI constraints to the original dataset. This model was similar to the illustration, except that the factor loading priors were normal with mean 4.5 and standard deviation of 1. These priors are informative and also do a good job of capturing the maximum likelihood estimates. Information criteria values for this model can be found in Table 4 and can be compared to those in Table 1. We see that, by all three information criteria, this model now appears better than all the previous models from Table 1.

In sum, we have observed that identification constraints can lead to differing information criteria, with the differences being explained by the prior specification and sample size. This means that a Bayesian model's metrics can potentially be improved by changing the identification constraint, something that has no impact under other estimation methods and that is often ignored in Bayesian SEM estimation. While we are unaware of researchers actually doing this, awareness of the issue is worthwhile.

Alternative Methods for Setting the Scale

The previous sections have illustrated the fact that Bayesian model comparison metrics can change depending on how a researcher sets the scale of the latent variables. In this section, we describe two methods for setting the scales of latent variables such that the issue is reduced or removed. The first method places priors on ratios of factor loadings, so that the specific loading fixed to 1 has less influence on model comparison metrics. The

second method identifies the model via effect coding, so that the user does not have to constrain any individual parameter. Example code for the implementation of these methods can be found at <https://osf.io/qvryj/>.

Ratio Method

Our first method involves maintaining the same identification constraint (e.g., fixing a single parameter to one), with prior distributions being placed on ratios of factor loadings instead of on individual factor loadings. In particular, Newsom (2015) describes the relationship between loadings estimated under the ULI method to loadings estimated under the UVI method. Multiplying a factor loading by the square root of the variance parameter under UVI yields the estimate of the loading under ULI. Conversely, the estimated factor loading under ULI is the ratio of the same loading and the referent indicator value under UVI. With these sorts of results, we can place priors on the ratios of loadings instead of the individual loadings, and then extract the loadings from the ratios, depending on the identification method used.

Referring to the one-factor, three-indicator model from Figure 1, we can use prior distributions similar to below, regardless of whether ULI or UVI is employed:

$$\frac{1}{\psi} \sim \text{Gamma}(1, .5) \quad (4)$$

$$r_1 = \frac{\lambda_1}{\lambda_2} \sim \text{N}(0, 10) \quad (5)$$

$$r_2 = \frac{\lambda_1}{\lambda_3} \sim \text{N}(0, 10). \quad (6)$$

The ratios r_1 and r_2 have priors centered at 0, instead of 1, to allow for the possibility that the individual loadings are negative as well as positive. If we expected all loadings to be positive, then a prior mean of 1 would be sensible.

We can now obtain parameter estimates under either ULI or UVI using the above

ratios. For example, say that we fix λ_1 to one. The remaining factor loadings are:

$$\lambda_1 = 1 \tag{7}$$

$$\lambda_2 = \frac{1}{r_1} \tag{8}$$

$$\lambda_3 = \frac{1}{r_2}. \tag{9}$$

Alternatively, if we fix the latent variance to one, the factor loadings can be obtained as:

$$\lambda_1 = \sqrt{\psi} \tag{10}$$

$$\lambda_2 = \frac{\sqrt{\psi}}{r_1} \tag{11}$$

$$\lambda_3 = \frac{\sqrt{\psi}}{r_2}, \tag{12}$$

where λ_1 is fixed to $\sqrt{\psi}$ so that the model-implied variance of the first indicator is the same as it would be if λ_1 had been fixed to one. This implies that λ_1 can only take positive values, which can be overly restrictive for a factor loading. However, such a sign constraint is necessary for the model to achieve global parameter identification under Bayesian estimation with UVI (Peeters, 2012b), so it serves a specific purpose here.

To study these ratio prior distributions, we conducted another simulation similar to the previous simulation, except using the ratio priors described above. We estimated four models with unique identification constraints: a model with each factor loading separately fixed to 1, and a fourth model with the latent variance fixed to 1. Model preferences were, once again, calculated by comparing the metrics of the UVI model to each of the ULI models. The simulation used the same set of priors specified above and sample sizes of 500 and 1000. Within each condition, we carried out 500 replications.

Model preference values for comparing each of the ULI models to the UVI model are found in Figure 4. Overall, as compared to initial simulations, models were not preferred as strongly as they were in the earlier simulations. For each of the combinations of priors and sample sizes, model preferences using DIC, WAIC, and LOOIC tended to approach .5 and were closest to exhibiting no preference at the largest sample size of 1000. Model

preference values varied slightly depending on which factor loading was fixed to 1, but were generally similar across choice of fixed loading.

As can be seen, placing priors on ratios of parameters can help to reduce differences in information criteria, so that choice of identification constraint matters less in practice. But this is not a full solution because, at sample sizes smaller than 500 (e.g., 100), the models with priors on ratios of loadings did not reliably converge for us. Perhaps a more reliable MCMC algorithm could be developed, but we did not pursue it here. Additionally, priors on ratios of factor loadings are not widely implemented in any software packages (they are not straightforward to implement), and the ratios can make it difficult for applied researchers to set informative priors.

Effect Coding Method

Instead of placing *priors* on functions (ratios) of parameters, we can place identification constraints on functions of parameters so that no individual parameter is singled out as the fixed parameter. The Little, Slegers, and Card (2006) effect coding method is well suited to this task. Under this method, the latent variable mean is estimated, and the model is identified via

$$\frac{\sum_{i=1}^I \lambda_i}{I} = 1 \quad (13)$$

$$\sum_{i=1}^I \nu_i = 0, \quad (14)$$

where I is the number of unique indicators for a latent variable and the ν_i are the indicator intercepts (see Figure 1). These equations state that factor loadings are constrained to average 1, and the intercepts become deviations from the grand mean of observed variables (the latent mean becomes the grand mean of observed variables). Referring specifically to Figure 1, we can use the constraints $\lambda_1 = 3 - \lambda_2 - \lambda_3$ and $\nu_1 = 0 - \nu_2 - \nu_3$. A similar strategy has been considered in Bayesian item response modeling (e.g., Fox & Glas, 2001), where the sum of difficulties is restricted to equal 0 and the product (as opposed to sum) of discriminations is restricted to equal 1.

Under effect coding, there are no arbitrary decisions about which parameter should be fixed to one. Thus, if a researcher were to consistently use effect coding, there are no issues with model comparison metrics systematically differing by choice of constraint. And implementation is straightforward, because packages *lavaan* and *blavaan* both allow the user to specify effect coding as the scaling option for latent variables, with default prior distributions being the same as they would be otherwise (with users being encouraged to set their own prior distributions).

However, effect coding may be unintuitive in situations where an observed variable is influenced by multiple latent variables (i.e., in a model with cross-loadings) and is not the ideal identification method here. In the case of cross-loadings, we can still specify the constraints from the previous paragraph separately for each latent variable. But now an observed variable can be involved in the constraints associated with multiple latent variables, making the parameter interpretation cloudy. For example, an observed variable's intercept could be constrained to sum to 0 with other intercepts associated with one latent variable, and that same intercept could be separately constrained to sum to 0 with other intercepts associated with a second latent variable. This potentially leads to indirect dependencies between observed variable intercepts that were specified to be conditionally independent; see Little et al. (2006) for some further discussion of this issue.

Example

To illustrate the identification methods' applications, we return to the Holzinger and Swineford (1939) data used in our original illustration. For the ratio priors, we estimate both the UVI and three ULI models, with the ULI models corresponding to each loading being fixed to 1. We also estimate the same model under effect coding, with the Raftery and Lewis (1995) automatic convergence method being used to determine the length of the chains (with three chains being used). Because the data set used for the illustration and example contains only 73 observations, we used informative prior distributions to aid in

convergence. We used $\text{Normal}(1,1)$ priors on ratios of loadings, and we used $\text{Gamma}(.3, .5)$ priors on precisions (ψ^{-1} and θ^{-1}).

Fit metrics for all models are presented in Table 5. The rows of the table denote how the model was identified and where the “1” was placed. The first three rows use ULI in conjunction with placing priors on ratios of factor loadings, the fourth row uses UVI with priors on ratios of factor loadings, and the last row uses effect coding for identification. Looking down each column, fit metric values still exhibit variability, though not as much as in the earlier illustration. Fixing the “Cubes” factor loading to 1 produces larger fit metrics than the other constraints, likely because this is the smallest loading (see Table 3) and the ratio prior disagrees more here, as compared to models where other loadings are fixed to 1.

General Discussion

In this paper, we highlighted the fact that model identification constraints can influence Bayesian model comparison metrics in ways that would not be observed under frequentist modeling. This is potentially a problem because model identification constraints are often chosen arbitrarily, which can lead to a Bayesian model’s metrics becoming artificially better or worse depending on choice of constraint. This could be viewed as an extra “researcher degree of freedom” (Simmons, Nelson, & Simonsohn, 2011) when conducting applied analyses: if a researcher’s preferred Bayesian model does not look the best under one identification constraint, the researcher could potentially choose an alternative identification constraint to influence the results.

We considered two solutions to the issue. The first involved identifying the model in the usual way (with a loading or latent variance fixed to 1), with priors being placed on ratios of factor loadings instead of individual loadings. While this solution is general, it might be difficult to conceptualize informative priors under this approach, the approach is not generally implemented in any software of which we are aware, and MCMC convergence is not optimal at small sample sizes. The second solution involved effect coding, which

avoids the need to arbitrarily fix a single parameter to one. We generally recommend the use of effect coding where applicable in Bayesian SEM, both because the parameter interpretations are intuitive and because effect coding can be easily handled in software such as *blavaan*. But effect coding becomes cumbersome in models with cross-loadings. Additionally, in any application, a researcher could potentially choose between effect coding, priors on ratios of loadings, or traditional priors and model parameterizations, reporting only the model that has the “best” values as compared to other candidate models. So awareness of the issues discussed here is important, paired with sensitivity analyses.

Some other methods related to parameter identification should be mentioned here. The first is the parameter expansion approach of Ghosh and Dunson (2009). Under this approach, we sample from an overparameterized model where the scale of the latent variable is not set. We then impose an identification constraint as a post-processing step, leading to posterior chains that have converged. While we would set prior distributions for all parameters of the overparameterized model, the implications for the priors of the post-processed, inferential parameters may still depend on the identification constraint chosen. The second approach involves building priors around the constrained parameters, as described by Levy and Mislevy (2016). For example, if we fix a factor loading to 1 for identification, the priors for the other loadings could have a mean of 1 instead of 0, stating that the other loadings are equally likely to be larger or smaller than the constrained loading. We may still encounter problems in situations where a single loading is considerably larger than the others. If we fix that large loading to one, then our priors centered at 1 may still lead to worse information criteria values, as compared to a model where we fix a different loading to one.

Recently, new Bayesian metrics of absolute fit, including BCFI, BTLI, and BRMSEA, have been studied by Garnier-Villareal and Jorgensen (2019) and are implemented in *blavaan*. They are designed to approximate the traditional model fit indices, and they can also be expected to vary by choice of identification constraint because they generally

involve the log-likelihood (which was examined in this paper). Thus, in addition to model comparison, the issues described here could influence a researcher’s assessment of absolute fit: researchers could, e.g., improve a model’s BRMSEA value by changing the identification constraint. Further research may be worthwhile here, especially in the context of the rules of thumb associated with these indices.

We also note that the identification/prior issues considered in this paper may become unimportant if one is willing to explicitly consider the uncertainty associated with model selection criteria. Vehtari, Simpson, Yao, and Gelman (2019), McElreath (2015), and others discuss the use of standard errors associated with WAIC and related metrics, pointing out that it is often inappropriate to select a single model when the differences between information criteria are small relative to the standard errors of the differences. Preacher and Merkle (2012) considered related issues in the context of maximum likelihood SEM, focusing heavily on BIC. As applied to the current problem, the differences in information criteria across identification constraints will usually be small relative to the posterior uncertainty in the criteria. This implies that, if we consider posterior uncertainty, then the priors and identification constraints examined in this paper will often lead to similar inferences. But these uncertainty estimates can be influenced by model predictions that are too similar, are misspecified models with outliers, or have small sample sizes, and their calculation is not as simple as computing the variance of posterior samples (Sivula, Magnusson, & Vehtari, 2020). Thus, uncertainty estimates for Bayesian information criteria are not always available in popular software, leading many SEM researchers to use point estimates of information criteria to select a single model from a set of candidates.

Computational Details

All results were obtained using the R system for statistical computing (R Core Team, 2018), version 4.1.0, employing the add-on packages *blavaan* (Merkle & Rosseel, 2018), *runjags* (Denwood, 2016), *loo* (Vehtari et al., 2020), and *ggplot2* (Wickham, 2016) are

freely available under the General Public License 2 from the Comprehensive R Archive Network at <http://CRAN.R-project.org/>. R code for replication of our results is available at <https://osf.io/qvryj/>.

References

- Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, 71(9), 1–25. doi: 10.18637/jss.v071.i09
- Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. NY: Springer.
- Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66, 271–288.
- Garnier-Villareal, M., & Jorgensen, T. D. (2019). Adapting fit indices for bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*.
- Ghosh, J., & Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2), 306–320.
- Holzinger, K. J., & Swineford, F. (1939). A study in factor analysis: The stability of a bi-factor solution. *Supplementary Educational Monographs*.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183–202.
- Jöreskog, K. G. (1979). Author’s addendum. In J. Magidson (Ed.), *Advances in factor analysis and structural equation models* (pp. 40–43). Cambridge: Abt Books.
- Kaplan, D. (2014). *Bayesian statistics for the social sciences*. NY: The Guildford Press.
- Kelley, K. (2018). MBESS: The MBESS R package [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=MBESS> (R package version 4.4.3)
- Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.
- Klößner, S., & Klopp, E. (2019). Explaining constraint interaction: How to interpret

- estimated model parameters under alternative scaling methods. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(1), 143–155.
- Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. Chichester, England: John Wiley & Sons.
- Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, 39(4), 653–686.
- Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Boca Raton, FL: Chapman & Hall.
- Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in sem and macs models. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(1), 59–72. doi: 10.1207/s15328007sem1301_3
- McElreath, R. (2015). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman & Hall/CRC.
- Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, 84, 802–829.
- Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software, Articles*, 85(4), 1–30. Retrieved from <https://www.jstatsoft.org/v085/i04> doi: 10.18637/jss.v085.i04
- Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological methods*, 17(3), 313.
- Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. Routledge.
- O’Brien, R. M., & Reilly, T. (1995). Equality in constraints and metric-setting measurement models. *Structural Equation Modeling: A Multidisciplinary Journal*, 2(1), 1–12.
- Peeters, C. F. W. (2012a). *Bayesian exploratory and confirmatory factor analysis:*

- Perspectives on constrained-model selection* (Unpublished doctoral dissertation).
Utrecht University.
- Peeters, C. F. W. (2012b). Rotational uniqueness conditions under oblique factor correlation metric. *Psychometrika*, 77, 288–292.
- Plummer, M., et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124).
- Preacher, K. J., & Merkle, E. C. (2012). The problem of model selection uncertainty in structural equation modeling. *Psychological Methods*, 17, 1–14. doi: 10.1037/a0026804
- R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Raftery, A., & Lewis, S. (1995). *The number of iterations, convergence diagnostics, and generic metropolis algorithms*. Chapman and Hall, London.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. Retrieved from <http://www.jstatsoft.org/v48/i02/>
- 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(3), 424–451.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. doi: 10.1177/0956797611417632
- Sivula, T., Magnusson, M., & Vehtari, A. (2020). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv preprint arXiv:2008.10296*.
- Song, X.-Y., & Lee, S.-Y. (2012). *Basic and advanced Bayesian structural equation*

- modeling: With applications in the medical and behavioral sciences*. Chichester, UK: Wiley.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, 64, 583–639.
- Steiger, J. H. (2002). When constraints interact: A caution about reference variables, identification constraints, and scale dependencies in structural equation modeling. *Psychological Methods*, 7(2), 210–227.
- van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, 22, 217–239.
- Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2020). *loo: Efficient leave-one-out cross-validation and waic for bayesian models*. Retrieved from <https://mc-stan.org/loo/> (R package version 2.4.1)
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432.
- Vehtari, A., Simpson, D. P., Yao, Y., & Gelman, A. (2019). Limitations of “limitations of bayesian leave-one-out cross-validation for model selection”. *Computational Brain & Behavior*, 2(1), 22–27.
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571–3594.
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from <http://ggplot2.org>

Table 1

Fit metrics under various choices of model identification constraints.

	LogL	DIC	WAIC	LOOIC
UVI	-687.68	1396.69	1397.87	1398.06
ULI: Visual	-693.39	1393.70	1401.89	1402.14
ULI: Cubes	-693.34	1394.95	1403.63	1403.72
ULI: Lozenges	-691.59	1395.03	1400.81	1401.04

Table 2

Simulation Mean Differences. The first three columns contain the sample sizes and prior specifications for each of the conditions. The last four columns present the average difference in fit metric values when subtracting the ULI model from the UVI model. Negative values suggest a preference for the UVI method.

Size	Lambda	iPsi	LogL	DIC	WAIC	LOOIC
500	N(0,10)	Gamma(1,.5)	-0.08	-0.13	-0.20	-0.20
1000	N(0,10)	Gamma(1,.5)	-0.04	-0.06	-0.09	-0.09
500	N(0,1)	Gamma(1,.5)	1.12	2.07	2.08	2.08
1000	N(0,1)	Gamma(1,.5)	0.61	1.11	1.11	1.11
500	N(0,10)	Gamma(.5,2)	-0.06	-0.06	-0.12	-0.12
1000	N(0,10)	Gamma(.5,2)	-0.03	-0.03	-0.06	-0.06
500	N(0,1)	Gamma(.5,2)	1.16	2.17	2.18	2.18
1000	N(0,1)	Gamma(.5,2)	0.62	1.15	1.16	1.16

Table 3

Asymptotic marginal maximum likelihood parameter estimates. The table below shows parameter estimates for models identified using traditional ULI and UVI methods and maximum likelihood estimations. The first column shows which parameter was set to 1, while the other columns show parameter estimates and standard errors.

Fixed Parameter	Psi	Lambda1	Lambda2	Lambda3
UVI: Latent Variance	1 (0)	4.91 (0.05)	2.96 (0.03)	6 (0.05)
ULI: Visual	24.13 (0.45)	1 (0)	0.6 (0.01)	1.22 (0.02)
ULI: Cubes	8.77 (0.17)	1.66 (0.02)	1 (0)	2.02 (0.03)
ULI: Lozenges	35.96 (0.64)	0.82 (0.01)	0.49 (0.01)	1 (0)

Table 4

Fit metric values for model with mean shifted, informative priors.

	Value
Log-likelihood	-688.46
DIC	1393.13
WAIC	1394.40
LOOIC	1394.54

Table 5

Fit metric values when applying the ratio identification and effect coding to the example spatial ability data. Identification method is presented in the first column, followed by each of the selected metrics.

	LogL	DIC	WAIC	LOOIC
ULI: Visual	-687.82	1394.11	1395.45	1395.61
ULI: Cubes	-687.95	1395.96	1396.83	1397.01
ULI: Lozenges	-687.84	1393.53	1394.84	1395.00
UVI: Latent Variance	-687.73	1394.39	1395.59	1395.76
Effect Coding	-687.77	1393.94	1395.05	1395.20

Figure 1. General path diagram used to fit the example models and generate simulation data.

Figure 2. Distributions of priors used for simulation. The points on the graphs show the parameter estimates from Holzinger and Swineford (1939) for the factor loadings and latent variance.

Figure 3. Model Preference Values for the simulation with a sample size of 500 in the top graph and 1000 in the bottom. Values closer to 1 suggest the UVI model is preferred, values closer to 0 suggest the ULI model is preferred, and values around .5 suggest no model preference.

Figure 4. Model preference comparing each ULI model to the UVI model using using ratio identification method. The headings on each graph show which factor loading was set to 1 for ULI.