METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Nuances of Information Criteria for Bayesian Psychometric Models

Edgar C. Merkle [1]

[1] *Department of Psychological Sciences, University of Missouri, Columbia, MO, USA.*

Methodology, 2026, Vol. 22(3), 195–224, https://doi.org/10.5964/meth.20361

Received: 2025-10-16 • **Accepted:** 2026-03-10 • **Published (VoR):** 2026-08-31

Handling Editor: Pablo Nájera Álvarez, Universidad Pontificia Comillas, Madrid, Spain

Corresponding Author: Edgar C. Merkle, 219 McAlester Hall, Columbia, MO 65211, USA. E-mail: merklee@missouri.edu

Supplementary Materials: Code, Data [see Index of Supplementary Materials]

Abstract

It is common practice to compare Bayesian psychometric models via information criteria such as DIC and WAIC. Especially because these criteria can be automatically computed by MCMC software, it is easy to ignore the intricacies related to their computation. This often leads researchers to use noisy criteria that may lead to suboptimal analysis decisions. In this paper, we first review different forms of Bayesian information criteria that could be computed for psychometric models. We then consider best practices, highlighting computational pitfalls that can occur even when one is attempting to follow best practices. Finally, we provide recommendations for the metrics' practical uses. The paper is intended to clarify conflicting recommendations from the literature and to raise awareness about ways that information criteria can behave unexpectedly.

Keywords

Bayesian information criteria, Bayesian SEM, cross-validation, DIC, MCMC, PSIS-LOO

Bayesian statistical models are commonly compared via information criteria such as the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) and the Widely Applicable Information Criterion (WAIC; Watanabe, 2010). These metrics are especially popular because they are convenient: software will often automatically compute the metrics, and the popular decision rule of "select the model with the lowest value" is as simple as it gets.

Convenient and simple metrics do not always lead to optimal outcomes. This is especially true when we apply Bayesian information criteria to models with "random" parameters, including psychometric models with latent variables and mixed models with random effects. In these situations, it is possible to compute multiple values of the information criteria for the same model, with these multiple values potentially leading us to select different models. The difference lies in whether or not the random parameters (latent variables or random effects) are counted alongside other model parameters. While Spiegelhalter et al. (2002) mentioned this issue in their original DIC paper (referring to it as the model "focus"), the issue has often been overlooked in the time since.

Software packages that supply Bayesian information criteria vary in what is computed, leading researchers to report different metrics without realizing it. The *blavaan* package (Merkle et al., 2021; Merkle & Rosseel, 2018) computes information criteria using likelihoods that are marginal over random parameters. But for some complex models, the criteria require a large amount of post-estimation computation time. The Mplus software often computes DIC using marginal likelihoods, but it switches to conditional (on random parameters) for some complex models that would require long computation time (see Asparouhov & Muthén, 2020). The Blimp software (Keller & Enders, 2023) often computes DIC using conditional likelihoods, but it switches to marginal for some multilevel models. The *brms* package (Bürkner, 2017) computes information criteria using conditional likelihoods. The same is true of models programmed in BUGS, JAGS, Stan, or custom MCMC algorithms, because the model specification nearly always involves the random parameters. Further, the specific DIC computations differ across software because there exist multiple definitions of DIC in the literature (e.g., Plummer, 2008). It is not necessarily a problem that these multiple metrics exist, but it is a problem that researchers do not know the differences between them. It is often easy to tell whether or not the random parameters are involved in likelihoods by looking at each criterion's "effective number of parameters" value: if this is close to the frequentist parameter count, then a marginal likelihood is probably being used. If this is close to the number of observations in the dataset, then it is more likely (but not a certainty!) that a conditional likelihood is being used.

Merkle et al. (2019) considered these issues in factor analysis and simple item response models, recommending use of marginal likelihoods (marginal over latent variables) when computing information criteria. They also showed that use of conditional likelihoods (conditioning on latent variables) can lead to suboptimal theoretical properties and large Monte Carlo error. These results followed on related work by Li et al. (2016), Millar (2018), and Vehtari et al. (2016). Recent work that is closer to psychometrics includes Du et al. (2023), who considered Bayesian information criteria of multilevel models with missing data, Graves and Merkle (2022), who considered how identification constraints influence Bayesian information criteria, and Zhang et al. (2019), who considered various versions of DIC for multilevel IRT models.

The goal of this paper is to provide further clarification and discussion on Bayesian information criteria for psychometric models, and to examine related problems in more complex psychometric models. With regard to the former, there still seems to be uncertainty in the psychometric literature about information criteria. For example, many authors continue to refer to "the" DIC (or WAIC or other) value of a model without realizing that multiple values exist for the same model. And the Zhang et al. (2019) simulation results lead them to recommend a "joint" criterion for multilevel IRT models, which differs from the Merkle et al. (2019) marginal recommendation. We consider these issues further below. With regard to complex psychometric models, the nuances of information criterion computation grow with model complexity. We will discuss computational problems that are easy to overlook in psychometric models of ordinal data and of two-level data.

In the pages below, we first define a modeling framework and review the work of Merkle et al. (2019). We then consider how related problems can manifest themselves in other models, both through imprecision in numerical approximations and through heterogeneity in two-level models. Finally, we conclude with recommendations for use of information criteria in practice and with general remarks about model comparison.

Theoretical Background

We start by introducing a simple structural equation model that can be extended to the models that we consider later. Say that we observe N individuals who each report p continuous variables. We can model the data from individual i, $\boldsymbol{y}_i$, as

$$\boldsymbol{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\,\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\boldsymbol{\Lambda}$ is a $p \times m$ ($m < p$) matrix of loadings, $\boldsymbol{\eta}_i$ is an $m \times 1$ vector of individual i's latent variables, and the remaining terms are $p \times 1$ vectors. This equation is sometimes called the *measurement submodel*.

A second equation, the *structural submodel*, allows for regression relationships between latent variables. We write it as

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{\eta}_i + \boldsymbol{\zeta}_i, \tag{2}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\zeta}_i$ are each $m \times 1$, and $\boldsymbol{B}$ is $m \times m$. We additionally require that the diagonal of $\boldsymbol{B}$ is $\boldsymbol{0}$. This is because it does not make sense for a latent variable to enter into a regression relationship with itself. The residuals $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\zeta}_i$ are then typically assumed to be multivariate normal:

$$\boldsymbol{\epsilon}_i \sim N_p(0, \boldsymbol{\Theta}) \tag{3}$$

$$\zeta_i \sim N_m(0, \boldsymbol{\Psi}), \tag{4}$$

which leads to a multivariate normal model of the $\boldsymbol{y}_i$.

Of importance for information criterion computations, there are two different likelihoods that are often used for model estimation. The first likelihood conditions on the η_i and resembles a multivariate regression model:

$$\boldsymbol{y}_i | \boldsymbol{\eta}_i \sim \mathrm{N}(\boldsymbol{\nu} + \boldsymbol{\Lambda}\,\boldsymbol{\eta}_i,\ \boldsymbol{\Theta}\,). \tag{5}$$

This likelihood is commonly used for Bayesian model estimation, where we have an additional posterior distribution of latent variables from which we draw samples. For some complex models with, e.g., ordinal variables or interactions between latent variables, this likelihood greatly simplifies model estimation (e.g., Lee et al., 2007). Use of the conditional likelihood in Bayesian estimation is a type of *data augmentation* (Tanner & Wong, 1987), as it "augments" the observed data with draws of the latent variables.

The second likelihood involves integrating the latent variables out of the model (or we could say *marginalizing over the latent variables*), yielding

$$\boldsymbol{y}_i \sim \mathrm{N}(\boldsymbol{\nu} + \boldsymbol{\Lambda}\,(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\alpha},\ \boldsymbol{\Lambda}\,(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Psi}(\boldsymbol{I} - \boldsymbol{B}')^{-1}\,\boldsymbol{\Lambda}' + \boldsymbol{\Theta}\,), \tag{6}$$

where we require $(\boldsymbol{I} - \boldsymbol{B})$ to be invertible. This likelihood is used for frequentist estimation via maximum likelihood and least squares methods.

The conditional likelihood from (5) is most often used for Bayesian model estimation, especially in Gibbs sampling contexts. But either of the above likelihoods could be used for Bayesian estimation via MCMC, and they will yield the same posterior distributions of model parameters. Importantly, though, the two likelihoods will yield different information criterion values for the same model. We further consider this issue in the next section.

Information Criteria

Popular information criteria like DIC and WAIC involve evaluations of the model likelihood at posterior parameter values sampled via MCMC. Their general aim is to approximate the model's ability to predict new ("out of sample") observations, where models with better out-of-sample predictions are selected over models with worse out-of-sample predictions. These ideas are highly related to the "expected cross-validation index" that was previously developed for SEM (Browne & Cudeck, 1989; Cudeck & Browne, 1983).

All of the criteria that we consider approximate out-of-sample prediction using only in-sample data. This is accomplished by first evaluating the likelihood of the observed data at the posterior samples, which represents the model's in-sample predictive accura-

cy. We know that this predictive accuracy will be better than out-of-sample predictive accuracy. Consequently, the information criteria involve an additional penalty term for model complexity. This penalty term is sometimes called the *optimism* or *effective number of parameters*. We consider specifics below.

DIC

DIC is the simplest of the criteria considered here. Assuming a generic parameter vector $\boldsymbol{\theta}$ and S posterior samples, it can be written as

$$\text{DIC} = -2\log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}}) + 2p_D \tag{7}$$

$$PD = \left[\frac{1}{S}\sum_{s=1}^{S} -2\log p(\boldsymbol{y}|\boldsymbol{\theta}^s)\right] + 2\log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}}), \tag{8}$$

where p_D is the effective number of parameters, $\boldsymbol{\theta}^s$ is the sth posterior sample, and $\overline{\boldsymbol{\theta}}$ is the posterior mean (or another posterior measure of central tendency). Notice that the in-sample predictive accuracy $\log p(\boldsymbol{y} \mid \overline{\boldsymbol{\theta}})$ is multiplied by −2 and that p_D is generally positive. So smaller DIC values are favored over larger values.

Celeux et al. (2006) considered many variations of DIC, some of which replace the $2\log p(\boldsymbol{y} \mid \overline{\boldsymbol{\theta}})$ term with other measures of in-sample predictive accuracy. Du et al. (2023) studied some of these variations as applied to multilevel models and found evidence that they can lead to improved model selection decisions over the original DIC. While we do not describe all the variations here, we do note that they can lead to additional confusion about what DIC metric is being reported in a given application.

WAIC

WAIC is a "fully Bayesian" information criterion that utilizes pointwise predictive accuracy terms, averaging them over the entire posterior distribution. This is in contrast to DIC, which uses the overall model likelihood to approximate predictive accuracy only at the posterior mean. WAIC assumes that each datapoint is independent of the others, which immediately holds for simple linear models but becomes more complicated for mixed models and for psychometric models. Assuming N independent datapoints, we can write WAIC as

$$\text{WAIC} = -2\sum_{i=1}^{N}\log\left(\frac{1}{S}\sum_{s=1}^{S} p(\boldsymbol{y}_i|\boldsymbol{\theta}^s)\right) + 2pw \tag{9}$$

$$pw = \sum_{i=1}^{N}\text{Var}_s\left(\log p(\boldsymbol{y}_i|\boldsymbol{\theta})\right) \tag{10}$$

$$= \sum_{i=1}^{N} \frac{1}{S} \sum_{s=1}^{S} \left(\log p(\boldsymbol{y}_i | \boldsymbol{\theta}^s) - \frac{1}{S} \sum_{s=1}^{S} \log p(\boldsymbol{y}_i | \boldsymbol{\theta}^s) \right)^2 \tag{11}$$

PSIS-LOO

Watanabe (2010) showed that WAIC is asymptotically equivalent to leave-one-out cross-validation, whereby we hold out each datapoint, estimate the model, and compute the likelihood of the held-out datapoint given the estimated parameters. Vehtari et al. (2017) propose a Pareto-smoothed importance sampling (PSIS) method for directly approximating leave-one-out cross-validation, as opposed to indirectly approximating leave-one-out cross-validation via WAIC. This method, PSIS-LOO, is implemented in the R package *loo* (Vehtari et al., 2024) and is commonly obtained alongside WAIC. Just like WAIC, the PSIS-LOO metric is based on the pointwise likelihoods evaluated at each posterior sample. The algorithm underlying PSIS-LOO produces diagnostic metrics ("Pareto k" metrics) that can signal problems with the approximation as well as influential observations in the dataset. The availability of diagnostics leads us to prefer PSIS-LOO over WAIC, though the two metrics are usually very similar to one another for the psychometric models considered in this paper.

Choice of Likelihood

The above criteria involve either the overall model likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ or the pointwise likelihoods $p(\boldsymbol{y}_i \mid \boldsymbol{\theta})$. As seen by Equations (5) and (6), a psychometric model with latent variables has multiple likelihoods. The specific choice of likelihood influences the values of the information criteria and the type of predictive generalization that the information criteria approximate (see Merkle et al., 2019).

MCMC estimation of Bayesian psychometric models typically involves sampling latent variables for each person. By conditioning on these latent variables, we usually attain independence between individual observations within a person. The likelihood $p(\boldsymbol{y}_i \mid \boldsymbol{\theta})$ is then based on Equation (5), where we use conditional independence to multiply across the univariate likelihoods $p(y_{ij} \mid \boldsymbol{\theta})$. When we compute information criteria using this conditional likelihood, we approximate the model's ability to predict new data from the same people in the original dataset because we have conditioned on the latent variables η_i. Merkle et al. (2019) referred to this as "leave one unit out" cross-validation.

Frequentists typically use a marginal likelihood for model estimation, where the latent variables have been integrated out of the model. In this case, the likelihood $p(\boldsymbol{y}_i \mid \boldsymbol{\theta})$ is based on Equation (6). For IRT models and some others, the marginal criteria require quadrature or related numerical methods because an analytic expression like Equation (6) cannot be obtained. When we compute information criteria using the marginal likelihood, we approximate the model's ability to predict new data from new people who were

not in the original dataset. Merkle et al. (2019) referred to this as "leave one cluster out" cross-validation.

In typical psychometric applications, there are many reasons to prefer marginal information criteria. First, we typically view individuals as exchangeable, and we wish to generalize beyond the people that were observed in our data (see Merkle, 2026b). Marginal information criteria approximate such generalization. Second, Millar (2018) discusses the fact that conditional forms of WAIC may not be asymptotically equivalent to leave-one-out cross-validation because the number of model parameters grows with the sample size. So there is less theoretical justification for conditional information criteria, as compared to marginal information criteria. Finally, Merkle et al. (2019) showed that marginal information criteria have less Monte Carlo error than conditional information criteria because the marginal likelihoods involve many fewer parameters. This means that we can obtain precise values of marginal information criteria using fewer posterior samples, as compared to conditional information criteria.

Despite these advantages, marginal information criteria are not always used in practice. This is especially true because traditional MCMC estimation methods sample the latent variables and rely on a conditional likelihood. So researchers will immediately think of the conditional likelihood as the one to use for information criteria, and some software like JAGS will automatically compute DIC for a specified model (which usually relies on the conditional likelihood). Additionally, the marginal likelihood is often more difficult to compute than the conditional likelihood. This can be frustrating to researchers who estimate a Bayesian model with a conditional likelihood, only to find out that they need further machinery to obtain marginal information criteria.

Comparison of Marginal and Conditional Criteria

If the conditional and marginal information criteria always led to the same conclusions, then their distinction could be ignored. On this point, Merkle et al. (2019) presented theoretical results that the conditional and marginal criteria differ, along with two examples where the criteria selected different psychometric models. The conditional criteria generally favored more complex models than marginal criteria, which makes sense when we consider the types of cross-validation that they represent. For example, we already stated that conditional criteria are related to predicting new data from people that we already observed. From this standpoint, we can afford additional model complexity that tunes parameters to the people in the observed data. In contrast, a marginal criterion needs to guard against overfitting to the people in the observed data because it is related to predicting unseen people.

On Use of Joint Likelihoods

Other researchers sometimes use information criteria that are not fully marginal or conditional. Most related to the current paper, Zhang et al. (2019) recommend a "joint" form of DIC in a two-level IRT model of, e.g., students nested in schools. They do not consider WAIC or PSIS-LOO, and we have not seen joint versions of WAIC or PSIS-LOO used in practice.

Using our notation, the joint DIC involves the likelihood $p(\mathbf{y}_i, \boldsymbol{\eta}_i \mid \boldsymbol{\theta}) = p(\boldsymbol{y}_i \mid \boldsymbol{\eta}_i, \boldsymbol{\theta})p(\boldsymbol{\eta}_i \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ would contain all model parameters except the latent variables η_i. This joint likelihood is typically easier to compute than the marginal likelihood, and it appears to have been proposed because it is convenient (see Celeux et al., 2006). It resembles a one-node quadrature approximation of the marginal likelihood (e.g., Rabe-Hesketh et al., 2005), where the specific node varies for each posterior sample and where the quadrature weights do not sum to 1. Zhang et al. (2019) describe some simulations where a joint DIC selects the true model more often than a marginal DIC.

From our point of view, use of the joint likelihood for information criteria is theoretically problematic because it does not have a clear relationship to cross-validation. That is, it is unclear what it means to hold out a model parameter (the $\boldsymbol{\eta}_i$) in addition to $\boldsymbol{y}_i$ when considering a model's predictive accuracy. Inclusion of the $\boldsymbol{\eta}_i$ in the likelihood will also lead to a large amount of noise in the resulting information criteria, a point that Zhang et al. (2019) mention in their paper. These considerations lead us to maintain the default recommendation of fully marginal information criteria, because these criteria match the type of generalization in which researchers are usually interested. We further consider these recommendations in the General Discussion.

While our information criterion recommendations differ from those of Zhang et al. (2019), we definitely agree with them that marginal forms of information criteria can be computationally problematic. In addition to long computation times, we describe here two examples where likelihood approximations and model misspecifications can lead to unexpected behavior of the marginal criteria. In the first example, we consider a model of ordinal variables that requires numerical approximation of the marginal likelihood. We show how crude numerical approximations lead to imprecise information criteria. In the second example, we consider two-level SEM and show how information criteria are impacted by heterogeneity. Beyond raising awareness for these specific problems, we hope that these examples make it clear that use of information criteria requires some care.

Example 1: Impact of Numerical Approximation

Numerical approximations of marginal likelihoods are often needed in psychometric models of ordinal data. Structural equation models of ordinal data can be motivated by

continuous, latent data vectors $\boldsymbol{y}_i^*$ underlying the ordinal data $\boldsymbol{y}_i$. We take the structural equation model from Equations (1) and (2) and place it on the $\boldsymbol{y}_i^*$, with threshold parameters $\boldsymbol{\tau}$ that chop entries of $\boldsymbol{y}_i^*$ into ordered categories. For example, in the case of an ordinal variable y_{ij} with four categories, we would have:

$$
\begin{aligned}
y_{ij} &= 1 \text{ if } y_{ij}^* < \tau_{j1} \\
y_{ij} &= 2 \text{ if } \tau_{j1} < y_{ij}^* < \tau_{j2} \\
y_{ij} &= 3 \text{ if } \tau_{j2} < y_{ij}^* < \tau_{j3} \\
y_{ij} &= 4 \text{ if } y_{ij}^* > \tau_{j3},
\end{aligned}
$$

where $\tau_{j1} < \tau_{j2} < \tau_{j3}$.

Model Likelihoods

Traditional MCMC methods for estimating this model involve sampling either the $\boldsymbol{y}_i^*$ or the $\boldsymbol{\eta}_i$, both of which are a form of data augmentation. In the former case, we can condition on the sampled $\boldsymbol{y}_i^*$ and pretend like we have continuous data when sampling the other model parameters. In the latter case, we achieve independence of the entries of $\boldsymbol{y}_i$ conditioned on $\boldsymbol{\eta}_i$. Then we can evaluate the (conditional) likelihood of the $\boldsymbol{y}_i$ via univariate normal CDFs, which is computationally easy.

Neither of these estimation methods is optimal from an information criterion point of view. This is because the marginal likelihood of the $\boldsymbol{y}_i$ is marginal over both the $\boldsymbol{y}_i^*$ and the $\boldsymbol{\eta}_i$. This marginal likelihood can be computed with the CDF of the multivariate normal distribution, but its evaluation is generally too slow to be useful for MCMC estimation. We can instead estimate the model (draw posterior samples) by sampling the $\boldsymbol{y}_i^*$ or the $\boldsymbol{\eta}_i$, then compute the fully marginal likelihood of the $\boldsymbol{y}_i$ for each posterior sample after estimation.

Even if the fully marginal likelihood is computed after estimation, it is still computationally expensive. This is especially the case for metrics like WAIC and PSIS-LOO that require pointwise likelihoods. For example, say that we have a dataset of 300 individuals, and we save 1,000 posterior samples for each of three chains. If it takes one second to evaluate the pointwise log-likelihoods for one posterior sample, then it will take 50 minutes to evaluate the log-likelihoods of all 3,000 posterior samples. It is clear that we need to be as economical as possible, while still doing accurate computations. There exist multiple options for approximating the marginal likelihood, and we consider two popular options below.

Adapted Gauss-Hermite Quadrature

Because we are approximating the model likelihood after model estimation, we have access to samples of the latent variables $\boldsymbol{\eta}_i$. This allows us to make informed choices

about quadrature nodes for each case i (see Merkle et al., 2019; Rabe-Hesketh et al., 2005) so that the procedure is "already adapted" instead of "adaptive". As the number of quadrature nodes per dimension increases, Gauss-Hermite quadrature is often treated as the gold standard approximation. But large numbers of quadrature nodes are computationally expensive, even when the nodes are already adapted.

Importance Sampling of $\boldsymbol{y}_i^*$

There exist sampling-based methods for approximating the truncated multivariate normal likelihood, the most popular of which is the GHK algorithm (e.g., Hajivassiliou & McFadden, 1998). This involves transforming the multivariate normal to have mean zero and identity covariance matrix, then generating samples of $\boldsymbol{y}_i^*$ via a series of truncated univariate distributions. For the purposes of information criteria, we can additionally obtain an estimate of the marginal likelihood of $\boldsymbol{y}_i$ by averaging across importance sampling weights obtained during sampling. Larger numbers of importance samples lead to more accurate approximations, similarly to larger numbers of quadrature nodes. The *tmvnsim* package (Bhattacjarjee, 2016) provides an implementation of the GHK method in R.

For most traditional SEMs, the dimension of $\boldsymbol{\eta}_i$ is less than the dimension of the $\boldsymbol{y}_i^*$, suggesting that quadrature is advantageous over importance sampling and other methods involving truncated multivariate normal distributions (because quadrature integrates over η, and others integrate over y^*). But all the methods involve a tradeoff between computational accuracy and speed, with this tradeoff potentially having implications for information criteria. This is examined in the next section.

Illustration

Here, we study how crude approximations of the marginal log-likelihood influence the resulting information criterion values. We fit a model to an ordinal dataset, approximate the pointwise contributions to the marginal log-likelihood in various ways, and compare the resulting information criteria to a gold standard.

Method

We fit an ordinal factor analysis model to items from the Nerdy Personality Attributes Scale (Open Source Psychometrics Project, 2016), which was previously used by Schneider et al. (2020) to study frequentist model comparison statistics. The scale consists of 26 five-point Likert items that are designed to measure a person's "nerdiness." We focus here on 1018 responses to the 6-item science subscale. An example item is "I prefer academic success to social success," with the lowest option being "disagree" and the highest option being "agree."

We fit a 1-factor model to the data, which, from an IRT point of view, can be called a graded response model with probit link function. We place the model from Equation (6) on the latent responses $\boldsymbol{y}_i^*$, with the individual y_{ij} being obtained via:

$$
\begin{aligned}
y_{ij} &= 1 \text{ if } y_{ij}^* < \tau_{j1} \\
y_{ij} &= 2 \text{ if } \tau_{j1} < y_{ij}^* < \tau_{j2} \\
y_{ij} &= 3 \text{ if } \tau_{j2} < y_{ij}^* < \tau_{j3} \\
y_{ij} &= 4 \text{ if } \tau_{j3} < y_{ij}^* < \tau_{j4} \\
y_{ij} &= 5 \text{ if } y_{ij}^* > \tau_{j4},
\end{aligned}
$$

where $\tau_{j1} < \tau_{j2} < \tau_{j3} < \tau_{j4}$. Prior distributions were intended to be mildly informative and were set as:

$$
\begin{aligned}
\lambda_j &\sim \text{Normal}(1, .4)\, j = 2, \ldots, 6 \\
\psi &\sim \text{Gamma}(1, .5) \\
\tau_{j1} &\sim \text{Normal}(0, 1.5)\, j = 1, \ldots, 6
\end{aligned}
$$

$$
\log(\tau_{jk} - \tau_{j(k-1)}) \sim \text{Normal}(0, 1.5)\, j = 1, \ldots, 6; k = 2, \ldots, 4.
$$

To identify the model, we fixed $\lambda_1 = 1$ and $\theta_j = 1$ for $j = 1,\ldots, 6$. For each of three chains, we used *blavaan* to obtain 2,000 samples following 500 burnin samples. This is more samples than we need for precise estimates of posterior means, but the information criteria generally have more Monte Carlo error than posterior means. In past work, if it takes S posterior samples for stable posterior means, then we take $2S$ posterior samples for the information criteria. A more formal examination can also be useful, say by monitoring the model log-likelihood during MCMC and examining its effective sample size.

Following model estimation, we computed DIC and PSIS-LOO using a variety of marginal likelihood approximations. Our gold standard marginal log-likelihood approximation is an adapted quadrature method with a large number of nodes. For the particular model that we consider here, 8 quadrature nodes per dimension was sufficient (further numbers of nodes left the log-likelihoods virtually unchanged). We compared the gold standard to GHK sampling with 5, 10, 20, and 50 samples per case, and to adapted quadrature with 1, 2, and 4 nodes.

Results

It is difficult to make exact statements about the computation times of the marginal likelihood approximations because the code that we used (see Merkle, 2026a) was not optimized for parallelization or for all available computational shortcuts. But to provide a rough idea of timings, each set of marginal likelihood computations took at least 5

minutes per chain on a mid-range laptop, with some taking closer to an hour per chain. GHK sampling was faster than quadrature, likely because the GHK code had better optimizations. Cruder approximations (fewer quadrature points or importance samples) provided speedups of up to 30%.

Figure 1 shows our primary results, with information criteria on the y-axes and number of quadrature points or importance samples on the x-axes. The columns show the adapted quadrature and GHK methods, and the rows show DIC and PSIS-LOO. The dashed line is the gold standard value with 8 quadrature points, and shaded areas represent 1 standard error around the criterion value (note that standard errors are not available for DIC). We see that the quadrature method generally yields information criterion values that are close to the gold standard. For GHK sampling, the criterion values are slightly larger when the approximation is crude. These differences look small, with the dashed line being well within one standard error of the PSIS-LOO values. But information criteria are generally correlated across models, so that differences between model criteria will have much smaller standard errors. For example, it is possible to compare the gold standard PSIS-LOO to other PSIS-LOO metrics, pretending like the metrics were coming from different models. In doing so, we find PSIS-LOO differences of 10 standard errors or more.

Figure 2 is arranged similarly to Figure 1, except it shows the estimated effective number of parameters instead of the information criteria. We see here that, under crude approximations of the marginal likelihood, the effective number of parameters is usually estimated to be too large. The exception is the upper right panel, which shows DIC with GHK sampling. There, effective number of parameters is too low until we obtain 40 samples or more for the approximation.

Discussion

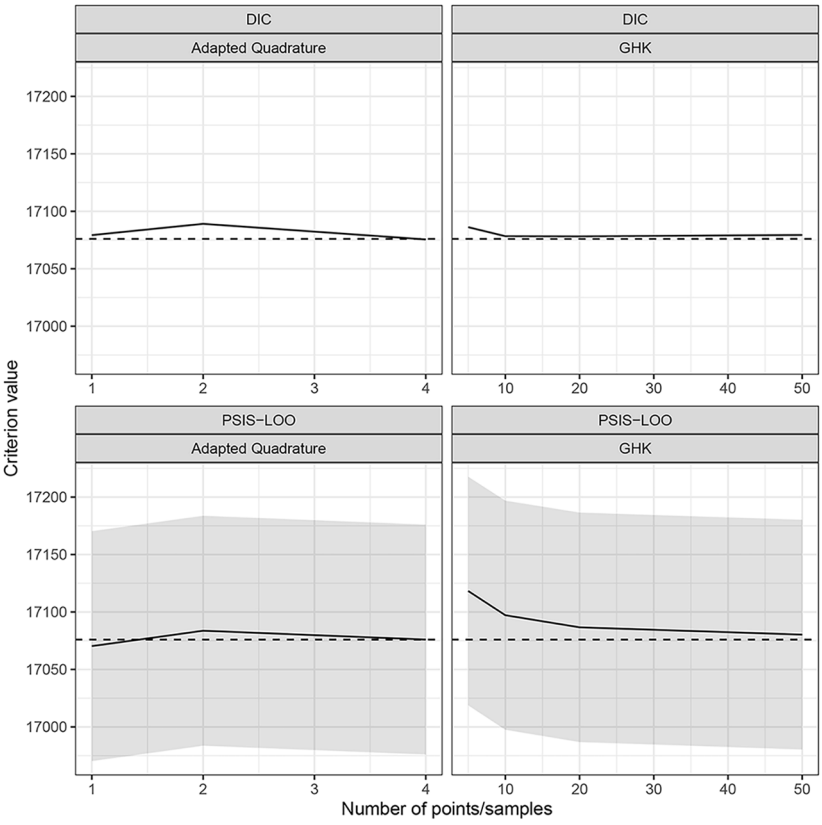
For the model considered here, crude approximations of the marginal log-likelihood led to information criteria that were sometimes too large. The GHK method with 5 samples is the crudest approximation here, and we found that its influence on effective number of parameter computations differed across DIC and PSIS-LOO (leading to a low value for DIC and a high value for PSIS-LOO). We believe that this is because, for PSIS-LOO, the effective number of parameter computations involve variances of pointwise log-likelihoods, so that imprecision from the crude integral approximations leads to upward bias. This result is related to the work of Timonen et al. (2023), who proposed importance weights that correct for crude numerical approximations in Ordinary Differential Equation models. DIC, on the other hand, involves averages of the full model log-likelihood that are differently affected by crude approximations.

In some cases, the marginal likelihood approximations will not influence model selection decisions. This is because the extra approximation imprecision will be small compared to the differences in the models' out-of-sample predictive accuracy. More cau-

tion may be warranted in situations where we compare models with different numbers of latent variables. For example, if one model involves a 1-dimensional latent variable and another model involves a three-dimensional latent variable, we may expect the second model to suffer more from the integral approximation. And in situations where models have the same number of latent variables, we may still be cautious about selecting one model when the information criterion values are close (also see Merkle et al., 2016).
In the next section, we consider how model misspecifications can lead to unexpected behavior of information criteria.

Figure 1

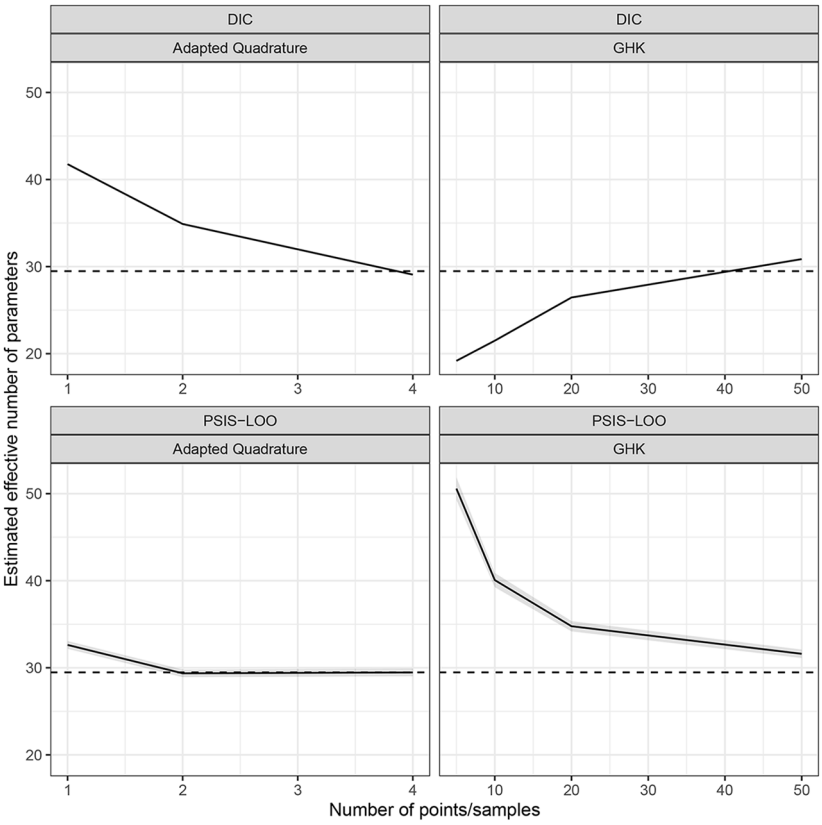
Information Criteria of Ordinal Factor Analysis Model



Note. Criteria are computed using marginal likelihoods approximated by adapted quadrature and by GHK sampling with differing numbers of nodes/samples. Shaded regions represent ±1 standard error.

Figure 2

Estimated Effective Numbers of Parameters Associated with DIC and PSIS-LOO



Note. Effective numbers of parameters are computed using marginal likelihoods approximated by adapted quadrature and by GHK sampling with differing numbers of nodes/samples. Shaded regions represent ±1 standard error.

Example 2: Heterogeneity in Two-Level Models

Two-level structural equation models lead to additional conditional/marginal distinctions in information criteria because they involve latent variables at multiple levels. These models are traditionally applied to datasets of students nested within schools or within countries. We observe multiple test scores and/or other variables per student, which leads us to specify student-level latent variables as in traditional SEM. But students within the same school are correlated with one another, and we want our model to

additionally account for these correlations. We review this modeling framework first before considering problems with information criteria.

Model

We now define the vector $\boldsymbol{y}_{ik}$ to be the data of Level 1 unit i within Level 2 unit k. In traditional applications, i would be student and k would be school, and we generally refer to students and schools below because it helps with intuition. Mixed modeling researchers may view this as a three-level dataset where individual observations are nested within students, and students are nested within schools. See Skrondal and Rabe-Hesketh (2004) for further discussion of these different viewpoints.

Specification of a two-level SEM with random intercepts can look very similar to a one-level SEM. We start out writing

$$\boldsymbol{y}_{ik} = \boldsymbol{\nu}_k + \boldsymbol{\Lambda}\,\boldsymbol{\eta}_{ik} + \boldsymbol{\varepsilon}_{ik} \tag{12}$$

$$\boldsymbol{\eta}_{ik} = \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{\eta}_{ik} + \boldsymbol{\zeta}_{ik}, \tag{13}$$

which is nearly the same as the traditional model from Equations (1) and (2). The only differences are that, (i) our intercept $\boldsymbol{\nu}$ has a k subscript, and (ii) some other vectors have ik subscripts because we are designating the model for student i in school k.

The $\boldsymbol{\nu}_{\mathrm{k}}$ term implies that the intercept is unique to each school k, which is what leads us to call this a "two-level" SEM. We complete the model by considering how to handle the $\boldsymbol{\nu}_{\mathrm{k}}$. If we come from a mixed modeling background, the natural thing to do is place a multivariate normal hyperdistribution on the $\boldsymbol{\nu}_{\mathrm{k}}$, i.e.,

$$\boldsymbol{\nu}_k \sim \mathrm{N}(\boldsymbol{\nu}_0, \boldsymbol{\Sigma}_\nu)\,. \tag{14}$$

This can add many parameters to the model, especially when each individual has many observed variables. For example, if each individual has $p = 9$ observed variables, then $\boldsymbol{\Sigma}_{\mathrm{v}}$ contains 45 free parameters.

Because an unrestricted multivariate normal can lead to many extra parameters, and also because we are already using latent variables, we could elect to place a second SEM on the $\boldsymbol{\nu}_{\mathrm{k}}$. This typically involves specifying school latent variables that are predictive of individual entries of $\boldsymbol{\nu}_{\mathrm{k}}$. We basically repeat our SEM, modeling $\boldsymbol{\nu}_{\mathrm{k}}$ instead of $\boldsymbol{y}_{ik}$:

$$\boldsymbol{\nu}_k = \boldsymbol{\nu}_0 + \boldsymbol{\Lambda}^{(c)}\,\boldsymbol{\eta}_k^{(c)} + \boldsymbol{\varepsilon}_k^{(c)} \tag{15}$$

$$\boldsymbol{\eta}_k^{(c)} = \boldsymbol{\alpha}^{(c)} + \boldsymbol{B}^{(c)}\boldsymbol{\eta}_k^{(c)} + \boldsymbol{\zeta}_k^{(c)}, \tag{16}$$

where the (c) superscripts distinguish the school-level matrices from the student-level matrices. The mean intercept across schools, $\boldsymbol{\nu}_0$, does not have a superscript because it

is similar to the intercept in a traditional, one-level SEM. And while we do not consider it here, the model may include school-level observed variables that do not exist for individual students. In that case, the $\boldsymbol{\nu}_k$ vector is augmented with the observed variables from school k (see Rosseel, 2021).

Conditional and Marginal Distributions

The model has student-level latent variables as well as school-level latent variables. Traditional model estimation via MCMC would proceed by sampling all the latent variables, which typically provides conditional independence between observed variables. This leads to a univariate conditional model likelihood that is relatively simple to evaluate, but it also has implications for information criteria. Depending on what latent variables are (and are not) marginalized out of the model, our information criteria approximate out-of-sample prediction at different levels (also see Zhang et al., 2019):

- No marginalization (fully conditional likelihood): Predict new observations from the same students that were in the original data (who are also in the same schools).
- Marginalize over student latent variables (partially marginal likelihood): Predict new observations from new students in the same schools that were in the original data.
- Marginalize over both student and school latent variables (fully marginal likelihood): Predict new observations from new students in new schools.

Note that our terminology refers only to marginalization of latent variables and should not be confused with the marginal likelihood used for Bayes factor computations. The latter marginalizes over all model parameters, including the parameters that would be viewed as fixed in a frequentist context.

In most psychometric modeling situations, researchers do not care about the specific students and/or schools in the data. So, for computing information criteria, the fully marginal likelihood would be the recommended default. Exceptions may be studies where researchers are assessing the specific students in the dataset, or school accountability studies where researchers care about the specific schools in the dataset. The former may warrant a fully conditional likelihood, and the latter may warrant a partially marginal likelihood.

Unfortunately, the two-level, fully marginal likelihoods involved in the default recommendation can be computationally difficult. The fully marginal likelihood accounts for the fact that students within a school are correlated, which typically leads to a high-dimensional multivariate normal distribution. As a specific example, imagine that we observe 30 students in each school, and each student provides 5 observed variables. Then our fully marginal likelihood above will involve a 150-dimensional normal distribution. It is difficult to evaluate this high-dimensional likelihood, which leads to excessively slow computations.

Fortunately, a good deal of attention has been devoted to this problem, and a large amount of progress has been made (e.g., du Toit & du Toit, 2008; McDonald, 1993; Rosseel, 2021). This is because the fully marginal likelihood is used for frequentist model estimation, and improved computations can lead to faster frequentist estimation. In the previous work, researchers took advantage of the structure of the model covariance matrix to evaluate the likelihood using matrices of low dimension. Let n_k be the number of students in school k, each of whom provides p observed variables. The previous authors have derived expressions that allow us to evaluate the likelihood using matrices of dimension p, instead of using matrices of dimension $n_k p$. The appendix includes technical detail on computing the fully marginal, pointwise log-likelihood of each Level-2 unit in a two-level SEM with random intercepts.

Illustration

In the two-level SEM framework, it is possible to observe multiple types of heterogeneity. Here, we consider a situation where the within-cluster covariance matrix differs across clusters. We generate data from a two-level model while manipulating the magnitude of heterogeneity. We then fit a model that ignores this heterogeneity and examine the resulting information criteria.

Method

We generated data from a model with 4 observed variables and 200 clusters of size 50. The data-generating model can be written as:

$$\boldsymbol{y}_{ik} \mid \boldsymbol{\nu}_k \sim \begin{cases} \mathrm{N}(\boldsymbol{\nu}_k, \boldsymbol{\Sigma}_w) & i = 1, ..., 50; k = 1, ..., 100 \\ \mathrm{N}(\nu_k, d \times \boldsymbol{\Sigma}_w) & i = 1, ..., 50; k = 101, ..., 200 \end{cases} \tag{17}$$

$$\boldsymbol{\nu}_k \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_b)\ k = 1, ..., 200 \tag{18}$$

$$\boldsymbol{\Sigma}_w = \begin{pmatrix} 1.5 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1.5 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1.5 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1.5 \end{pmatrix} \tag{19}$$

$$\boldsymbol{\Sigma}_b = \begin{pmatrix} 0.6 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.6 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.6 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.6 \end{pmatrix} \tag{20}$$

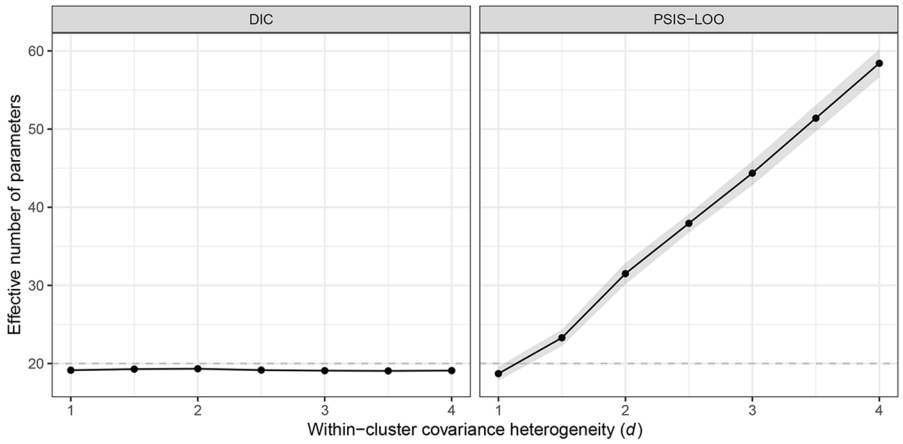
where we manipulated the scalar d to introduce heterogeneity in the within-cluster covariance matrix. The d parameter assumed values from 1 to 4 in steps of 0.5, where a

new dataset was generated for each value of *d*. We then fit a traditional two-level model that assumes a single within-cluster covariance matrix for all clusters. In *(b)lavaan*, this model could be specified as

```
model <-'
    level: within
        fw =~ y1 + y2 + y3 + y4
    level: between
        fb =~ y1 + y2 + y3 + y4
'
```

Figure 3

Magnitude of Heterogeneity in Within-Cluster Covariance Matrix (X-Axis) vs. Estimated Effective Number of Parameters (Y-Axis) of DIC and PSIS-LOO.



The prior distributions that we used for model estimation were

$$
\begin{aligned}
\boldsymbol{\nu}_0 &\sim \text{Normal}(0, 32) \\
\lambda_j &\sim \text{Normal}(1, .4) \quad j = 2, \ldots, 4 \\
\lambda_j^{(c)} &\sim \text{Normal}(1, .4) \quad j = 2, \ldots, 4 \\
\psi &\sim \text{Gamma}(1, .5) \\
\psi^{(c)} &\sim \text{Gamma}(1, .5) \\
\theta_j &\sim \text{Gamma}(1, .5) \quad j = 1, \ldots, 4 \\
\theta_j^{(c)} &\sim \text{Gamma}(1, .5) \quad j = 1, \ldots, 4.
\end{aligned}
$$

where $\lambda_1 = \lambda_1^{(c)} = 1$ and the ψ and θ parameters are standard deviations, as opposed to variances. The priors on loadings are intended to be mildly informative about the direction of correlation between observed variables (i.e., we expect all observed variables to be positively related). The remaining priors are relatively uninformative, considering the data generating model. For each of three chains, we used *blavaan* to obtain 2,000 samples following 500 burnin samples.

Results

Results are shown in Figure 3. The value of d is on the x-axis, where 1 represents no heterogeneity and values above 1 represent increasing magnitudes of heterogeneity. The estimated effective numbers of parameters are on the y-axis, panels represent different information criteria, and shaded areas represent ±1 standard error (which is unavailable for DIC). The dashed, horizontal line in each panel is the frequentist parameter count, which is 20 for this model. We see that the DIC effective number of parameters is nearly flat and stays near the frequentist parameter count of 20. This is common behavior under mildly informative priors. But under PSIS-LOO, the model appears to gain parameters as the heterogeneity increases. While it is difficult to see in the figure, the standard error for the PSIS-LOO effective number of parameters nearly doubles from $d = 1$ (no heterogeneity) to $d = 4$. The standard error of the full PSIS-LOO criterion (as opposed to the standard error for effective number of parameters) increases more dramatically, from a value of 282 at $d = 1$ to a value of 1,717 at $d = 4$. And even at $d = 4$, the Pareto k diagnostics are all in the range that the *loo* package designates as "good."

When using the marginal likelihood, PSIS-LOO approximates predictive accuracy of new students in new schools. But when there is large heterogeneity from school to school, it becomes difficult to use the existing data to make predictions for new schools. The inflated effective number of parameters reflects the idea that, under the heterogeneity considered here, we cannot effectively pool information across schools. Our result is similar to the "scaled 8 schools" example of Vehtari et al. (2017), who considered a related problem in a simple two-level model. They showed that WAIC and PSIS-LOO break down when there is no pooling across schools, that is, when data from one school are not informative about data from other schools. We have essentially illustrated a similar issue in a larger model. A related issue can be observed when estimating two-level models with "within only" variables that appear only at Level 1. See the example in Appendix B.

It may be tempting to view DIC as advantageous in this example because its effective number of parameters remains similar, regardless of the amount of heterogeneity in the data. But we find the behavior of PSIS-LOO worthwhile because it can be used as a diagnostic that signals model misspecification. The unexpected increase in effective number of parameters can potentially lead researchers to do further posterior checking and to improve the model.

General Discussion

In this paper, we first reviewed the idea that multiple versions of Bayesian information criteria are available for the same model, depending on how one handles the latent variables and random effects. We recommend use of marginal likelihoods (marginal over latent variables and random effects) as the default because the resulting information criteria approximate predictive generalization to new people and to new Level-2 units, in the two-level case. This type of generalization is usually of primary interest, as compared to generalization to new data from the same people and/or from the same Level-2 units that were in the original dataset. The marginal information criteria also have better theoretical justification and less Monte Carlo error, as compared to conditional information criteria.

We then highlighted the idea that computation of Bayesian information criteria can be more intricate than one would expect. Multiple criteria exist for the same model, numerical approximations that are not involved in model estimation may still influence the information criteria, and heterogeneity in two-level models may lead to unexpected behavior in the information criteria. These results are especially problematic for researchers who view information criteria as simple methods for model comparison that do not require much extra thought. In the sections below, we further consider applied use of information criteria and related methods in psychometrics.

Computational Issues

Marginal information criteria are generally more difficult to compute, as compared to conditional information criteria, because they often require extra computations after model estimation that may not be needed otherwise. Software tools can facilitate some of these extra computations. These include the *blavaan* package, which automatically computes marginal information criteria for the models that it is able to estimate, and the *bleval* package (Luo & Dong, 2025; Luo et al., 2026), which contains a quadrature method that can be flexibly applied to the posterior samples of many Bayesian models. The latter package requires users to specify a function defining the model's conditional likelihood.

While the above packages make marginal likelihood evaluations easier, the computations will still be intensive and/or infeasible for some models. While this may lead researchers to use conditional or joint likelihoods, we caution that those criteria do not necessarily approximate the generalization that would often be expected of psychometric models. Researchers may instead consider alternative model evaluation methods and computations. For example, the *loo* package authors and their colleagues recently developed a subsampling approach to computing PSIS-LOO for large datasets (Magnusson et al., 2020; Magnusson et al., 2025). This approach does not require us to compute the likelihood for every observation in the dataset so can lead to faster computations, while also providing estimates of additional uncertainty due to subsampling. Alternatively,

it is possible to directly carry out a cross-validation analysis by sequentially holding out data, fitting models, and examining predictions of the held-out data via metrics other than log-likelihoods. This may be advantageous when model estimation is fast and marginal likelihood computation is slow. As a third alternative, there might be an encompassing model that contains all candidate models as special cases. We can use the encompassing model to account for our uncertainty in candidate models, and the encompassing model also facilitates other model comparison metrics such as the Bayes factor (e.g., Wagenmakers et al., 2010). Finally, Bayesian model averaging, stacking, and related methods can also help us address model uncertainty without the need to select a single model (e.g., Kaplan, 2021; Yao et al., 2018). We acknowledge that none of these solutions provides a fast and simple replacement for conditional information criteria that are automatically computed by some software.

Metric Selection

Throughout the paper, we highlighted that a specific information criterion (marginal or otherwise) should be chosen based on the type of cross-validation that each criterion approximates. Other researchers sometimes focus on a criterion's ability to select the data generating model. In our view, the model's ability to select the true model is not necessarily relevant in situations where the true model does not exist. This topic was discussed at length in a special issue of *Computational Brain & Behavior* by authors with diverse points of view (e.g., Gronau & Wagenmakers, 2019; Vehtari et al., 2019). Piironen and Vehtari (2017) relatedly considered the use of different metrics depending on whether or not we expect the true model to be in the set of candidate models (see their Table 1). We also recall the earlier remarks of Browne (2000): "Any attempt to use cross-validation to detect which of a set of models is correct disregards the fundamental intent of the approach. This is to find a model that yields as small an overall discrepancy as is possible given a specific sample size, not to find a model with no error of approximation. In other words, the intent is to find a model that yields a calibration that is worth further examination despite a possibly small sample, not to seek a correct model with an unknown calibration." (p. 130).

More generally, as pointed out by Navarro (2019), all the information criteria and other metrics mentioned in this paper emphasize quantitative measures of model performance at the neglect of whether any of the models help solve scientific problems. The overall goal of solving scientific problems should not be lost among model selection metrics.

Recommendations

We offer a variety of practical recommendations for researchers who estimate Bayesian models via MCMC and report Bayesian information criteria. First, verify that there

are random model parameters (that is, parameters that would be called "random" in a frequentist context), which will often be called random effects, latent variables, or factors. If there are no random parameters, then the issues described in this paper are not applicable. Second, ensure that a large number of posterior samples are taken to reduce noise in the information criteria. We recommend starting with twice the number of samples that one would use for posterior means of individual parameters, as well as monitoring the effective sample size of the log-likelihood values that are used to compute information criteria. One may encounter a speed-accuracy tradeoff in this step, and conditional information criteria will have more noise than marginal.

Following estimation, researchers should verify the type of information criteria that their software supplies. To our knowledge, Mplus and *blavaan* are the only programs that supply marginal information criteria of psychometric models by default. Most other software will produce conditional information criteria because the computations are easier. The effective number of parameters associated with a criterion often provides a hint about what likelihood is being used: if the effective number of parameters is close to the frequentist count of model parameters, then a marginal likelihood is being used. Fourth, if one has conditional information criteria but does not have a good reason to prefer conditional information criteria, consider whether it is possible to compute marginal information criteria (with subsampling, if necessary), to compute a related cross-validation metric, or to estimate an encompassing model; see the Computational Issues section above. Finally, be transparent in reporting: provide details about the specific types of metrics that were reported, and be careful about using the metrics to select one model in light of the nuances described in this paper.

Conclusions

The popularity of Bayesian information criteria partly stems from the fact that they can be byproducts of MCMC estimation, allowing researchers to compare models without needing to consider post-estimation checks or computations. This simplicity becomes problematic as models increase in complexity, because many researchers do not expect the information criteria to require extra consideration or computations. In this paper, we highlighted some of these issues and made some recommendations for computing information criteria and related cross-validation metrics. The issues might be disappointing to some researchers because they imply that the information criteria require more work and attention than is desired. This could lead researchers to consider whether they are using information criteria because they are simple or because they are the best tools for their work. Because computation and effort are not infinite, researchers might sometimes find that their resources are better spent on posterior checks, assessment of qualitative differences between models, and encompassing models.

Computational Details

All results were obtained using the R system for statistical computing (R Core Team, 2025), Version 4.5.2, especially relying on packages *blavaan* (Merkle et al., 2021), *loo* (Vehtari et al., 2024), and *rstan* (Stan Development Team, 2025). The code for reproducing the results accompanies this submission (see Merkle, 2026a).

Funding: This work was made possible through funding from the Institute of Education Sciences, U.S. Department of Education, Grant R305D210044.

Acknowledgments: The author has no additional (i.e., non-financial) support to report.

Competing Interests: The author has declared that no competing interests exist. The author wrote the paper himself and excluded AI.

Data Availability: The code (see Merkle, 2026a) and data (see Merkle, 2026b) for the paper are openly available in the Supplementary Materials.

Supplementary Materials

Type of supplementary material	Availability/Access
Data	
Merkle_2026_Nuances_SUPPL_data	Merkle (2026b)
Code	
Merkle_2026_Nuances_SUPPL_code.R	Merkle (2026a)
Material	
No supplementary material provided	—
Study/Analysis preregistration	
No preregistration	—
Other	
No other material provided	—

References

Asparouhov, T., & Muthén, B. (2020). Comparison of models for the analysis of intensive longitudinal data. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(2), 275–297. https://doi.org/10.1080/10705511.2019.1626733

Bhattacjarjee, S. (2016). *tmvnsim: Truncated multivariate normal simulation* [R Package Version 1.0-2]. R Project for Statistical Computing. https://CRAN.R-project.org/package=tmvnsim

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44*(1), 108–132. https://doi.org/10.1006/jmps.1999.1279

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24*(4), 445–455. https://doi.org/10.1207/s15327906mbr2404_4

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Celeux, G., Forbes, F., Robert, C. P., & Titterington, D. M., (2006). Deviance information criteria for missing data models. *Bayesian Analysis, 1*(4), 651–673. https://doi.org/10.1214/06-BA122

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18*(2), 147–167. https://doi.org/10.1207/s15327906mbr1802_2

Du, H., Keller, B., Alacam, E., & Enders, C. (2023). Comparing DIC and WAIC for multilevel models with missing data. *Behavior Research Methods, 56*(4), 2731–2750. https://doi.org/10.3758/s13428-023-02231-0

du Toit, S. H. C., & du Toit, M. (2008). Multilevel structural equation modeling. In J. De Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 435–478). https://doi.org/10.1007/978-0-387-73186-5_12

Graves, B., & Merkle, E. C. (2022). A note on identification constraints and information criteria in Bayesian latent variable models. *Behavior Research Methods, 54*, 795–804. https://doi.org/10.3758/s13428-021-01649-8

Gronau, Q. F., & Wagenmakers, E.-J. (2019). Limitations of Bayesian leave-one-out cross-validation for model selection. *Computational Brain & Behavior, 2*(1), 1–11. https://doi.org/10.1007/s42113-018-0011-7

Hajivassiliou, V. A., & McFadden, D. L. (1998). The method of simulated scores for the estimation of LDV models. *Econometrica, 66*(4), 863–896. https://doi.org/10.2307/2999576

Kaplan, D. (2021). On the quantification of model uncertainty: A Bayesian perspective. *Psychometrika, 86*(1), 215–238. https://doi.org/10.1007/s11336-021-09754-5

Keller, B. T., & Enders, C. K. (2023). *Blimp user's guide* (3rd ed.). https://www.appliedmissingdata.com/blimp.

Lee, S.-Y., Song, X.-Y., & Tang, N.-S. (2007). Bayesian methods for analyzing structural equation models with covariates, interaction, and quadratic latent variables. *Structural Equation Modeling, 14*(3), 404–434. https://doi.org/10.1080/10705510701301511

Li, L., Qui, S., & Feng, C. X. (2016). Approximating cross-validatory predictive evaluation in Bayesian latent variable models with integrated IS and WAIC. *Statistics and Computing, 26*(4), 881–897. https://doi.org/10.1007/s11222-015-9577-2

Luo, X., & Dong, J. (2025). *bleval: Bayesian evaluation for latent variable models* [R package Version 0.0.0.9000]. GitHub. https://github.com/luoxh3/bleval

Luo, X., Dong, J., Liu, H., Liu, Y., & Merkle, E. C. (2026). *Bayesian evaluation for latent variable models: A tutorial on computing information criteria and Bayes factors with the R package* bleval [Manuscript in preparation].

Magnusson, M., Andersen, M. R., Jonasson, J., & Vehtari, A. (2020). *Leave-one-out cross-validation for Bayesian model comparison in large data* [arXiv 2001.00980]. arXiv. https://arxiv.org/abs/2001.00980

Magnusson, M., Bürkner, P., Vehtari, A., & Gabry, J. (2025, December 23). Using leave-one-out cross-validation for large data. https://mc-stan.org/loo/articles/loo2-large-data.html

McDonald, R. P. (1993). A general model for two-level data with responses missing at random. *Psychometrika, 58*(4), 575–585. https://doi.org/10.1007/bf02294828

Merkle, E. C. (2026a). *Supplementary Materials to* "Nuances of information criteria for Bayesian psychometric models" [R codes]. PsychOpen GOLD. http://dx.doi.org/10.23668/psycharchives.22424

Merkle, E. C. (2026b). *Supplementary Materials to* "Nuances of information criteria for Bayesian psychometric models" [Full study data]. PsychOpen GOLD. http://dx.doi.org/10.23668/psycharchives.22423

Merkle, E. C., Fitzsimmons, E., Uanhoro, J., & Goodrich, B. (2021). Efficient Bayesian structural equation modeling in Stan. *Journal of Statistical Software, 100*(6), 1–22. https://doi.org/10.18637/jss.v100.i06

Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika, 84*(3), 802–829. https://doi.org/10.1007/s11336-019-09679-0

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

Merkle, E. C., You, D., & Preacher, K. J. (2016). Testing non-nested structural equation models. *Psychological Methods, 21*(2), 151–163. https://doi.org/10.1037/met0000038

Millar, R. B. (2018). Conditional vs. marginal estimation of predictive loss of hierarchical models using WAIC and cross-validation. *Statistics and Computing, 28*(2), 375–385. https://doi.org/10.1007/s11222-017-9736-8

Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior, 2*, 28–34. https://doi.org/10.1007/s42113-018-0019-z

Open Source Psychometrics Project. (2016). *Data from* "The Nerdy Personality Attributes Scale" [Dataset]. https://openpsychometrics.org/%5C_rawdata/

Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing, 27*, 711–735. https://doi.org/10.1007/s11222-016-9649-y

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics, 9*(3), 523–539. https://doi.org/10.1093/biostatistics/kxm049

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics, 128*(2), 301–323. https://doi.org/10.1016/j.jeconom.2004.08.017

Rosseel, Y. (2021). Evaluating the observed log-likelihood function in two-level structural equation modeling with missing data: From formulas to R code. *Psych, 3*(2), 197–232. https://doi.org/10.3390/psych3020017

Schneider, L., Chalmers, R. P., Debelak, R., & Merkle, E. C. (2020). Model selection of nested and non-nested item response models using Vuong tests. *Multivariate Behavioral Research, 55*(5), 664–684. https://doi.org/10.1080/00273171.2019.1664280

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation modeling*. Chapman & Hall.

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

Stan Development Team. (2025). *RStan: The R interface to Stan* [R package Version 2.32.7]. Stan Project. https://mc-stan.org/

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*(398), 528–540. https://doi.org/10.1080/01621459.1987.10478458

Timonen, J., Siccha, N., Bales, B., Lähdesmäki, H., & Vehtari, A. (2023). An importance sampling approach for reliable and efficient inference in Bayesian ordinary differential equation models. *Stat, 12*(1), Article e614. https://doi.org/10.1002/sta4.614

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2024). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models* [R package Version 2.8.0]. Stan Project. https://mc-stan.org/loo/

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. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., & Winther, O. (2016). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *Journal of Machine Learning Research, 17*(103), 1–38.

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. https://doi.org/10.1007/s42113-018-0020-6

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology, 60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

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. https://doi.org/10.48550/arXiv.1004.2316

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis, 13*(3), 917–1007. https://doi.org/10.1214/17-BA1091

Zhang, X., Tao, J., Wang, C., & Shi, N.-Z. (2019). Bayesian model selection methods for multilevel IRT models: A comparison of five DIC-based indices. *Journal of Educational Measurement, 56*(1), 3–27. https://doi.org/10.1111/jedm.12197

Appendix

Appendix A: Clusterwise Marginal Likelihood Contributions in Two-Level SEM With Random Intercepts

In this appendix, we provide technical details about how to obtain cluster-level contributions to the marginal likelihood of a two-level SEM with random intercepts. Many relevant derivations are presented by Rosseel (2021), who builds on work by du Toit and du Toit (2008) and McDonald (1993). For simplicity, we exclude observed variables that only occur at Level 2 (e.g., school-level observed variables). For intuitive discussion, we generically refer to Level 1 units as "students" and Level 2 clusters as "schools."

The model is defined by Equations (12) and (15). The model from those equations includes the latent variables at both levels, and we wish to obtain a likelihood that is marginalized over those latent variables. But in marginalizing over the latent variables, we must account for correlations between students who are in the same school. Because our model involves multivariate normal distributions, we can obtain analytic expressions for the marginal likelihood.

An initial step towards these analytic expressions involves obtaining the mean and covariance of $\boldsymbol{y}_{ik}$, i.e., the mean and covariance of the observed variables from a particular student. These are similar to expressions from one-level SEM and are given by

$$\boldsymbol{\mu}_y = \boldsymbol{\nu}_0 + \boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\alpha} + \boldsymbol{\Lambda}^{(c)}(\boldsymbol{I} - \boldsymbol{B}^{(c)})^{-1}\boldsymbol{\alpha}^{(c)} \tag{21}$$

$$\boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_w + \boldsymbol{\Sigma}_b, \tag{22}$$

where the covariance matrix $\boldsymbol{\Sigma}_y$ is familiarly decomposed into within and between components. Some treatments of multilevel SEM with random intercepts define the model as this decomposition. This may be useful when the mean vector is not of interest, but it is also not the most intuitive model definition. Regardless, the between and within components mirror one another, involving parameter matrices at the student and school levels, respectively:

$$\boldsymbol{\Sigma}_w = \boldsymbol{\Lambda}(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Psi}(\boldsymbol{I} - \boldsymbol{B})^{-1'}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \tag{23}$$

$$\boldsymbol{\Sigma}_b = \boldsymbol{\Lambda}^{(c)}(\boldsymbol{I} - \boldsymbol{B}^{(c)})^{-1}\boldsymbol{\Psi}^{(c)}(\boldsymbol{I} - \boldsymbol{B}^{(c)})^{-1'}\boldsymbol{\Lambda}^{(c)'} + \boldsymbol{\Theta}^{(c)} \tag{24}$$

Next, let n_k be the number of individuals in school k. Then our marginal likelihood for all observed variables in school k is multivariate normal with mean

$$\boldsymbol{\mu}_k = \left(\mathbf{1}'_{n_k} \otimes \boldsymbol{\mu}'_y\right)' \tag{25}$$

and covariance matrix

$$\boldsymbol{V}_k = \left(\mathbf{1}_{n_k}\mathbf{1}_{n_k}' \otimes \boldsymbol{\Sigma}_b + \boldsymbol{I}_{n_k} \otimes \boldsymbol{\Sigma}_w\right). \tag{26}$$

where $\mathbf{1}_{n_k}$ is a column vector of n_k 1s. This leads us to express school k's marginal log-likelihood as

$$\ell_k(\boldsymbol{\theta} \mid \boldsymbol{y}_k) = -\frac{1}{2}\left(P_k \log(2\pi) + \log(|\boldsymbol{V}_k|) + (\boldsymbol{y}_k - \boldsymbol{\mu}_k)'\boldsymbol{V}_k^{-1}(\boldsymbol{y}_k - \boldsymbol{\mu}_k)\right). \tag{27}$$

where $P_k = n_k p$.

While the above expression is all we need to evaluate the marginal log-likelihood contribution of each school, it will usually be very slow to evaluate. It involves both the inverse and determinant of $\boldsymbol{V}_k$, with the dimension of this matrix equaling the total number of observed variables in school k (P_k). But the structure of $\boldsymbol{V}_k$ allows for a variety of simplifications (see Rosseel, 2021). In particular, the log-determinant can be written as

$$\log(|\boldsymbol{V}_k|) = (n_k - 1)\log(|\boldsymbol{\Sigma}_w|) + \log(|n_k\boldsymbol{\Sigma}_b + \boldsymbol{\Sigma}_w|), \tag{28}$$

which involves determinants of much smaller matrices. Relatedly, we have

$$\begin{aligned}(\boldsymbol{y}_k - \boldsymbol{\mu}_k)'\boldsymbol{V}_k^{-1}(\boldsymbol{y}_k - \boldsymbol{\mu}_k) = \quad & \mathrm{tr}\left[(Y_k - \mathbf{1}_{n_k} \otimes \boldsymbol{\mu}_y')\boldsymbol{\Sigma}_w^{-1}(Y_k - \mathbf{1}_{n_k} \otimes \boldsymbol{\mu}_y)'\right] \; - \\ & n_k(\overline{\boldsymbol{y}}_k - \boldsymbol{\mu}_y)'\boldsymbol{\Sigma}_w^{-1}(\overline{\boldsymbol{y}}_k - \boldsymbol{\mu}_y) \; + \\ & n_k(\overline{\boldsymbol{y}}_k - \boldsymbol{\mu}_y)'(n_k\boldsymbol{\Sigma}_b + \boldsymbol{\Sigma}_w)^{-1}(\overline{\boldsymbol{y}}_k - \boldsymbol{\mu}_y),\end{aligned} \tag{29}$$

where $\boldsymbol{Y}_k$ is an $n_k \times p$ matrix of observations in Cluster k. This expression involves inverses of $\boldsymbol{\Sigma}_w$ and $\boldsymbol{\Sigma}_b$, both of which are of much smaller dimension than $\boldsymbol{V}_k$. (Note that the $\boldsymbol{Y}_k$ matrix should not be confused with $\boldsymbol{y}_k$, with the latter being a vector of length P_k containing all observations in Cluster k.)

As discussed by Rosseel (2021), additional simplifications are available when we wish to compute the overall log-likelihood across schools. These additional simplifications are especially worthwhile when n_k is equal for most or all k. The expressions can also be modified to handle Level-2 observed variables and missing data. In addition to being used for information criterion computations, the expressions can be used for fast and efficient MCMC estimation as demonstrated by recent versions of *blavaan*.

Appendix B: "Within-Only" Variables in Two-Level SEM

In this section, we illustrate how "within-only" variables in two-level SEM can lead to unexpected behavior of PSIS-LOO. As an example of a within-only variable, consider a situation where students complete three tests and also report the amount of anxiety that they felt during the tests. We might attribute test anxiety exclusively to individual students, as there is no reason to believe that anxiety is a property of a school. In this case, the observed anxiety variable still appears in Equation (15), but the corresponding row of $\boldsymbol{\Lambda}^{(c)}$ and the corresponding entry of $\boldsymbol{\epsilon}_k^{(c)}$ are fixed to 0. When this "within-only" observed variable actually exhibits meaningful across-school variability in the observed data, the effective number of parameters in WAIC and PSIS-LOO are again inflated.

Method

We generated data from a model with 4 observed variables and 200 clusters of size 50. The data-generating model can be written as:

$$\boldsymbol{y}_{ik} \mid \boldsymbol{\nu}_k \sim \mathrm{N}(\nu_k, \boldsymbol{\Sigma}_w) \quad i = 1, \ldots, 50; k = 1, \ldots, 200 \tag{30}$$

$$\boldsymbol{\nu}_k \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_b) \quad k = 1, \ldots, 200 \tag{31}$$

$$\boldsymbol{\Sigma}_w = \begin{pmatrix} 1.5 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1.5 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1.5 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1.5 \end{pmatrix} \tag{32}$$

$$\boldsymbol{\Sigma}_{\mathbf{b}} = \begin{pmatrix} 0.6 & 0.2 & 0.2 & 0 \\ 0.2 & 0.6 & 0.2 & 0 \\ 0.2 & 0.2 & 0.6 & 0 \\ 0 & 0 & 0 & \sigma_{44} \end{pmatrix} \tag{33}$$

where we manipulated the value of σ_{44} in $\boldsymbol{\Sigma}_b$. The σ_{44} parameter assumed values from 0 to 4 in steps of 0.5, where a new dataset was generated for each value of σ_{44}. We then fit a model where the fourth observed variable, y_4, is treated as "within only." In *(b)lavaan*, this "within only" model could be specified as

```
model <-'
    level: within
        fw =~ y1 + y2 + y3 + y4
    level: between
        fb =~ y1 + y2 + y3
'
```

Prior distributions were the same as those used in Example 2. For each of three chains, we used *blavaan* to obtain 2,000 samples following 500 burnin samples.

Results

Results are shown in Figure B1. The true "between" variance of y_4 is on the x-axis, effective number of parameter estimates are on the y-axis, and panels represent different information criteria. The horizontal line in each panel is the frequentist parameter count, which is 18 for this model. Similar to Example 2, we see that the DIC effective number of parameters is nearly flat and stays near the frequentist parameter count of 18. But under PSIS-LOO, the model appears to gain parameters as the between variability increases.

Figure B1

Magnitude of Between Variability in "Within-Only" Variable (X-Axis) vs. Estimated Effective Number of Parameters (Y-Axis) of DIC and PSIS-LOO