Received: 4 February 2021 | Revised: 22 September 2021 | Accepted: 23 September 2021

DOI: 10.1111/stan.12256

ORIGINAL ARTICLE

WILEY

Bayesian model selection for multilevel mediation models

Oludare Ariyo[1,2] | Emmanuel Lesaffre[1] | Geert Verbeke[1] | Martijn Huisman[3] | Martijn Heymans[3,4] | Jos Twisk[3]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), KU Leuven, Leuven, Belgium

[2] Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria

[3] Department of Epidemiology and Data Science, Amsterdam Public Health Research Institute, Amsterdam UMC, The Netherlands

[4] Department of Sociology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Correspondence
Oludare Ariyo, Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), KU Leuven, Leuven, Belgium.
Email: ariyoso@funaab.edu.ng

Funding information
The research of the first author was funded by Tertiary Education Trust Fund (TETFund) - AS&D grant of the Federal University of Agriculture, Abeokuta, Nigeria.

Mediation analysis is often used to explore the complex relationship between two variables through a third mediating variable. This paper aims to illustrate the performance of the deviance information criterion, the pseudo-Bayes factor, and the Watanabe–Akaike information criterion in selecting the appropriate multilevel mediation model. Our focus will be on comparing the conditional criteria (given random effects) versus the marginal criteria (averaged over random effects) in this respect. Most of the previous work on the multilevel mediation models fails to report the poor behavior of the conditional criteria. We demonstrate here the superiority of the marginal version of the selection criteria over their conditional counterpart in the mediated longitudinal settings through simulation studies and via an application to data from the Longitudinal Aging Study of the Amsterdam study. In addition, we demonstrate the usefulness of our self-written R function for multilevel mediation models.

KEYWORDS

deviance information criterion, marginalized likelihood, multilevel mediation models, pseudo Bayes factor, Watanabe–Akaike information criterion

1 | INTRODUCTION

Mediation analysis enables researchers to investigate the complex relationship between two variables through a third "mediating" variable. This indirect pathway through a mediating variable (or mediator) helps explain how exposure affects an outcome (MacKinnon, 2008). The concept of mediation being used in the estimation of single-level mediation for independent subjects from random sampling (Hayes, 2017; MacKinnon, 2008) and multilevel mediation (McNeish, 2017; Zigler & Ye, 2019) has broad applications in both biomedical and social science research.

Multilevel data is usually encountered in medicine where patients are nested within hospitals, or repeated measurements are nested within patients. However, this type of multilevel data violates the assumption of independence necessary for traditional regression methods (Zigler & Ye, 2019). Hence, several authors have examined the use of mediation in multilevel data using multilevel modeling (MLM)(Bauer, Preacher, & Gil, 2006; Krull & MacKinnon, 1999; MacKinnon, 2008; Preacher, Zyphur, & Zhang, 2010; Rusá, Komárek, Lesaffre, & Bruyneel, 2018) and the multilevel structural equation model (Lee, 2007; McNeish, 2017; Preacher et al., 2010; Yanuar, Ibrahim, & Jemain, 2013; Zigler & Ye, 2019).

Krull and MacKinnon (2001) established the terminology for a multilevel mediation design. They suggested the Predictor-Mediator-Outcome format, wherein a number indicates the level of data where each variable is located. For example, a 1-1-1 means all three variables are measured at level-1. Level 1 represents the lowest measurement level, for example, repeated measurements within a person, and that level 2 represents the cluster level, for example, the subjects. In a longitudinal study, this means that all variables are time dependent. Other designs include the 2-1-1 design where the predictor is at level 2, and the outcome and mediator have been measured as level 1. This is, for instance, true in an RCT (randomized controlled trial) with more than one follow-up measurement. The intervention variable is not time-independent, that is, measured at level 2. Figure 1 displays a 1-1-1 design for an MLM where the associations between the variables are split into between-cluster and within-cluster effects. For further details regarding this terminology, see Krull and MacKinnon (2001).

Several MLMs have been proposed in the literature; therefore, evaluating the most appropriate model that best fits the data would be a useful exercise. Model selection is a task of selecting an appropriate statistical model from the list of candidate models, given data. It is an important step in a statistical modeling exercise. In most mediation analyses, MLM models are typically evaluated for bias, coverage probability, and power (Blood & Cheng, 2011; Gao & Albert, 2018; Wang,

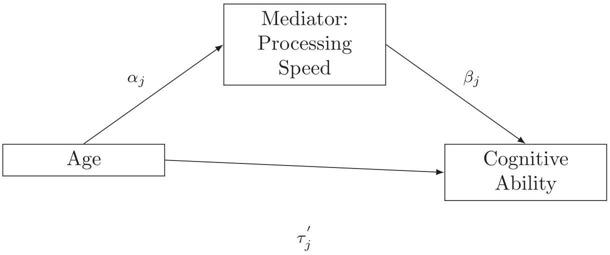

FIGURE 1 Longitudinal Aging Study of the Amsterdam (LASA) data example: Diagram for a two-level mediation model in which the effect of Age on Cognitive Ability is partially mediated by Processing Speed: Age affect Processing Speed (path α_j), Processing Speed affect Cognitive Ability (path β_j), and Age affect Cognitive Ability (path τ'_j)

Chen, Goldstein, Buck Louis, & Gilman, 2019; Zigler & Ye, 2019). Few authors use model selection criteria using a frequentist (Wu, Carroll, & Chen, 2018) or the Bayesian (Rusá et al., 2018) approach. Rusá et al. (2018) used the Watanabe–Alkaike Information criteria (WAIC) to compare their flexible, moderated mediation model with competing models. However, the authors did not consider the marginal version of the WAIC. Here, we compare the conditional and marginal Bayesian model selection criteria performance to select the most appropriate MLM model for the data at hand.

The choice between conditional and marginal criteria should be based on the aim of the study (Vaida & Blanchard, 2005). For instance, the marginal model selection criteria should be used when the study's aim is to estimate the predictiveness of the model when new subjects (as in longitudinal study) are involved. While the conditional criteria should be used when the aim is to estimate the predictiveness of the model when new elements in the cluster (in longitudinal studies, new observations from the existing subjects) are involved. The model selection criteria based on the likelihood from which the random effects have been integrated out are referred to as the marginal criteria whereas the model selection criteria based on the likelihood including the random effects, that is, the conditional likelihood, result in conditional criteria. This paper will compare the performance (the ability of the criteria to select an appropriate data-generating model) of the conditional and marginal versions of three popular Bayesian model selection criteria: the deviance information criteria (DIC), the Pseudo-Bayes factor (PSBF), and WAIC.

The conditional versions of the three model selection criteria discussed above are the most popular and easiest to compute from the generated Markov Chain Monte Carlo (MCMC) samples. Consequently, the marginal versions of the selection criteria are almost never reported. However, it has been demonstrated theoretically (see, e.g. Li & Yu, 2012; Chan & Grant, 2016a; Celeux, Forbes, Robert, & Titterington, 2006) and via simulation studies (see, e.g. Ariyo & Adeleke, 2021; Ariyo, Lesaffre, Verbeke, & Quintero, 2021; Ariyo et al., 2019a, 2019; Chan & Grant, 2016b; Merkle, Furr, & Rabe-Hesketh, 2018) that the use of the conditional likelihood is questionable. In the theoretical argument, for instance, Li and Yu (2012) proved that the conditional likelihood of the augmented data is nonregular and disprove the standard asymptotic arguments that are needed to justify the DIC. Celeux et al. (2006) compared different DIC constructions and found through theoretical studies that some of these constructions are simply not adequate for evaluating the complexity and fit of a model. Likewise, Chan and Grant (2016a) show that the numerical *SE*s of the conditional DICs are often too large to be useful for models' comparison. On practical grounds (Ariyo & Adeleke, 2021; Ariyo et al., 2019a, 2019, 2021; Chan & Grant, 2016b; Merkle et al., 2018; Millar, 2009) provide simulation studies in which the conditional DIC almost always favors an overfitted model.

In practice, especially in medical and social science research, the researchers often rely on the default software, presumably because of a lack of awareness of the marginal and conditional criteria' differences. In fact, in the analysis of mediation models, we are not aware of any previous work that distinguishes between the marginal and conditional criteria in the context of a multilevel mediation model. Hence, this paper illustrates the marginal version of the selection criteria' superiority over their conditional counterparts in mediated longitudinal settings. As such, we have aimed to conduct simulation studies to illustrate the performance of the conditional and marginal criteria in selecting the true data-generating models under different scenarios: (a) when the mediation paths are allowed to vary randomly across clusters (see, e.g. Raudenbush & Bryk, 2002; Zigler & Ye, 2019), (b) when the mediation paths are fixed across clusters (see, e.g. McNeish, 2017), (c) when the mediation path is zero, with "no mediation effects" (see, e.g. Zigler and Ye (2019) and (d) when the distributions of the mediation paths are misspecified.

The outline of the paper is as follows. In Section 2, we introduce the Longitudinal Aging Study of the Amsterdam (LASA) dataset. The basic concepts of mediation, moderation, and the combination thereof are summarized in Section 3. We briefly discuss the model selection criteria in Section 4. Different simulation settings and scenarios are presented in Section 5, while we illustrate the comparison between the conditional and marginal criteria on LASA data in Section 6. Finally, concluding remarks are given in Section 7.

2 | THE LONGITUDINAL AGING STUDY AMSTERDAM

LASA is a prospective cohort study intended to determine the predictors and consequences of ageing, specifically physical, cognitive, emotional, and social functioning in older adults (aged 55–85) in the Netherlands. The participants were sampled from the registries of urban and rural municipalities in different parts of the country. The baseline measurement took place in 1992/1993, and follow-up measurements have been conducted since then about every 3 years. The data collection consists of the main interview, a self-reported questionnaire, and a medical interview. The example in this paper was initially analyzed and published by Robitaille, Piccinin, Muniz-Terrera, et al. (2013), who examined processing speed (M) as a mediator between age (X) and cognitive abilities (Y). Processing speed was based on a coding task adapted from the Alphabet Coding Task-15. The reasoning was based on the adapted version of the Raven Colored Progressive Matrices test, and the cognitive ability was based on the 15 Word Test. The cognition has been based on three different measures: (a) immediate recall, (b) delayed recall, and (c) reasoning as the outcome variable. More information can be found in Robitaille et al. (2013).

Each model was based on data from the first five waves of the LASA study. Respondents were excluded from the analyses if they had a 23 or lower score on the Mini-Mental State Examination during any of the five waves ($n = 798$) or if they had missing education information ($n = 3$). The analytical cohort consisted of 2,306 respondents in the first wave, of which 1,883 also participated in the second wave (81.7%), with a further 1,562 in the third wave (83.0%), 1,300 in the fourth wave (83.2%), and 1,021 in the fifth wave (78.5%). Detailed information on the LASA can be found in (Hoogendijk et al., 2016; Huisman et al., 2011; Robitaille et al., 2013).

Robitaille et al. (2013) applied a lower-level mediation 1-1-1 model since all variables M, X, and Y were measured at level-1. Here, we aim to illustrate the performance of the conditional and marginal criteria in the context of a multilevel mediation model using a 1-1-1 MLM model.

3 | MULTILEVEL MEDIATION MODEL

Before discussing the 1-1-1 multilevel mediation model in detail, we explain a multilevel mediation model's basic framework. When an outcome Y and a predictor X are mediated by a mediator M, it means that M is correlated with X and explains the effect of X on Y. With a continuous outcome Y and a mediator M, a single-level mediation equation is given as

$$\begin{aligned} M_i &= \beta_1 + \alpha X + e_M \\ Y_i &= \beta_2 + \beta M_i + \tau' X + e_Y, \end{aligned} \tag{1}$$

where β_1 and β_2 denote intercept for mediator and outcome, respectively, and e_M and e_Y denote error terms in the equations. The direct effect of X on Y is denoted as τ' and the indirect effect of $\mathbf{X}$

on **Y** through the mediator **M** is expressed as the product of α and β, that is, $\alpha\beta$. It is important to note here the four key assumptions in traditional mediation analysis, including (i) no unmeasured confounder between the exposure variable X and the response variable Y; (ii) no unmeasured confounder between the exposure variable X and the mediator M; (iii) no unmeasured confounder between the mediator M and the response variable Y; and (iv) any mediator M_i is not causally prior to M_{-i}, (the vector of mediators M without M_i.) Recently, Cao, Li, and Yu (2021) performed sensitivity analysis of the impact of violation of assumptions on the estimation of mediation effects using Yu et al.'s mediation analysis method.

Given that all variables are measured at level 1, the estimate of model parameters is straightforward using standard OLS-regression or maximum likelihood methods (Preacher & Selig, 2012; Song, 2018). However, the direct application of such techniques to multilevel data (such as mediation data) thereby ignoring the nested structure of the data will statistically bias (see also Raudenbush & Bryk, 2002; Tom, Bosker, & Bosker, 2012) the estimates and the conclusions of the analysis. Hence, we consider a set of standard MLM equations predicting Y from X including a random effects structure (random intercept and slope(s)). Following the notation in Yuan and MacKinnon (2009), one can write a two-level lower mediation model (as given in Figure 1) with level 1 equations as:

$$\begin{aligned} M_{ij} &= \beta_{1j} + \alpha_j X_{ij} + e_{M_{ij}} \\ Y_{ij} &= \beta_{2j} + \beta_j M_{ij} + \tau_j' X_{ij} + e_{Y_{ij}}, \end{aligned} \tag{2}$$

and level 2 is given as

$$\begin{aligned} \beta_{1j} &= \beta_3 + u_{1j} \\ a_j &= \alpha + u_{2j} \\ \beta_{2j} &= \beta_4 + u_{3j} \\ \beta_j &= \beta + u_{4j} \\ \tau_j' &= \tau' + u_{5j}, \end{aligned}$$

where $e_{M_{ij}}$ and $e_{Y_{ij}}$ are level 1 error terms for M and Y, respectively; subscript i and j refer to individual and level-2 units; the parameters β_{1j} and β_{2j} are random intercepts, and α_j, β_j and τ_j' are random slopes. The parameters β_2 and β_3 are population (or average) effects. For MLM, the first-level residuals $e_{M_{ij}}$ and $e_{Y_{ij}}$ are assumed to be independent and follow normal distribution, that is, $e_{M_{ij}} \sim N(0, \sigma^2_{e_{M_{ij}}})$ and $e_{Y_{ij}} \sim N(0, \sigma^2_{e_{Y_{ij}}})$ and the second-level residuals $\mathbf{u}_j = (u_{1j}, u_{2j}, u_{3j}, u_{4j}, u_{5j})^T$ follow a multivariate normal distribution $\mathbf{u}_j \sim N(\mathbf{0}, \mathbf{D})$ where $\mathbf{D}$ is 5×5 covariance matrix.

In multilevel mediation, the average indirect effects in the population are often of primary interest. Yuan and MacKinnon (2009) gave the average indirect effects (applies to models with only random slopes) formula to be

$$ab = E(\alpha_j \beta_j) = \alpha\beta + \sigma_{\alpha_j \beta_j}, \tag{3}$$

where $\sigma_{\alpha_j \beta_j}$ denotes the covariance between α_j and β_j.

MacKinnon (2008) and Kenny, Korchmaros, and Bolger (2003) also showed that the total effect in a fully random, lower mediated multilevel model is

$$c = \tau' + \alpha\beta + \sigma_{\alpha_j \beta_j}, \tag{4}$$

and the relative average indirect effect can be expressed as

$$ab/c = \frac{\alpha\beta + \sigma_{\alpha_j\beta_j}}{\tau' + \alpha\beta + \sigma_{\alpha_j\beta_j}}. \tag{5}$$

Equation (5) is often referred to as the proportion mediated in the mediation analysis literature (Ananth, 2019; Ditlevsen, Christensen, Lynch, Damsgaard, & Keiding, 2005). This statistic has some important disadvantages. For example, the proportion mediated effect cannot be used when the mediation model is inconsistent (i.e., the direct and indirect effect have a different sign), which is actually the case in the LASA data example (see Robitaille et al. (2013). In these situations, the proportion mediated can exceed 1 and can be below 0 and the interpretation may become meaningless (as the limits of a proportion are 0 and 1).

In practice, the heterogeneity in the causal effects across level 2 units may be of scientific interest. For example, in Section 6, we analyze the LASA dataset to investigate whether the processing speed mediates between age and cognitive abilities in older adults. The importance of random effects in the lower level mediation (1-1-1 model in particular) was pointed out first in Kenny et al. (2003). For model represented in Equation (2) to be estimable and ensure that the mediational effects are unbiased, some assumptions are required as given below:

1. The predictors X_{ij} must be uncorrelated with the random intercepts and slopes and with β_{2j}, β_j, τ_j', and $e_{Y_{ij}}$.
2. The residuals e_{M_j} and $e_{Y_{ij}}$ are normally distributed with mean zero and uncorrelated with each other.
3. The level 1 residuals are uncorrelated with the random effects that is, e_{M_j} is uncorrelated with β_{1j}, a_j, β_{2j}, β_j, and τ_j'.

It is important to note that some of these assumptions may not hold in practice. In this paper, we considered the performance of the conditional and marginal selection criteria when these assumptions are violated.

4 | BAYESIAN MODEL SELECTION

We considered three different Bayesian model selection criteria for evaluating a multilevel mediation model: PSBF, DIC, and WAIC. We further distinguished between the marginal and conditional version of these criteria. For MLM, let $\mathbf{\Theta}$ represent all the model parameters, the distinction is that the marginal MLM includes the fixed effects (i.e., the intercept for mediator and outcome, when assume fixed) and the parameters making up the covariance matrix of the random effects. Conversely, the conditional MLM includes the random effects (i.e., the direct and indirect pathway of the model) in the $\mathbf{\Theta}$.

Further, we denote the collected (longitudinal) mediated outcomes by $\mathbf{y}$ and the obtained covariate values by the matrix $\mathbf{X}$ moderated by $\mathbf{M}$. The posterior distribution is $p(\theta|\mathbf{y}, \mathbf{X}, \mathbf{M}) = p(\mathbf{y}|\mathbf{\Theta}, \mathbf{X})p(\mathbf{\Theta})/p(\mathbf{y}|\mathbf{X}, \mathbf{M})$. When the closed-form of this posterior distribution does not exist then it is appropriate to use MCMC methods. Namely, K (dependent) values $\mathbf{\Theta}^1, \dots, \mathbf{\Theta}^K$ are sampled from the posterior distribution. The true posterior summary measures can then be approximated by their sampled versions.

Recently, some authors have compared the performance of these criteria in different application studies (see e.g. Ariyo & Adeleke, 2021; Dey, Delampady, & Gopalaswamy, 2019; Millar, 2018). However, there is still no consensus about the best criterion for model selection in a Bayesian context. For the distinction between the performance of the marginal against the conditional criteria, other authors have shown that marginal criteria outperform the conditional criteria in most settings for LMMs with some extensions (Ariyo et al., 2019a, 2019) and generalised linear mixed models (GLMMs) (Ariyo et al., 2021; Millar, 2018; Quintero & Lesaffre, 2018). This is also true for an item response model (Li, Qiu, Zhang, & Feng, 2016; Merkle et al., 2018; Millar, 2018). However, to our knowledge, the marginal criteria' superiority over the conditional criteria has not been demonstrated in the mediation analysis literature. We will (briefly) discuss the three Bayesian model selection criteria in the subsequent sections for reasons of completeness.

4.1 | The pseudo Bayes factor

Model comparison using Bayes' factors requires the computation of the marginal likelihood of two competing models. Given a model $\mathcal{M}$ and model parameters $\theta_{\mathcal{M}}$, we assume that the data $y_1, \dots, y_n$ are conditionally independent. The marginal likelihood is given by:

$$p(\mathbf{y}|\mathcal{M}) = \int_{\theta_{\mathcal{M}}} \prod_{i=1}^{n} p(y_i|\theta_{\mathcal{M}}, \mathcal{M})p(\theta_{\mathcal{M}})d\theta_{\mathcal{M}}. \tag{6}$$

However, Equation (6) is not analytically available in general. Therefore, Geisser and Eddy (1979) suggested replacing (6) by the pseudo marginal likelihood

$$\hat{p}(\mathbf{y}|\mathcal{M}) = \prod_{i=1}^{n} p(y_i|y_{-i}, \mathcal{M}), \tag{7}$$

where $\prod_{i=1}^{n} p(y_i|y_{-i}, \mathcal{M})$ is the ith conditional predictive ordinate (CPO$_i$) and the predictive density calculated at the observed y_i given y_{-i}, which is the set of all data except the ith observation. The PSBF is then obtained by taking the ratio $\hat{p}(\mathbf{y} \mid \mathcal{M}_1)/\hat{p}(\mathbf{y} \mid \mathcal{M}_2)$ to evaluate the preference of model $\mathcal{M}_1$ over model $\mathcal{M}_2$. Low value of this ratio reflect preference of model $\mathcal{M}_2$ based on the current data. In practice, one often evaluates the logarithm of expression (7), leading to the log pseudo marginal likelihood (LPML) for model $\mathcal{M}_r$ is given as $\text{LPML}_r = \sum_{i=1}^{n} \log(\text{CPO}_{r,i})$ where

$$\text{CPO}_{r,i} \approx \left[\frac{1}{K}\sum_{k=1}^{K} \frac{1}{p(\mathbf{y}_i|\theta_r^k, \mathcal{M}_r)}\right]^{-1},$$

and θ_r^k represents the model parameters for model $\mathcal{M}_r$.

4.2 | The deviance information criterion

Deviance is defined as $D(\boldsymbol{\Theta}) = -2\log p(\mathbf{y}|\theta)$. The DIC is then defined as $\text{DIC} = -2\log p(\mathbf{y}|\overline{\theta}) + 2p_{\text{DIC}}$, where $\overline{\theta}$ is the posterior mean of the model parameter (parameter in focus), that is, $\overline{\theta} =$

$E(\theta|y)$ and p_{DIC} corresponds to the effective number of parameters, given by

$$p_{\text{DIC}} = -2\, E_{\theta|\mathbf{y}}[\log p(\mathbf{y}|\boldsymbol{\theta})] + 2\log[p(\mathbf{y}|\overline{\boldsymbol{\theta}})]. \tag{8}$$

Two versions of p_{DIC} are generally used (Gelman, Hwang, & Vehtari, 2014; Spiegelhalter, Best, Carlin, & van der Linde, 2014): (i) p_{DIC} in (8) which is considered to be numerically stable, and (ii) $p_{\text{DIC}_2} = 2\ \text{Var}_{\theta|\mathbf{y}}[(\mathbf{y}|\overline{\boldsymbol{\theta}})]$ which has the advantage of being always positive (Gelman et al., 2014). Consequently, Celeux et al. (2006) suggested several forms for the DIC that can be used for different hierarchical models and Ariyo et al. (2019a) have compared the performance of the marginal and conditional versions of these DIC versions and have shown that there are inconsistencies in the performance of the conditional versions of different forms of DIC whereas the marginal versions perform similarly. Supposed an additional vector of latent variables $\boldsymbol{\mu}$ with density $p(\boldsymbol{\mu}|\overline{\boldsymbol{\theta}})$ is added to the model $p(y|\overline{\boldsymbol{\theta}})$ the we have:

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\theta})d\boldsymbol{\mu} = \int p(\boldsymbol{y},\boldsymbol{\mu}|\boldsymbol{\theta})d\boldsymbol{\mu}, \tag{9}$$

where $p(\boldsymbol{y}|\theta, \boldsymbol{\mu})$ is the conditional likelihood and $p(\boldsymbol{y}|\theta)$ is the integrated likelihood. The marginal DIC (mDIC) is obtained by integrating the likelihood in Equation (9). As such, the definition of mDIC from integrated likelihood is given as

$$mDIC = -4\boldsymbol{E}_{\theta}[\log p(\boldsymbol{y}|\boldsymbol{\theta})|\boldsymbol{y}] + 2\log p(\boldsymbol{y}|\tilde{\boldsymbol{\theta}}). \tag{10}$$

Consequently, the alternative definition of DIC via conditional likelihood (cDIC) is given as

$$cDIC = -4\boldsymbol{E}_{\theta,\mu}[\log p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\mu})|\boldsymbol{y}] + 2\log p(\boldsymbol{y}|\tilde{\boldsymbol{\theta}},\tilde{\boldsymbol{\mu}}),$$

where $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\mu}})$ is the joint maximum a posterior estimate of the pair $(\boldsymbol{\mu}, \theta)$ given the data $\boldsymbol{y}$ (see Celeux et al., 2006). Despite its popularity and availability in Bayesian software, DIC has been criticized, see Spiegelhalter et al. (2014) for details. For instance, DIC is not invariant to nonlinear transformations of $\boldsymbol{\theta}$ and negative value for p_{DIC} can occur in some cases. The major setback of mDIC is computational difficulties since the integral in Equation (10) is generally intractable, notwithstanding, mDIC as been found to show superior performance in most cases (Ariyo & Adeleke, 2021; Ariyo et al., 2019a, 2021; Ariyo, Lesaffre, et al., 2019; Chan & Grant, 2016a; Quintero & Lesaffre, 2018).

4.3 | Watanabe–Akaike information criterion

The WAIC is a Bayesian version of the AIC (Watanabe, 2010) and a worthy successor of DIC (Spiegelhalter et al., 2014) as it uses the posterior predictive distribution of the data to estimate the out-of-sample predictive accuracy of the model. The WAIC is then defined as

$$\text{WAIC} = -2\widehat{\text{lppd}} + 2p_{\text{WAIC}},$$

where p_{WAIC} corresponds to an estimate of the effective number of parameters given by

$$p_{\text{WAIC}} = 2\sum_{i=1}^{n}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K} p(\mathbf{y}_i|\boldsymbol{\theta}^k)\right) - \frac{1}{K}\sum_{k=1}^{K}\ \log\ p(\mathbf{y}_i|\boldsymbol{\theta}^k)\right].$$

and $\widehat{\text{lppd}}$ which can be estimated using an MCMC sample from the posterior distribution as

$$\widehat{\text{lppd}} = \sum_{i=1}^{n} \log \left[\frac{1}{K} \sum_{k=1}^{K} p(\mathbf{y}_i | \boldsymbol{\theta}^k) \right].$$

Similar to DIC, a model with smaller WAIC is preferred. One advantage of WAIC is its invariability to the scale of the model parameters, which implies that WAIC does not change when $\boldsymbol{\theta}$ is replaced by $\boldsymbol{\psi} = h(\boldsymbol{\theta})$, where h a strictly monotone function. For all the three model selection criteria, the model with the smallest value is preferred.

5 | SIMULATION STUDIES

We performed simulation studies to illustrate the conditional and marginal criteria' performance in MLMs. We opted to use 1-1-1 mediation models with random slopes for our simulation studies because this kind of model has been a model of interest in the fundamental work behind multilevel mediation (see Preacher, Zhang, & Zyphur, 2011) and has been commonly used in empirical studies (McNeish, 2017). A 1-1-1 design allows for the modeling of both the between and within components of the indirect effects and model pathways (Zhang, Zyphur, & Preacher, 2009). Additionally, we were motivated by previous investigations using LASA data (see Robitaille et al., 2013)

5.1 | Simulation study 1

Following the example of Kenny et al. (2003) and subsequently used in Bauer et al. (2006); we conducted a simple simulation study. Here, we generated 500 datasets based on model (2) with the following parameters: the random intercept for the mediator β_{1j} had a mean 0 and variance of 0.6, that is, $\beta_{2j} \sim N(0, 0.6)$, the random intercept for the response $\beta_{2j} \sim N(0, 0.4)$. These two random intercept β_{1j}, β_{2j} were normally distributed. The level 1 variance of response (σ^2_{eY}) and the mediator (σ^2_{eM}) were set to 0.45 and 0.65, respectively, the α_j and β_j paths were normally distributed with a mean of 0.6 and a variance of 0.16, while $\tau'_j \sim N(0.2, 0.4)$. The covariance between α_j and β_j was 0.113, that is, $\sigma_{\alpha_j\beta_j} = 0.113$, yielding a correlation of 0.706. We assume that neither α_j and β_j was correlated with τ'_j. We further simulated a predictor X_{ij} from $X_{ij} = \overline{X}_i + e_{X_{ij}}$, where $\overline{X}_i \sim N(0, 1)$ and $e_{X_{ij}} \sim N(0, 1)$.

In addition to Equation (2) as the true model, we fitted two alternative models: (i) Equation (2) without mediation effects $Y_{ij} = \beta_{2j} + \tau'_j X_{ij} + e_{Y_{ij}}$, (ii) Equation (2) without direct effect component (i.e., $\tau'_j = 0$), to evaluate the ability of the conditional and marginal criteria to select the data-generating model. We further varied the number of clusters understudy from 10 to 100 (10, 25, 50, 100), as this number of clusters is similar to those previously used in the literature (Bauer et al., 2006; McNeish, 2017). We also set the number of observations per level 2 units to $m_j = 4, 8, 16, \text{and} 32$, which is consistent with those used by Krull and MacKinnon (2001) and Bauer et al. (2006).

Regarding the choice of the prior distributions, we assigned independent noninformative uniform priors on regression parameters $p(\beta_2, \alpha, \beta, \tau')$. For the first-level variance parameters $e_{M_j}, e_{Y_{ij}}$ we assigned an inverse-gamma prior, while we assigned an Inverse-Wishart (IW$(k, \boldsymbol{S})$) for the second-level covariance matrix $\boldsymbol{U}$ ($\boldsymbol{U}_j \sim (\boldsymbol{0}, \boldsymbol{D})$). As suggested in JAGS (Plummer, 2013), the value of the degree of freedom k is taken as 5 (the rank of $\boldsymbol{U}$) while the scale matrix is a diagonal matrix with small values such as 0.001 at the diagonal. All model parameters in the simulation studies were estimated based on three chains of 10,000 iterations (discarding the first 5,000) and a thinning factor equal to 10 to avoid correlation problems in the generated chains. The convergence of the MCMC samples was assessed using the Brooks–Gelman–Rubin (BGR) diagnostic (Brooks & Gelman, 1998; Gelman & Rubin, 1992), and in cases where the BGR was larger than 1.1, a new MCMC sample was selected with 10,000 extra iterations until convergence was obtained. The models were implemented using rjags (Plummer, 2016) software.

Table 1 displays the performance of the conditional and marginal criteria for a 1-1-1 multilevel model in identifying the correct model when the mediation pathway was allowed to vary in a normal, random fashion across clusters. As expected, the performance of both versions of the criteria gets better as sample sizes increase. However, this increase is less obvious when the sample size is larger than 25. Similarly, as the number of observations increases from 8 to 32, the performance of both versions of the criteria are less obvious. An increase in observation units of more than eight has a minimal impact on both versions of the criteria' performance. As such, these results are in agreement with the results previously described in the literature. For instance, McNeish (2017) concluded that, if the right precautions have been taken, only a few clusters and observations are needed to provide reliable results.

Additionally, McNeish and Stapleton (2016) suggested that 20 clusters with five or more observations per cluster might be sufficient if the model is estimated with restrictive maximum likelihood. Overall, the marginal criteria outperformed the conditional ones, which is in line with the other results previously obtained in the literature (see Ariyo et al., 2019a, 2019, 2021; Chan & Grant, 2016b; Merkle et al., 2018; Quintero & Lesaffre, 2018).

5.2 | Simulation study 2

We evaluated the conditional and marginal criteria' performance when the distribution of the mediation paths was misspecified. As such, we generated 500 data sets based on Equation (2) with the modification that the distributions of the random slopes α_j and β_j are generated from $\chi^2(3)$ distribution, which has skewness 1.63 and kurtosis 4, which closely resembles the skew-normal distribution (see e.g. Wang, Boyer, Genton, et al., 2004). In addition to the data-generating model, we fit two alternative models: A model with mediation paths (i) assumed normal and (ii) assumed skew-normal.

The percentage of the times that each criterion selected a data generating model is displayed in Table 2. The results demonstrate that when the mediation paths are assumed to be skewed, both criteria' performance is better than if normality is assumed for the mediation pathways. These results show that the assumption that the average mediation paths (especially the indirect effects) follow a normal distribution might be unrealistic. This is why several authors warn researchers against making these assumptions for hypothesis testing (Hayes & Scharkow, 2013; MacKinnon, Lockwood, & Williams, 2004; Preacher & Selig, 2012; Song, 2018) and recommend approaches that relax the normality assumption, such as using bootstrap confidence interval (Efron & Tibshirani, 1994; MacKinnon et al., 2004; MacKinnon, Fritz, Williams, & Lockwood, 2007) and Monte Carlo (MC) simulations (Preacher & Selig, 2012) among others. Regardless of the mediation pathways' assumptions, the marginal criteria display superior performance compared to the conditional criteria.

TABLE 1 The percentage of times the conditional and marginal criteria select the true model when mediation path are random vary across clusters under different sample sizes and number of observation per units

	Sample size					
Number of observation/units	**Criteria**	**10**	**25**	**50**	**100**	**200**
4	cDIC	65.4	71.4	77.2	77.8	78.2
	cWAIC	62.0	68.6	73.4	74.6	75.6
	cPSBF	62.0	67.6	73.6	73.4	78.2
	mDIC	71.4	79.8	84.4	86.0	89.2
	mWAIC	72.2	80.2	84.8	85.8	88.6
	mPSBF	71.6	80.4	82.6	82.8	89.0
8	cDIC	69.4	76.8	78.2	78.2	79.8
	cWAIC	67.6	77.4	79.6	80.0	83.6
	cPSBF	67.4	77.6	78.8	80.2	82.6
	mDIC	79.4	80.4	84.4	86.2	88.0
	mWAIC	73.6	84.8	86.8	86.4	89.2
	mPSBF	73.8	88.6	88.8	89.0	89.8
16	cDIC	76.0	79.4	80.0	80.2	81.4
	cWAIC	75.6	78.2	80.2	80.0	81.4
	cPSBF	73.6	74.8	79.2	80.0	80.8
	mDIC	86.6	89.4	89.8	89.8	89.8
	mWAIC	89.6	89.8	89.6	90.0	90.6
	mPSBF	86.0	89.8	89.8	90.2	90.8
32	cDIC	76.2	78.8	80.4	82.4	84.2
	cWAIC	74.8	78.6	81.2	84.0	84.6
	cPSBF	73.0	78.0	80.6	83.6	84.8
	mDIC	86.2	89.0	90.0	90.8	93.0
	mWAIC	87.8	89.2	90.2	93.8	93.4
	mPSBF	85.8	89.0	90.0	94.2	95.0

5.3 | Simulation study 3

In the 1-1-1 model, the mediation paths can be estimated as fixed effects. However, when the paths are not allowed to randomly vary across clusters, a large number of clusters may lead to convergence problems while also diminishing the quality of the estimates (McNeish, 2017). Here, we have illustrated the performance of the conditional and marginal DIC, WAIC, and PSBF when the mediation pathways are not allowed to be random. Furthermore, we evaluated this condition under a variety of clusters and a different number of observations per level 2 units, as described in Section 5.1. As such, we generated 500 datasets based on Equation (1) with the following value for each parameters: (i) indirect effect component $\alpha = \beta = 0.40$, and (ii) direct effect component $\tau' = 0.40$. Additionally, $\beta_1 = 0.45$ and $\beta_2 = 0.45$.

TABLE 2 The percentage of times the conditional and marginal criteria select true model when mediation are not normally distributed across clusters under different sample sizes

	Sample					
Path distribution	**Criteria**	**10**	**25**	**50**	**100**	**200**
Skew-normal	cDIC	69.4	76.8	78.2	78.2	79.8
	cWAIC	67.6	77.4	79.6	80.0	83.6
	cPSBF	67.4	77.6	78.8	80.2	82.6
	mDIC	79.4	80.4	84.4	86.2	88.0
	mWAIC	73.6	84.8	86.8	86.4	89.2
	mPSBF	73.8	88.6	88.8	89.0	89.8
Normal	cDIC	45.4	54.4	67.2	72.8	78.2
	cWAIC	42.0	55.6	73.4	72.6	75.6
	cPSBF	42.0	57.6	63.6	70.4	78.2
	mDIC	51.4	68.8	76.4	81.0	87.2
	mWAIC	52.2	76.2	73.8	83.8	85.6
	mPSBF	51.6	78.0	70.6	82.8	88.0

TABLE 3 The percentage of time selection criteria select true model when the mediation pathways are fixed

Sample size					
Criteria	**10**	**25**	**50**	**100**	**200**
cDIC	86.3	93.6	98.6	99.2	100.0
cWAIC	83.0	94.4	98.6	99.0	100.0
cPSBF	83.6	94.0	98.6	99.6	100.0
mDIC	86.3	93.6	98.6	99.2	100.0
mWAIC	83.0	94.4	98.6	99.0	100.0
mPSBF	83.6	94.0	98.6	99.6	100.0

We fitted three alternative models: (i) Model (1) (ii) Model (1) without mediation effect (i.e., $\beta = 0$), (iii) Model (1) without direct effect component (i.e., $\tau' = 0$). The performance of the marginal and conditional criteria in identifying the correct data-generating model when the distribution of the mediation paths have been fixed has been presented in Table 3. The results show that regardless of the sample sizes, there is no difference between the performance of the conditional and marginal criteria.

6 | ANALYSIS OF LASA DATA

Here, we illustrate the performance of the conditional and marginal criteria using the LASA data described in Section 2. We fitted four different models: (i) Model "A" based on equation (2), (ii)

Model "B" based on Equation (11), (iii) Model "C" based on Equation (12) and (iv) Model "D" based on Equation (2) without mediation effects.

$$\begin{aligned} M_{ij} &= \beta_{1j} + \alpha_j X_{ij} + e_{M_{ij}} \\ Y_{ij} &= \beta_{2j} + \beta_j M_{ij} + e_{Y_{ij}}. \end{aligned} \tag{11}$$

$$\begin{aligned} M_{ij} &= \beta_{1j} + \alpha_j X_{ij} + e_{M_{ij}} \\ Y_{ij} &= \beta_{2j} + \beta_j M_{ij} + \tau_j' X_{ij} + e_{Y_{ij}}, \end{aligned} \tag{12}$$

where $\begin{pmatrix} a_j \\ b_j \end{pmatrix} \sim (\mathbf{0}, \mathbf{D})$ and $\mathbf{D}$ is a 2×2 covariance matrix. For each of the models, we calculate the level-specific indirect effects (Equation (3)) and the level-specific total effects (Equation (4)).

The priors used has been described in Section 5.1. We used 10,000 iterations, which after discarding the first 5,000 as burn-in and thinning was set to 10. The convergence of the MCMC samples was assessed using the BGR criteria. In addition to the estimates of model parameters, we compute the conditional and marginal criteria for each model. The results are displayed in Table 4. It can be observed that the conditional criteria support Model "C" and "D." These models assume that the level two parameters β_{1j} and β_{2j} are fixed. This seems to be an incorrect model since the 1-1-1 model assumed these parameters to be random across subjects. In contrast, the marginal criteria favor model "A" which seems to be the most appropriate 1-1-1 mediation model. This confirmed the results of the simulation that marginal criteria often outperform the conditional criteria in selecting the most appropriate model.

7 | CONCLUSION

We compared three Bayesian selection criteria in the context of multilevel mediation models. Our focus was on illustrating the conditional and marginal criteria' performance in selecting the true, data-generating model under different distributional assumptions for mediation pathways. The simulation studies' results demonstrated the superior performance of the marginal criteria when the mediation pathways are assumed to be random. The conditional criteria often select over-specified mediation pathways, while the marginal criteria select the correct model often. Conversely, when the mediation pathways are assumed to be fixed for the 1-1-1 mediation model, the marginal and conditional criteria' performance is essentially the same. Both the conditional and marginal criteria prefer (often) the correct model when the mediation pathways are fixed. However, the motivation for assuming fixed or random pathways should be based on the research question. The result from the LASA dataset analysis also confirmed the results obtained in the simulation studies as the results of the marginal criteria were consistent with the summary statistics.

These results confirm the results of (Ariyo et al., 2019a, 2019) for LMMs, Chan and Grant (2016a) for volatility models, (Li et al., 2016; Millar, 2018) for item response models, Merkle et al. (2018) for latent variable model and (Ariyo et al., 2021; Quintero & Lesaffre, 2018) for GLMM. To encourage the applied researchers to use the marginal criteria, we provide an R function that computes not only the marginal criteria but also the conditional criteria with minimal computational effort. The function is available at https://github.com/OludareAriyo/Bayesselect

The choice of a 1-1-1 multilevel mediation model with three variables was motivated by its popularity in the literature and motivated by the LASA dataset. We believe that the results are

TABLE 4 Longitudinal Aging Study of the Amsterdam (LASA) dataset: Posterior mean (estimates of a lower level mediation model), 95% probability interval and the conditional and marginal criteria under four fitted models

	Model A			Model B			Model C			Model D		
Effects	Estimates	25%	95%	Estimates	25%	95%	Estimates	25%	95%	Estimates	25%	95%
$\alpha\beta$	0.4306	0.4160	0.4437	0.3771	0.3742	0.8021	0.2013	0.1980	0.2046	0.2690	0.2571	0.2778
τ	0.3655	0.3612	0.3696	0.3771	0.3742	0.8021	0.2013	0.1980	0.2046	0.1570	0.1477	0.1699
α	0.6804	0.6778	0.6830	0.6809	0.6786	0.6832	0.3895	0.3825	0.3962	0.3900	0.3831	0.3969
β	0.6323	0.6115	0.6519	0.5534	0.5489	0.5581	0.5145	0.5098	0.5190	0.6877	0.6620	0.7041
τ'	−0.0623	−0.0783	−0.0508	—	—	—	−0.0130	−0.1980	−0.2046	−0.1120	−0,1477	−0,1699
$\sigma^2_{\alpha_j}$	0.0004	0.0004	0.0005	0.0004	0.0003	0.0005	0.0037	0.0032	0.0043	0.0037	0.0032	0.0010
$\sigma^2_{\beta_j}$	0.0017	0.0015	0.0020	0.0061	0.0002	0.0002	0.0015	0.0012	0.0018	0.0012	0.0010	0.0014
$\sigma^2_{\alpha_j\beta_j}$	0.0003	0.0002	0.0004	0.0003	0.0003	0.0004	0.0009	0.0008	0.0010	0.0008	0.0006	0.0011
$\sigma^2_{\tau'_j}$	−0.0561	−0.0783	−0.0508	—	—	—	—	—	—	−0.0034	−0.0031	−0.0042
cDIC		−32,226.53			−32,226.89			−32,234.27			−32,229.60	
cWAIC		−32,429.84			−32,430.11			−32,436.01			−32,432.23	
cLPPD		−16,032.55			−16,029.91			−16,033.89			−16,032,05	
mDIC		−2,907.24			−2,626.53			−2,892.17			−2,569.05	
mWAIC		−2,916.62			−2,614.84			−2,887.56			−2,567.78	
mLPPD		−1,283.78			−1,246.82			−1,243.78			−1,284.47	

valid for other multilevel mediation models as well. When more variables are involved in mediation analysis, the clusters' effects are likely to impact model selection criteria performance. Hence, further studies could derive the likelihood for more complex mediation models to evaluate Bayesian model selection's performance in more complex settings.

ORCID

Oludare Ariyo https://orcid.org/0000-0003-3375-1831
Emmanuel Lesaffre https://orcid.org/0000-0002-3747-6905
Geert Verbeke https://orcid.org/0000-0001-8430-7576
Martijn Huisman https://orcid.org/0000-0002-9285-6759
Martijn Heymans https://orcid.org/0000-0002-3889-0921
Jos Twisk https://orcid.org/0000-0001-9617-1020

REFERENCES

Ananth, C. V. (2019). Proportion mediated in a causal mediation analysis: How useful is this measure? *BJOG: An International Journal of Obstetrics & Gynaecology*, *126*(8), 983–983.

Ariyo, O., & Adeleke, M. (2021). Simultaneous Bayesian modelling of skew-normal longitudinal measurements with non-ignorable dropout. *Computational Statistics*, 1–23. https://link.springer.com/article/10.1007/s00180-021-01118-y

Ariyo, O., Lesaffre, E., Verbeke, G., & Quintero, A. (2019). Model selection for Bayesian linear mixed models with longitudinal data: Sensitivity to the choice of priors. *Communications in Statistics - Simulation and Computation*, 1–25. https://www.tandfonline.com/doi/abs/10.1080/03610918.2019.1676439

Ariyo, O., Lesaffre, E., Verbeke, G., & Quintero, A. (2021). Bayesian model selection for longitudinal count data. *Sankhya B*, (0), 1–24.

Ariyo, O., Quintero, A., Muñoz, J., Verbeke, G., & Lesaffre, E. (2019). Bayesian model selection in linear mixed models for longitudinal data. *Journal of Applied Statistics*, *47*(5), 890–913.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, *11*(2), 142.

Blood, E. A., & Cheng, D. M. (2011). The use of mixed models for the analysis of mediated data with time-dependent predictors. *Journal of Environmental and Public Health*, *2011*, 1–12.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Cao, W., Li, Y., & Yu, Q. (2021). Sensitivity analysis for assumptions of general mediation analysis. *Communications in Statistics-Simulation and Computation*, 1–18. https://www.tandfonline.com/doi/abs/10.1080/03610918.2021.1908556?journalCode=lssp20

Celeux, G., Forbes, F., Robert, C. P., & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, *1*(4), 651–673.

Chan, J. C. C., & Grant, A. L. (2016a). Fast computation of the deviance information criterion for latent variable models. *Computational Statistics & Data Analysis*, *100*, 847–859.

Chan, J. C. C., & Grant, A. L. (2016b). On the observed-data deviance information criterion for volatility modeling. *Journal of Financial Econometrics*, *14*(4), 772–802.

Dey, S., Delampady, M., & Gopalaswamy, A. M. (2019). Bayesian model selection for spatial capture-recapture models. *Ecology and Evolution*, *9*(20), 11569–11583. https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.5551

Ditlevsen, S., Christensen, U., Lynch, J., Damsgaard, M. T., & Keiding, N. (2005). The mediation proportion: A structural equation approach for estimating the proportion of exposure effect on outcome explained by an intermediate variable. *Epidemiology*, *16*(1), 114–120.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.

Gao, T., & Albert, J. M. (2018). Bayesian causal mediation analysis with multiple ordered mediators. *Statistical Modelling*, *19*(6), 1471082–18798067.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*(365), 153–160.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Publications.

Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological Science*, *24*(10), 1918–1927.

Hoogendijk, E. O., Deeg, D. J., Poppelaars, J., van der Horst, M., van Groenou, M. I. B., Comijs, H. C., … van Tilburg, T. G. (2016). The Longitudinal aging study Amsterdam: Cohort update 2016 and major findings. *European Journal of Epidemiology*, *31*(9), 927–945.

Huisman, M., Poppelaars, J., Horst, M., Beekman, A., Brug, J., Tilburg, T. G., & Deeg, D. J. (2011). Cohort profile: The longitudinal aging study Amsterdam. *International Journal of Epidemiology*, *40*, 868–876.

Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, *8*(2), 115.

Krull, J. L., & MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review*, *23*(4), 418–444.

Krull, J. L., & MacKinnon, D. P. (2001). Multivarite modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, *36*, 249–277.

Lee, S. Y. (2007). *Structural equation modeling: A Bayesian approach* (Vol. *711*). Hoboken, NJ: John Wiley & Sons.

Li, L., Qiu, S., Zhang, B., & 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.

Li, Y., & Yu, J. (2012). Bayesian hypothesis testing in latent variable models. *Journal of Econometrics*, *166*(2), 237–246.

MacKinnon, D. (2008). *Introduction to statistical mediation analysis*. London, UK: Routledge.

MacKinnon, D. P., Fritz, M. S., Williams, J., & Lockwood, C. M. (2007). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behavior Research Methods*, *39*(3), 384–389.

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*(1), 99–128.

McNeish, D. (2017). Multilevel mediation with small samples: A cautionary note on the multilevel structural equation modeling framework. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 609–625.

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295–314.

Merkle, E, Furr, D, & Rabe-Hesketh, S. (2018). Bayesian model assessment: Use of conditional vs marginal likelihoods. *arXiv preprint arXiv:180204452*.

Millar, R. B. (2009). Comparison of hierarchical Bayesian models for overdispersed count data using DIC and Bayes' factors. *Biometrics*, *65*(3), 962–969.

Millar, R. B. (2018). Conditional vs marginal estimation of the predictive loss of hierarchical models using WAIC and cross-validation. *Statistics and Computing*, *28*(2), 375–385.

Plummer M. (2013). JAGS user manual, 3.4.0 ed.

Plummer M. (2016). rjags: Bayesian graphical models using MCMC. *R Package Version* 4(6).

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*(2), 77–98.

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling*, *18*(2), 161–182.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*(3), 209.

Quintero, A., & Lesaffre, E. (2018). Comparing hierarchical models via the marginalized deviance information criterion. *Statistics in Medicine*, *37*(16), 2440–2454.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. *1*). USA: Sage.

Robitaille, A., Piccinin, A. M., Muniz-Terrera, G., Hoffman, L., Johansson, B., Deeg, D. J., … Hofer, S. M. (2013). Longitudinal mediation of processing speed on age-related change in memory and fluid intelligence. *Psychology and Aging*, *28*(4), 887.

Rusá, Š., Komárek, A., Lesaffre, E., & Bruyneel, L. (2018). Multilevel moderated mediation model with ordinal outcome. *Statistics in Medicine*, *37*(10), 1650–1670.

Song, H. (2018). A primer on multilevel mediation models for egocentric social network data. *Communication Methods and Measures*, *12*(1), 1–24.

Spiegelhalter, D. J., Best, N., Carlin, N., & van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society - Series B*, *76*(3), 485–493.

Tom, A., Bosker, T. A. S. R. J., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. USA: Sage.

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, *92*(2), 351–370.

Wang, J., Boyer, J., & Genton, M. M. G. (2004). A note on an equivalence between chi-square and generalized skew-normal distributions. *Statistics & Probability Letters*, *66*(4), 395–398.

Wang, Y. B., Chen, Z., Goldstein, J. M., Buck Louis, G. M., & Gilman, S. E. (2019). A Bayesian regularized mediation analysis with multiple exposures. *Statistics in Medicine*, *38*(5), 828–843.

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.

Wu, W., Carroll, I. A., & Chen, P. Y. (2018). A single-level random-effects cross-lagged panel model for longitudinal mediation analysis. *Behaviour Research Methods*, *50*(5), 2111–2124.

Yanuar, F., Ibrahim, K., & Jemain, A. A. (2013). Bayesian structural equation modeling for the health index. *Journal of Applied Statistics*, *40*(6), 1254–1269.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*(4), 301.

Zhang, Z., Zyphur, M. J., & Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods*, *12*(4), 695–719.

Zigler, C. K., & Ye, F. (2019). A comparison of multilevel mediation modeling methods: Recommendations for applied researchers. *Multivariate Behavioral Research*, *54*(3), 338–359.

How to cite this article: Ariyo, O., Lesaffre, E., Verbeke, G., Huisman, M., Heymans, M., & Twisk, J. (2022). Bayesian model selection for multilevel mediation models. *Statistica Neerlandica*, *76*(2), 219–235. https://doi.org/10.1111/stan.12256