Running head: COMPARING DIC AND WAIC FOR MULTILEVEL MODELS WITH MISSING DATA

Comparing DIC and WAIC for Multilevel Models with Missing Data

Han Du, Brian Keller, Egamaria Alacam, Craig Enders

Department of Psychology, University of California, Los Angeles

Educational, School, & Counseling Psychology, University of Missouri

Department of Psychology, University of California, Los Angeles

Department of Psychology, University of California, Los Angeles

Correspondence should be addressed to Han Du, Pritzker Hall, 502 Portola Plaza, Los Angeles, CA 90095.

Email: hdu@psych.ucla.edu.

This work was supported by Institute of Educational Sciences award R305D1900002.

Du, H., Keller, B. T., Alacam, E., & Enders, C. K. (2023). Comparing DIC and WAIC for multilevel models with missing data. Behavior Research Methods. Advance online publication.

Abstract

In Bayesian statistics, the most widely used criteria of Bayesian model assessment and comparison are Deviance Information Criterion (DIC) and Watanabe–Akaike Information Criterion ($WAIC$). We use a multilevel mediation model as an illustrative example to compare different types of DIC and $WAIC$. More specifically, we aim to compare the performance of conditional and marginal DICs and $WAIC$s, and investigate their performance with missing data. We focus on two versions of DIC (DIC_1 and DIC_2) and one version of $WAIC$. In addition, we explore whether it is necessary to include the nuisance models of incomplete exogenous variables in likelihood. Based on the simulation results, whether DIC_2 is better than DIC_1 and $WAIC$ and whether we should include the nuisance models of exogenous variables in likelihood functions depend on whether we use marginal or conditional likelihoods. Overall, we find that the marginal likelihood based DIC_2 that excludes the likelihood of covariate models generally had the highest true model selection rates.

Comparing DIC and WAIC for Multilevel Models with Missing Data

In the recent decades, Bayesian statistics have been more widely used given the development of Markov chain Monte Carlo ($MCMC$) sampling techniques and computational power. As in frequentist statistics, researchers need to compare candidate models and evaluate which model fits the data better. The most widely used criteria of Bayesian model assessment and comparison are Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) and Watanabe–Akaike Information Criterion ($WAIC$; Watanabe & Opper, 2010). Model selection usually involves models with latent variables and random effects. More specifically, in item response theory (IRT) models and structural equation models (SEM), latent variables represent latent abilities or traits. In multilevel models (MLM), random effects represent group differences (e.g., families, schools, and countries) when participants are grouped, or represent individual differences in longitudinal data.

DIC and $WAIC$ evaluate model fit using likelihood. For models with latent variables and random effects, there are two options of likelihood calculation: marginal likelihood and conditional likelihood. Marginal likelihood does not rely on random effects and latent variables, but considers the variances of the random effects and latent variables. Conditional likelihood is conditional on random effects and latent variables without directly using their variances. Comparing marginal and conditional likelihood information criteria is widely discussed in DIC, but only a few researchers pay attention to the marginal likelihood version of $WAIC$ (e.g., Li et al., 2016; Merkle et al., 2019; Millar, 2018; Vehtari et al., 2016). There are two perspectives in choosing between marginal and conditional likelihoods. First, theoretically and conceptually, Merkle et al. (2019) suggest that if we want to make predictions for new observations within the same groups as in the original data, the conditional likelihood is more appropriate; whereas if predictions are for new observations in groups outside of the original data, the marginal likelihood is more appropriate. The latter application is usually what we anticipate because we can generalize the model to new clusters (e.g., schools and countries) that are not in the original data. Spiegelhalter et al. (2002)

suggest that if the focal parameters of interest include latent variables and random effects, we need to use the conditional likelihood; otherwise, we can use the marginal likelihood. Second, recent studies (e.g., Merkle et al., 2019; Tong et al., 2022; Zhang et al., 2019) have found that marginal likelihood information criteria have smaller Monte Carlo errors and better model selection than conditional likelihood information criteria in various models (e.g., multi-group confirmatory factor analysis, mixture growth curve model, multilevel item response theory model). When comparing DIC and $WAIC$, the conclusion also depends on marginal and conditional likelihood information. DIC and $WAIC$ performed more similarly with marginal likelihoods compared to conditional likelihoods (Tong et al., 2022).

Although there is a big distinction between marginal and conditional likelihood information criteria, substantive researchers do not necessarily notice that there are two versions of information criteria because they usually rely on the default settings of software. Merkle et al. (2019) summarized that Stan (Carpenter et al., 2017), BUGS (Spiegelhalter et al., 1996), and jags (Plummer et al., 2003) report conditional likelihood information criteria, whereas blavaan (Merkle & Rosseel, 2015) reports marginal likelihood information criteria. Mplus (Muthén & Muthén, 1998–2017) reports conditional likelihood or marginal likelihood information criteria depending on the specified model (if integrating the random effects is easy, marginal likelihood information criteria are reported).

Besides marginal likelihood versus conditional likelihood, missing data are another important issue in DIC and $WAIC$ computation. Bayesian information criteria with missing data has not received widespread attention. Especially, when predictors/covariates are incomplete, we need to estimate the missing predictors/covariates. As exogenous variables, we cannot directly estimate predictors but need to specify models for predictors. Although SEM software such as Mplus and lavaan allows users to treat predictors/covariates as random and a multivariate normal distribution $f(\boldsymbol{X}, Y)$ is specified for all exogenous and endogenous variables, this approach is not suggested when the model contains nonlinear covariate effects, such as quartic terms and interaction effects, random slopes, or categorical variables (e.g.,

Bartlett et al., 2015; Enders et al., 2018; Erler et al., 2016; Grund et al., 2018; Kim et al., 2015; Seaman et al., 2012; Van Buuren et al., 2006). In these cases, the joint distribution of all exogenous and endogenous variable is not a multivariate normal distribution (Du et al., 2022; Enders et al., 2020, 2018). Due to the misspecification, estimation and inference will be misleading.[1]

To solve the misspecification issue, researchers have proposed various substantive-model-compatible covariate models $f(\boldsymbol{X})$ that specify the relationship between missing predictors/covariates and help impute their missing values. The first option is the substantive-model-compatible joint modeling that specifies a joint model (usually a multivariate normal distribution) for $f(\boldsymbol{X})$ and the joint distribution of all variables is given by $f(Y|\boldsymbol{X})\, f(\boldsymbol{X})$ where $f(Y|\boldsymbol{X})$ is the model implied distribution of the substantive analysis model. However, when the relation between the exogenous variables are nonlinear or with random slopes, $f(\boldsymbol{X})$ is not normal or any familiar distribution. The second option is the substantive-model-compatible fully conditional specification (Grund et al., 2021) and is referred by us as the separate specification to emphasize that the covariate model $f(X_K|X_{-k})$ (the kth predictor regressed on all other predictors) only computes the (marginal or conditional) likelihood of each predictor without considering the joint distribution. When $f(X_K|X_{-k})$ is specified to be normal, the assumption is the same as the substantive-model-compatible joint modeling and as a consequence $f(\boldsymbol{X})$ is misspecified when the relation between the exogenous variables are nonlinear or with random slopes. The third option is the sequential specification that computes the joint (marginal or conditional) likelihood of all variables by factoring the joint distribution of predictors into a sequence of univariate regression models (i.e., $f(\boldsymbol{X}) = f(X_K)\, f(X_{K-1}|X_K)\, f(X_{K-2}|X_K, X_{K-1}) \ldots f(X_1|X_{>1})$; Ibrahim et al., 1999). Then the joint distribution may or may not be normal. Different from the joint modeling approach, the sequential approach can ensure the existence of a distribution (i.e., compatibility; Bartlett et al., 2015; Du et al., 2022; Enders et al., 2018; Erler et al., 2016). Therefore, the sequential specification can accommodate nonlinear terms or random slopes between the exogenous variables.

Regardless of the specific covariate models employed, it is important to recognize that these models are nuisance models. The nuisance models lack substantive interest, but they incorporate likelihood functions and may contribute to evaluating the fit of the overall model. Although these nuisance models help impute missing data, it is unclear whether they play an important role in assessing model fit. To the best of our knowledge, whether to include nuisance models in information criteria computation has not been studied yet. In comparing the aforementioned two competing frameworks, the sequential factorization naturally incorporates all models in fit summaries via $f(\boldsymbol{X})$. On the other hand, the separate specification uses $f(X_K|X_{-k})$ and does not directly offer $f(\boldsymbol{X})$, and thus excludes the possibility of including nuisance models in the likelihoods. In parameter estimation and inference, when the compatibility assumptions are satisfied (Du et al., 2022; Grund et al., 2021), both methods are similar. However, they can differ significantly in terms of information criteria. Surprisingly, this issue has not yet been explored in the existing literature. Thus, the primary objective of our paper is to provide preliminary insights into this problem. We will elaborate on the compatibility issue later and explore whether it is critical to include the nuisance models in information criteria.

This paper presents a multilevel mediation model as an illustrative example to address our research questions. First, the multilevel mediation model as a multilevel model allows us to compare conditional likelihood information criteria (conditional on random effects) and marginal likelihood criteria (conditional on the level-2 variances). Second, this model encompasses essential elements that lie at the core of the marginal vs. conditional likelihood debate. It incorporates random intercepts, random slopes, latent cluster means, and missing data. Therefore, this model serves as a foundation for comprehending various multilevel path models of interest. This model is more complex than a large proportion of multilevel models with univariate outcomes in the literature. The insights gained from this fundamental model naturally generalize to more complex models, which may exhibit similar issues and features explored in this study. Third, we are careful in selecting the magnitudes of the effects based on real data. We aim to

design scenarios that could generate realistic misspecifications. These misspecifications encompass both overfitting and underfitting scenarios. Given the meticulous determination of effects, a smaller model was necessary.

Overall, in this paper, we aim to compare the performance of conditional and marginal DICs and $WAIC$s in multilevel modeling. More specifically, we are interested in their performance with missing data. To the best of our knowledge, $WAIC$ with missing data has not been studied yet. In addition, we want to explore whether it is necessary to include the nuisance models of incomplete exogenous variables (i.e., the aforementioned covariate models) in likelihood. Various versions of DIC and $WAIC$ have been proposed by researchers, differing in their approaches to handling missing data and marginal versus conditional likelihood (e.g., Celeux et al., 2006; Lu & Zhang, 2022). This has resulted in challenges when using these criteria. Our objective is to identify the best-performing option and provide guidance to researchers utilizing these criteria.

The outline of this paper is as follows. In the "Multilevel Mediation Model" section, we introduce a multilevel mediation model. In the "Conditional Versus Marginal Likelihoods" section, we compare conditional and marginal likelihoods and illustrate equations for the multilevel mediation model. In the "Likelihoods with Missing Data" section, we compare the separate and sequential specifications and illustrate how the specification influences likelihoods. In the "DIC and $WAIC$ Calculation" section, we propose a new way to compute DIC, and present DIC and $WAIC$ equations with conditional and marginal likelihoods. In the "Simulation Study" section, we examine the performance of conditional and marginal DICs and $WAIC$s in model selection. In the "Real Data Example" section, we use a real data example to compare conditional and marginal DICs and $WAIC$s. We end the paper with some concluding remarks in the "Discussion" section.

Illustrative Multilevel Model

We consider a two-level mediation model with one level-1 predictor (x_{1ij}), one level-2 predictor (x_{2j}), one level-1 mediator (m_{ij}), and one outcome (y_{ij}) where i indicates the ith individual and j indicates the jth cluster ($j = 1, ..., J$). The mediation model consists of three sub-models: y_{ij}'s regression model, m_{ij}'s regression model, and the regression model between two predictors (x_{1ij} and x_{2j}). We will estimate all sub-models simultaneously as one big model using Bayesian statistics. y_{ij}'s regression model is in Equation (1) where predictors do not have a direct effect on y_{ij}.

$$
\begin{aligned}
y_{ij} &= \beta_{y,0j} + \beta_{y,1j} m_{ij} + e_{y,ij}, \qquad e_{y,ij} \sim N\left(0, \sigma^2_{y,e}\right) \\
\beta_{y,0j} &= \beta_{y,00} + u_{y,0j}, \\
\beta_{y,1j} &= \beta_{y,10} + u_{y,1j}, \qquad \begin{pmatrix} u_{y,0j} \\ u_{y,1j} \end{pmatrix} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma_{y,u}}\right)
\end{aligned} \tag{1}
$$

$\beta_{y,0j}$ and $\beta_{y,1j}$ indicate the cluster specific random intercept and slope in y_{ij}'s regression model, respectively. $e_{y,ij}$ is the level-1 residual variance with a variance of $\sigma^2_{y,e}$. $\beta_{y,00}$ and $\beta_{y,10}$ are the overall intercept and slope averaged over clusters, respectively. $u_{y,0j}$ and $u_{y,0j}$ are the level-2 residuals that capture the intercept and slope differences between clusters with a covariance matrix of $\boldsymbol{\Sigma_{y,u}}$.

m_{ij}'s regression model is in Equation (2) with a random intercept where x_{2j} predicts the random intercept.

$$
\begin{aligned}
m_{ij} &= \beta_{m,0j} + \beta_{m,1} x_{1ij} + e_{m,ij}, \quad e_{m,ij} \sim N\left(0, \sigma^2_{m,e}\right) \\
\beta_{m,0j} &= \beta_{m,00} + \beta_{m,01} x_{2j} + u_{m,0j}, \quad u_{m,0j} \sim N\left(0, \sigma^2_{m,u}\right)
\end{aligned} \tag{2}
$$

$\beta_{m,0j}$ indicates the cluster specific intercept, whereas $\beta_{m,1}$ indicates that the effect of x_{1ij} on m_{ij} is fixed across clusters. $e_{m,ij}$ indicates the level-1 residual variance with a variance of $\sigma^2_{m,e}$. $\beta_{m,00}$ and $\beta_{m,01}$ are

the overall intercept and the effect of level-2 predictor x_{2j} on the random intercept, respectively. $u_{m,0j}$ is the level-2 residual on random intercepts with a variance of $\sigma^2_{m,u}$.

Nuisance Models of Incomplete Exogenous Variables

Besides the focal model in Equations (1) and (2), when x_{1ij} and x_{2j} are incomplete, we need to estimate the missing values using nuisance models for x_{1ij} and x_{2j}. As previously mentioned, there are two ways to specify the covariate models (we do not focus on the joint modeling approach in this paper): sequential specification and separate specification. They have been systematically summarized and compared in Du et al. (2022), Grund et al. (2021), and Lüdtke et al. (2020). The sequential specification can ensure *compatibility*, meaning that the joint distribution of all variables exists, whereas the separate specification has the risk of failing to ensure the existence of the joint distribution of all variables (Du et al., 2022). More specifically, when all models are linear (i.e., without any polynomial or interactive terms) with normally distributed errors, the implied joint distribution of variables is multivariate normal. However, models are not necessarily linear in practice. When models contain nonlinear covariate effects, such as quartic terms and interaction effects, or random slopes (e.g., Bartlett et al., 2015; Enders et al., 2018; Erler et al., 2016; Grund et al., 2018; Kim et al., 2015; Seaman et al., 2012; Van Buuren et al., 2006), the separate specification cannot guarantee the existence of the joint distribution of all variables and thus parameter estimation and inference are wrong. The sequential specification is more flexible in dealing with the compatibility issue.

The sequential specification factors the joint distribution of predictors into a sequence of univariate distributions (Ibrahim et al., 1999). For the sequential specification, the covariate model can be specified as

$f(x_1, x_2) = f(x_1|x_2) f(x_2)$ with x_{1ij}'s and x_{2j}'s models defined as followed.

$$
\begin{aligned}
x_{1ij} &= \beta_{x1,0j} + e_{x1,ij}, \quad e_{x1,ij} \sim N\left(0, \sigma^2_{x1,e}\right) \\
\beta_{x1,0j} &= \beta_{x1,00} + \beta_{x1,01} x_{2j} + u_{x1,0j}, \quad u_{x1,0j} \sim N\left(0, \sigma^2_{x1,u}\right) \\
x_{2j} &= \beta_{x2,0} + u_{x2,0j}, \quad u_{x2,0j} \sim N\left(0, \sigma^2_{x2,u}\right)
\end{aligned} \tag{3}
$$

In x_{1ij}'s model (i.e., $f(x_1|x_2)$), $\beta_{x1,0j}$ indicates the cluster mean of x_{1ij}. $e_{x1,ij}$ indicates the level-1 residual variance of x_{1ij} with a variance of $\sigma^2_{m,e}$. $\beta_{x1,00}$ and $\beta_{x1,01}$ are the intercept and the effect of level-2 predictor x_{2j} on the cluster mean, respectively. $u_{x1,0j}$ is the level-2 residual on random intercepts with a variance of $\sigma^2_{x1,u}$. In x_{2j}'s model (i.e., $f(x_2)$), $\beta_{x2,0}$ and $\sigma^2_{x2,u}$ indicate the mean and variance of x_{2j}, respectively. When we further consider y_{ij}'s and m_{ij}'s models, we can compute the joint distribution of y, m, x_1, and x_2 by multiplying Equations (1), (2), and (3),
$f(y, m, x_1, x_2) = f(y|m) f(m|x_1, x_2) f(x_1|x_2) f(x_2)$.

On the other hand, the separate specification specify the univariate conditional distribution as regressing each predictor on all other predictors (Bartlett et al., 2015; Enders et al., 2020). In this multilevel mediation model, we will need $f(x_1|x_2)$ and $f(x_2|x_1)$. Since x_1 and x_2 are at different levels, x_2 is conditional on the cluster mean of x_1. The covariate model is as follows.

$$
\begin{aligned}
x_{1ij} &= \beta_{x1,0j} + e_{x1,ij}, \quad e_{x1,ij} \sim N\left(0, \sigma^2_{x1,e}\right) \\
\beta_{x1,0j} &= \beta_{x1,00} + \beta_{x1,01} x_{2j} + u_{x1,0j}, \quad u_{x1,0j} \sim N\left(0, \sigma^2_{x1,u}\right) \\
x_{2j} &= \beta_{x2,0} + \beta_{x2,1}\beta_{x1,0j} + u_{x2,0j}, \quad u_{x2,0j} \sim N\left(0, \sigma^2_{x2,u}\right)
\end{aligned} \tag{4}
$$

x_{1ij}'s model keeps the same as the one in Equation (3), but x_{2j}'s model is $f(x_2|x_1)$ instead of $f(x_2)$. $\beta_{x2,0}$ and $\beta_{x2,1}$ are the intercept and the effect of level-2 predictor's cluster mean $\beta_{x1,0j}$, respectively.

u_{x120j} is the residual with a variance of $\sigma^2_{x2,u}$. We cannot directly multiply $f\left(x_1|x_2\right)$ and $f\left(x_2|x_1\right)$ to compute the joint distribution of x_1 and x_2, $f\left(x_1, x_2\right)$, although without nonlinear terms and random slopes we can derive $f\left(x_1, x_2\right)$ based on $f\left(x_1|x_2\right)$ and $f\left(x_2|x_1\right)$. Hence, the separate specification cannot directly provide a joint likelihood of y, m, x_1, and x_2. Instead, by multiplying Equations (1) and (2) we can compute the joint distribution of y and m, $f\left(y, m|x_1, x_2\right) = f\left(y|m\right) f\left(m|x_1, x_2\right)$.

Since this multilevel mediation model contains random slopes, the joint distribution of y, m, x_1, and x_2 is not a multivariate normal distribution (Du et al., 2022; Enders et al., 2020, 2018). The joint modeling approach of directly specifying a normal distribution of all variables will cause noticeable biases of estimation.

Conditional Versus Marginal Likelihoods

As previously mentioned, conditional likelihoods and marginal likelihoods are important in order to calculate information criterion. For each model in the multilevel mediation analysis, we can compute a conditional likelihood and a marginal likelihood. Suppose there are J clusters. The conditional likelihood of y_{ij}'s model is conditional on the random effects ($\boldsymbol{u_{y,j}} = (u_{y,0j}, u_{y,1j})$; see Equation 5). Let $\boldsymbol{\beta_y}$ denote $(\beta_{y,00}, \beta_{y,10})$, $\boldsymbol{y_j}$ denote the outcome scores in the jth cluster, and $\boldsymbol{m_j}$ denote the mediator scores in the jth cluster.

$$f\left(\boldsymbol{y}|\sigma^2_{y,e}, \boldsymbol{\beta_y}, \boldsymbol{m}, \boldsymbol{u_y}\right) = \prod_{j=1}^{J} f\left(\boldsymbol{y_j}|\sigma^2_{y,e}, \boldsymbol{\beta_y}, \boldsymbol{m_j}, \boldsymbol{u_{y,j}}\right) \tag{5}$$

The marginal likelihood of y_{ij}'s model integrates out the random effects and is conditional on the level-2 variances ($\boldsymbol{\Sigma_{y,u}}$; see Equation 6).

$$f\left(\boldsymbol{y}|\sigma^2_{y,e}, \boldsymbol{\beta_y}, \boldsymbol{m}, \boldsymbol{\Sigma_{y,u}}\right) = \prod_{j=1}^{J} \int f\left(\boldsymbol{y_j}|\sigma^2_{y,e}, \boldsymbol{\beta_y}, \boldsymbol{m_j}, \boldsymbol{u_{y,j}}\right) f\left(\boldsymbol{u_{y,j}}|\boldsymbol{\Sigma_{y,u}}\right) d\boldsymbol{u_{y,j}} \tag{6}$$

The conditional likelihood of m_{ij}'s model is conditional on the random effect ($u_{m,0j}$; see Equation 7). Let $\boldsymbol{\beta_m}$ denote $(\beta_{m,00}, \beta_{m,01}, \beta_{m,1})$, $\boldsymbol{x_{1j}}$ denote the level-1 predictor scores in the jth cluster, and x_{2j} denote the level-2 predictor score in the jth cluster.

$$f\left(\boldsymbol{m}|\sigma^2_{m,e}, \boldsymbol{\beta_m}, \boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{u_{m,0}}\right) = \prod_{j=1}^{J} f\left(\boldsymbol{m_j}|\sigma^2_{m,e}, \boldsymbol{\beta_m}, \boldsymbol{x_{1j}}, x_{2j}, u_{m,0j}\right) \tag{7}$$

The marginal likelihood of the mediator model integrates out $\beta_{m,0j}$ and is conditional on the level-2 variance ($\sigma^2_{m,u}$; see Equation 8).

$$\begin{aligned} f\left(\boldsymbol{m}|\sigma^2_{m,e}, \boldsymbol{\beta_m}, \boldsymbol{x_1}, \boldsymbol{x_2}, \sigma^2_{m,u}\right) = \prod_{j=1}^{J} \int & f\left(\boldsymbol{m_j}|\sigma^2_{m,e}, \boldsymbol{\beta_m}, \boldsymbol{x_{1j}}, x_{2j}, u_{m,0j}\right) \\ & \times f\left(u_{m,0j}|\sigma^2_{m,u}\right) du_{m,0j} \end{aligned} \tag{8}$$

As mentioned in the previous section, we only can compute the joint distribution of x_{1ij} and x_{2j} with the sequential specification but not the separate specification. Based on the sequential specification, the conditional likelihood of x_{1ij} and x_{2j}'s regression model is conditional on the random effect ($u_{x1,0j}$; see Equation 9). Let $\boldsymbol{\beta_{x1}}$ denote $(\beta_{x1,00}, \beta_{x1,01})$, $\boldsymbol{x_{1j}}$ denote the level-1 predictor scores in the jth cluster, and x_{2j} denote the level-2 predictor score in the jth cluster.

$$f\left(\boldsymbol{x_1}, \boldsymbol{x_2}|\sigma^2_{x1,e}, \boldsymbol{\beta_{x1}}, \beta_{x2,0}, u_{x1,0j}, U_{x2}\right) = \prod_{j=1}^{J} f\left(\boldsymbol{x_{1j}}|\sigma^2_{x1,e}, \boldsymbol{\beta_{x1}}, u_{x1,0j}\right) f\left(x_{2j}|\beta_{x2,0}, U_{x2}\right) \tag{9}$$

The marginal likelihood of x_{1ij} and x_{2j}'s regression model integrates out $u_{x1,0j}$ and is conditional on the level-2 variance ($\sigma^2_{x1,u}$; see Equation 10).

$$f\left(\boldsymbol{x_1}, \boldsymbol{x_2}|\sigma^2_{x1,e}, \boldsymbol{\beta_{x1}}, \beta_{x2,0}, U_{x1}, U_{x2}\right) = \prod_{j=1}^{J} \int f\left(\boldsymbol{x_{1j}}|\sigma^2_{x1,e}, \boldsymbol{\beta_{x1}}, u_{x1,0j}\right) f\left(x_{2j}|\beta_{x2,0}, U_{x2}\right) \times f\left(u_{x1,0j}|\sigma^2_{x1,u}\right) du_{x1,0j} \quad (10)$$

Likelihoods with Missing Data

Celeux et al. (2006) proposed three ways to incorporate missing data in likelihoods, which lead to different information criteria. First, we can ignore missing variables and calculate the likelihood only based on the fully observed variables that contain no missing data ($f\left(observed.var|\theta\right)$) where $observed.var$ stands for all fully observed variables. This likelihood is called observed likelihood. Second, we can calculate the joint likelihood of missing variables and fully observed variables given parameters ($f\left(observed.var, missing.var|\theta\right)$) where $missing.var$ stands for all variables that have missing data. This likelihood is called complete likelihood. Third, we can calculate the likelihood of the fully observed variables given the missing variables and parameters ($f\left(observed.var|\theta, missing.var\right)$). This likelihood is called conditional likelihood. The complete and conditional likelihoods require either estimating or integrating out missing values of incomplete variables. Celeux et al. (2006) further proposed various DICs within each type of likelihood. Celeux et al. (2006) considered a random effect model where the random effects can be viewed as a variable that is completely missing. In our model, besides random effects and latent variables, we have predictors, mediator, and outcome which are partially missing. In this paper, we impute missing data using MCMC estimation and propose to define likelihoods based on the analysis model. For example, in terms of the conditional likelihood of y_{ij}'s model $f\left(\boldsymbol{y}|\sigma^2_{y,e}, \boldsymbol{\beta_y}, \boldsymbol{m}, \boldsymbol{u_y}\right)$, if both m and y have missing observations, we will define the likelihood as $f\left(\boldsymbol{y^o}, \boldsymbol{y^m}|\sigma^2_{y,e}, \boldsymbol{\beta_y}, \boldsymbol{m^o}, \boldsymbol{m^m}\right)$ where superscript o denotes observed data and superscript m denotes missing data. Hence, our likelihood is a

combination of Celeux's complete and conditional likelihoods.

To impute missing data, we use a specific type of fully Bayesian imputation called the model-based imputation that can guarantee that the conditional distribution of the incomplete variables is mathematically correct and compatible with each other (e.g., Bartlett et al., 2015; Enders et al., 2020; Erler et al., 2019; Goldstein et al., 2014; Kim et al., 2018). The rationale of the model-based imputation method is that we use the substantive model and the so-called covariate models that capture the relationship among the predictors to construct the imputation model for incomplete predictors. In this multilevel mediation model, the substantive analysis models are y_{ij}'s regression model (Eq. 1) and m_{ij}'s regression model (Eq. 2) since these two models address the substantive mediation research question.

In terms of the covariate model, there are two options: sequential specification and separate specification. Only the sequential specification can ensure compatibility between all the covariate models whereas the separate specification may cause an incompatibility issue (Du et al., 2022). Without random slopes, nonlinear terms, and nonnormal data, the sequential and separate specifications are equivalent and provide very similar parameter estimates. Hence, in this multilevel mediation example, the sequential and separate specifications both ensure compatibility and are safe to use. Additionally, in our pilot simulation, we find that the sequential and separate specifications provide almost the same parameter estimates in y_{ij}'s and m_{ij}'s regression models. In the current multilevel mediation example, the difference between the sequential and separate specifications is that the separate specification only can provide $f\left(y, m|x_1, x_2\right)$ but not $f\left(y, m, x_1, x_2\right)$ since we cannot directly obtain $f\left(x_1, x_2\right)$ with the separate specification. We compare the sequential and separate specifications in DIC and $WAIC$ calculation to explore whether it is necessary to include the predictor or covariate models (e.g., $f\left(x_1, x_2\right)$) in information criteria calculation. For more background on the fully Bayesian imputation specifically using Blimp (Keller & Enders, 2021), we refer readers to Enders (2022).

DIC and WAIC Calculation

We focus on two widely used information criteria in this paper: DIC and $WAIC$. In the fully Bayesian approach, likelihoods (defined in the previous sections) and prior distributions lead to posterior distributions via the Bayes' rule. The Bayesian estimation views parameters as random variables and uses $MCMC$ sampling procedures to draw parameters from a series of posterior distributions. Therefore, Bayesian estimation is not a single value as in maximum likelihood estimation, but a set of posterior distribution samples. When comparing DIC and $WAIC$, recent research regards $WAIC$ as an improvement over DIC since it has multiple desirable properties and theoretical support. For example, $WAIC$ averages over the posterior distribution rather than only uses a point estimate, and DIC can have negative estimates of the effective number of parameters (Gelman, Hwang, & Vehtari, 2014; Vehtari et al., 2017). Hence, we propose DIC_2 that shares the similar property as $WAIC$. In addition, the original $WAIC$ does not consider multilevel structures or missing data. Given the preference for the $WAIC$ in the literature, investigating the performance of the $WAIC$ with random effects and missing data is an important avenue for future research.

DIC is reported in all Bayesian software packages. WAIC is available in Stan (Carpenter et al., 2017), BUGS (Spiegelhalter et al., 1996), jags (Plummer et al., 2003), and the R packages blavaan (Merkle & Rosseel, 2015) and loo (Vehtari et al., 2023). While $WAIC$ and DIC are primarily designed to optimize prediction rather than selecting the true data generating model, they are still frequently employed to identify the best fitting model from among several competitive candidates (Gronau & Wagenmakers, 2019; Spiegelhalter et al., 2014).

DIC

Spiegelhalter et al. (2002) proposed the Deviance Information Criterion (DIC). DIC is based on a deviance $D(\theta) = -2logf(y|\theta)$ where y is the data and θ is the population parameter (suppose we only

have one parameter). $f\left(y|\theta\right)$ can be either the marginal or conditional likelihood. Since the population parameter is unknown, we can replace θ with posterior mean, mode, or median to calculate $D\left(\hat{\theta}\right) = -2logf\left(y|\hat{\theta}\right)$. Posterior mean is the most widely used in software and real data analysis and $D\left(\hat{\theta}\right)$ is called the posterior mean deviance. Besides focusing on point estimates, we also can average over the parameter space to calculate the likelihood, $\overline{D\left(\theta\right)} = E_\theta\left[-2logf\left(y|\theta\right)\right]$. In practice, it can be calculated as $\overline{D\left(\theta\right)} = \frac{1}{K}\sum_{k=1}^{K} -2logf\left(y|\hat{\theta}_k\right)$ where K indicates the K iterations and $\hat{\theta}_k$ is the kth posterior sample. The difference between $D\left(\hat{\theta}\right)$ and $\overline{D\left(\theta\right)}$ measures the effective number of parameters in the model, $p_D = \overline{D\left(\theta\right)} - D\left(\hat{\theta}\right)$. The general equation of DIC is

$$\begin{aligned} DIC &= \overline{D\left(\theta\right)} + p_D \\ &= 2\overline{D\left(\theta\right)} - D\left(\hat{\theta}\right) \end{aligned} \tag{11}$$

We denote the widely used DIC with $D\left(\hat{\theta}\right) = -2logf\left(y|\hat{\theta}\right)$ as DIC_1.

$$DIC_1 = -4\frac{1}{K}\sum_{k=1}^{K} logf\left(y|\hat{\theta}_k\right) + 2logf\left(y|\hat{\theta}\right) \tag{12}$$

We use DIC_1 including the nuisance models of exogenous variables ($f\left(\boldsymbol{x_{1j}}, x_{2j}|\hat{\boldsymbol{\theta}}\right)$) as an example in this multilevel mediation model. More specifically, we can calculate DIC_1 in Equation (13) where $\boldsymbol{\theta}$ indicates parameters in marginal likelihoods, or parameters and random effects in conditional likelihoods, $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$ indicates the kth posterior sample of $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}$ indicates the posterior mean in the multilevel mediation

model. We also can extend Equation (13) to DIC_1 excluding the nuisance models.

$$
\begin{aligned}
DIC_1 &= -4\frac{1}{K}\sum_{k=1}^{K}\log f\left(\boldsymbol{y},\boldsymbol{m},\boldsymbol{x_1},\boldsymbol{x_2}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)+2\log f\left(\boldsymbol{y},\boldsymbol{m},\boldsymbol{x_1},\boldsymbol{x_2}|\hat{\boldsymbol{\theta}}\right) \\
&= -4\sum_{j=1}^{J}\left\{\frac{1}{K}\sum_{k=1}^{K}\log f\left(\boldsymbol{y_j},\boldsymbol{m_j},\boldsymbol{x_{1j}},x_{2j}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)\right\}+2\log f\left(\boldsymbol{y},\boldsymbol{m},\boldsymbol{x_1},\boldsymbol{x_2}|\hat{\boldsymbol{\theta}}\right) \\
&= -4\sum_{j=1}^{J}\left\{\frac{1}{K}\sum_{k=1}^{K}\log\left[f\left(\boldsymbol{y_j}|,\boldsymbol{m_j},\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)f\left(\boldsymbol{m_j}|\boldsymbol{x_{1j}},x_{2j},\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)f\left(\boldsymbol{x_{1j}},x_{2j}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)\right]\right\} \\
&\quad +2\sum_{j=1}^{J}\log\left[f\left(\boldsymbol{y_j}|,\boldsymbol{m_j},\hat{\boldsymbol{\theta}}\right)f\left(\boldsymbol{m_j}|\boldsymbol{x_{1j}},x_{2j},\hat{\boldsymbol{\theta}}\right)f\left(\boldsymbol{x_{1j}},x_{2j}|\hat{\boldsymbol{\theta}}\right)\right]
\end{aligned}
\tag{13}
$$

Suppose $\boldsymbol{x_1}$ and $\boldsymbol{m}$ have missing data, we impute missing data at each iteration and calculate DIC_1 based on imputed data, where $\hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\boldsymbol{1,k}}$, $\hat{\boldsymbol{m}}^{\boldsymbol{m}}_{\boldsymbol{k}}$, and $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$ indicate the kth posterior samples of $\boldsymbol{x_1}$, $\boldsymbol{m}$, and $\boldsymbol{\theta}$ respectively in the multilevel mediation model, $\hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\boldsymbol{1j,k}}$ and $\hat{\boldsymbol{m}}^{\boldsymbol{m}}_{\boldsymbol{j,k}}$ indicate the kth posterior samples of $\boldsymbol{x_1}$ and $\boldsymbol{m}$ in the jth cluster, and $\hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\boldsymbol{1}}$, $\hat{\boldsymbol{m}}^{\boldsymbol{m}}$, and $\hat{\boldsymbol{\theta}}$ indicate the posterior means. Equation (14) includes the nuisance models, and we also can extend Equation (14) to DIC_1 excluding the nuisance models.

$$
\begin{aligned}
DIC_1 &= -4\frac{1}{K}\sum_{k=1}^{K}\log f\left(\boldsymbol{y},\boldsymbol{m^o},\hat{\boldsymbol{m}}^{\boldsymbol{m}}_{\boldsymbol{k}},\boldsymbol{x^o_1},\hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\boldsymbol{1,k}},\boldsymbol{x_2}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)+2\log f\left(\boldsymbol{y},\boldsymbol{m^o},\hat{\boldsymbol{m}}^{\boldsymbol{m}},\boldsymbol{x^o_1},\hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\boldsymbol{1}},\boldsymbol{x_2}|\hat{\boldsymbol{\theta}}\right) \\
&= -4\sum_{j=1}^{J}\left\{\frac{1}{K}\sum_{k=1}^{K}\log\left[f\left(\boldsymbol{y_j}|\boldsymbol{m^o_j},\boldsymbol{m^m_{j,k}},\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)f\left(\boldsymbol{m^o_j},\boldsymbol{m^m_{j,k}}|\boldsymbol{x^o_{1j}},\boldsymbol{x^m_{1j,k}},x_{2j},\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)f\left(\boldsymbol{x^o_{1j}},\boldsymbol{x^m_{1j,k}},x_{2j}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)\right]\right\} \\
&\quad +2\sum_{j=1}^{J}\log\left[f\left(\boldsymbol{y_j}|\boldsymbol{m^o_j},\boldsymbol{m^m_j},\hat{\boldsymbol{\theta}}\right)f\left(\boldsymbol{m^o_j},\boldsymbol{m^m_j}|\boldsymbol{x^o_{1j}},\boldsymbol{x^m_{1j}},x_{2j},\hat{\boldsymbol{\theta}}\right)f\left(\boldsymbol{x^o_{1j}},\boldsymbol{x^m_{1j}},x_{2j}|\hat{\boldsymbol{\theta}}\right)\right]
\end{aligned}
\tag{14}
$$

There are other variants of DIC. For example, Celeux et al. (2006) proposed to calculate $D\left(\hat{\theta}\right)$ as $D\left(\hat{\theta}\right)=-2log\frac{1}{K}\sum\limits_{k=1}^{K}f\left(y|\hat{\theta}_k\right)$, which employs all posterior samples instead of relying solely on posterior means. However, $f\left(y|\hat{\theta}_k\right)$ can be small, especially when we multiply the likelihoods across clusters. Taking the logarithm of an averaged small value can lead to inaccurate results. Follow Celeux's idea, in multilevel modeling, we propose to calculate the average across iterations first and compute the

joint likelihood across clusters (denoted as DIC_2). Hence, the primary distinction between DIC_1 and DIC_2 lies in the calculation of $D\left(\hat{\theta}\right)$: DIC_1 computes the likelihood based on posterior means, whereas DIC_2 calculates the average likelihood across iterations. Therefore, DIC_2 shares the same property as $WAIC$ that it averages over the posterior distribution rather than conditioning on a point estimate. Note that our DIC_2 is following Celeux's idea, not the alternative version of DIC in Gelman et al. (2014). Using the likelihood including the nuisance models as an example, without missing data, $D\left(\hat{\theta}\right)$ and DIC_2 can be computed as in Equation (15). We also can extend Equation (15) to DIC_2 excluding the nuisance models ($f\left(\boldsymbol{x_{1j}}, x_{2j}|\hat{\boldsymbol{\theta}}\right)$) .

$$
\begin{aligned}
DIC_2 = -4 \sum_{j=1}^{J} \left\{ \frac{1}{K} \sum_{k=1}^{K} \log \left[f\left(\boldsymbol{y_j}|, \boldsymbol{m_j}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) f\left(\boldsymbol{m_j}|\boldsymbol{x_{1j}}, x_{2j}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) f\left(\boldsymbol{x_{1j}}, x_{2j}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) \right] \right\} \\
+ 2 \sum_{j=1}^{J} \log \left[\frac{1}{K} \sum_{k=1}^{K} f\left(\boldsymbol{y_j}|, \boldsymbol{m_j}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) f\left(\boldsymbol{m_j}|\boldsymbol{x_{1j}}, x_{2j}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) f\left(\boldsymbol{x_{1j}}, x_{2j}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) \right]
\end{aligned} \tag{15}
$$

When $\boldsymbol{x_1}$ and $\boldsymbol{m}$ have missing data, DIC_2 based on the likelihood including the nuisance models is Equation (16). We also can extend Equation (16) to DIC_2 excluding the nuisance models.

$$
\begin{aligned}
DIC_2 = -4 \sum_{j=1}^{J} \left\{ \frac{1}{K} \sum_{k=1}^{K} \log \left[f\left(\boldsymbol{y_j}|\boldsymbol{m_j^o}, \boldsymbol{m_j^m}, \hat{\boldsymbol{\theta}}\right) f\left(\boldsymbol{m_j^o}, \boldsymbol{m_j^m}|\boldsymbol{x_{1j}^o}, \boldsymbol{x_{1j}^m}, x_{2j}, \hat{\boldsymbol{\theta}}\right) f\left(\boldsymbol{x_{1j}^o}, \boldsymbol{x_{1j}^m}, x_{2j}|\hat{\boldsymbol{\theta}}\right) \right] \right\} \\
+ 2 \sum_{j=1}^{J} \log \left[\frac{1}{K} \sum_{k=1}^{K} f\left(\boldsymbol{y_j}|\boldsymbol{m_j^o}, \boldsymbol{m_j^m}, \hat{\boldsymbol{\theta}}\right) f\left(\boldsymbol{m_j^o}, \boldsymbol{m_j^m}|\boldsymbol{x_{1j}^o}, \boldsymbol{x_{1j}^m}, x_{2j}, \hat{\boldsymbol{\theta}}\right) f\left(\boldsymbol{x_{1j}^o}, \boldsymbol{x_{1j}^m}, x_{2j}|\hat{\boldsymbol{\theta}}\right) \right]
\end{aligned} \tag{16}
$$

WAIC. The Watanabe–Akaike Information Criterion ($WAIC$; also called widely available information criterion) is firstly proposed by Watanabe and Opper (2010). The typical equation of $WAIC$ is not for multilevel models, which uses individual scores as units. In multilevel models, Merkle et al. (2019) also uses individual scores (e.g., y_{ij}) as units, whereas we propose to use cluster vector scores (e.g., $\boldsymbol{y_j}$) as the unit. In addition, there is no literature about $WAIC$ with missing data to the best of our knowledge.

We extend $WAIC$ to incorporate missing data. Similar to DIC, we need to compute the effective number of parameters (p_{waic}) in $WAIC$. There are two ways to compute p_{waic} (Gelman et al., 2014), and we extend them to multilevel models with cluster vector scores as units. First, similar to computing DIC, p_{waic} can be estimated in practice as $p_{WAIC} = 2\sum_{j=1}^{J}\left(log\left[\frac{1}{K}\sum_{k=1}^{K} f\left(y_j|\hat{\theta}_k\right)\right] - \frac{1}{K}\sum_{k=1}^{K} logf\left(y_j|\hat{\theta}_k\right)\right)$. Second, $p_{WAIC} = \sum_{j=1}^{J} var_{k=1}^{K}\left(logf\left(y_j|\hat{\theta}_k\right)\right)$. The second way of computing is more stable than the first way because it computes the variance separately for each data point and sums up the variances (Gelman et al., 2014). Hence, we will focus on the second method for computing p_{WAIC}. We will also need to compute the log point-wise predictive density (lppd), $lppd = \sum_{j=1}^{J} log\left(\frac{1}{K}\sum_{k=1}^{K} f\left(y_j|\hat{\theta}_k\right)\right)$. Therefore, the general equation of $WAIC$ is as follows.

$$\begin{aligned} WAIC &= -2\left(lppd - p_{WAIC}\right) \\ &= -2\sum_{j=1}^{J} log\left(\frac{1}{K}\sum_{k=1}^{K} f\left(y_j|\hat{\theta}_k\right)\right) + 2\sum_{j=1}^{J} var_{k=1}^{K}\left(logf\left(y_j|\hat{\theta}_k\right)\right) \end{aligned} \tag{17}$$

In the multilevel mediation model, without missing data, we can calculate $WAIC$ in Equation (18) where $\boldsymbol{\theta}$ indicates parameters in marginal likelihoods, or parameters and random effects in conditional likelihoods, $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$ indicates the kth posterior sample of $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}$ indicates the posterior mean.

$$\begin{aligned} WAIC = &-2\sum_{j=1}^{J} log\left(\frac{1}{K}\sum_{k=1}^{K} f\left(\boldsymbol{y_j}, \boldsymbol{m_j}, \boldsymbol{x_{1j}}, x_{2j}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)\right) \\ &+ 2\sum_{j=1}^{J} var_{k=1}^{K}\left(logf\left(\boldsymbol{y_j}, \boldsymbol{m_j}, \boldsymbol{x_{1j}}, x_{2j}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)\right) \\ = &-2\sum_{j=1}^{J}\left\{log\left[\frac{1}{K}\sum_{k=1_{\theta_k}}^{K}\left(f\left(\boldsymbol{y_j}|, \boldsymbol{m_j}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) f\left(\boldsymbol{m_j}|\boldsymbol{x_{1j}}, x_{2j}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) f\left(\boldsymbol{x_{1j}}, x_{2j}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)\right)\right]\right\} \\ &+ 2\sum_{j=1}^{J} var_{k=1}^{K}\left(\log\left(f\left(\boldsymbol{y_j}|, \boldsymbol{m_j}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) f\left(\boldsymbol{m_j}|\boldsymbol{x_{1j}}, x_{2j}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) f\left(\boldsymbol{x_{1j}}, x_{2j}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)\right)\right) \end{aligned} \tag{18}$$

Suppose $\boldsymbol{x_1}$ and $\boldsymbol{m}$ have missing data, we impute missing data at each iteration and calculate $WAIC$

based on imputed data, where $\hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\mathbf{1},\boldsymbol{k}}$, $\hat{\boldsymbol{m}}^{\boldsymbol{m}}_{\boldsymbol{k}}$, and $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}$ indicate the kth posterior samples of $\boldsymbol{x_1}$, $\boldsymbol{m}$, and $\boldsymbol{\theta}$ respectively, $\hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\mathbf{1}\boldsymbol{j},\boldsymbol{k}}$ and $\hat{\boldsymbol{m}}^{\boldsymbol{m}}_{\boldsymbol{j},\boldsymbol{k}}$ indicate the kth posterior samples of $\boldsymbol{x_1}$ and $\boldsymbol{m}$ in the jth cluster, and $\hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\mathbf{1}}$, $\hat{\boldsymbol{m}}^{\boldsymbol{m}}$, and $\hat{\boldsymbol{\theta}}$ indicate the posterior means.

$$
\begin{aligned}
WAIC = & -2\sum_{j=1}^{J} log\left(\frac{1}{K}\sum_{k=1}^{K} f\left(\boldsymbol{y}, \boldsymbol{m^o}, \hat{\boldsymbol{m}}^{\boldsymbol{m}}_{\boldsymbol{k}}, \boldsymbol{x^o_1}, \hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\mathbf{1},\boldsymbol{k}}, \boldsymbol{x_2}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)\right) \\
& + 2\sum_{j=1}^{J} var^{K}_{k=1}\left(logf\left(\boldsymbol{y}, \boldsymbol{m^o}, \hat{\boldsymbol{m}}^{\boldsymbol{m}}, \boldsymbol{x^o_1}, \hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\mathbf{1}}, \boldsymbol{x_2}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)\right) \\
= & -2\sum_{j=1}^{J}\left\{log\left[\frac{1}{K}\sum_{k=1_{\theta_k}}^{K}\left(f\left(\boldsymbol{y_j}|\boldsymbol{m^o_j}, \hat{\boldsymbol{m}}^{\boldsymbol{m}}_{\boldsymbol{j},\boldsymbol{k}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) f\left(\boldsymbol{m^o_j}, \hat{\boldsymbol{m}}^{\boldsymbol{m}}_{\boldsymbol{j},\boldsymbol{k}}|\boldsymbol{x^o_{1j}}, \hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\mathbf{1}\boldsymbol{j}}, x_{2j}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) f\left(\boldsymbol{x^o_{1j}}, \hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\mathbf{1}\boldsymbol{j}}, x_{2j}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)\right)\right]\right\} \\
& + 2\sum_{j=1}^{J} var^{K}_{k=1}\left(\log\left(f\left(\boldsymbol{y_j}|\boldsymbol{m^o_j}, \hat{\boldsymbol{m}}^{\boldsymbol{m}}_{\boldsymbol{j},\boldsymbol{k}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) f\left(\boldsymbol{m^o_j}, \hat{\boldsymbol{m}}^{\boldsymbol{m}}_{\boldsymbol{j},\boldsymbol{k}}|\boldsymbol{x^o_{1j}}, \hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\mathbf{1}\boldsymbol{j}}, x_{2j}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right) f\left(\boldsymbol{x^o_{1j}}, \hat{\boldsymbol{x}}^{\boldsymbol{m}}_{\mathbf{1}\boldsymbol{j}}, x_{2j}|\hat{\boldsymbol{\theta}}_{\boldsymbol{k}}\right)\right)\right)
\end{aligned}
\tag{19}
$$

Simulation Study

Simulation Design

The data generating model (i.e., true model) is given in Equations (1) to (3). We rewrite the true model here again. The regression model for y_{ij} is $y_{ij} = \beta_{y,0j} + \beta_{y,1j}m_{ij} + e_{y,ij}$, $\beta_{y,0j} = \beta_{y,00} + u_{y,0j}$, and $\beta_{y,1j} = \beta_{y,10} + u_{y,1j}$. The regression model for m_{ij} is $m_{ij} = \beta_{m,0j} + \beta_{m,1}x_{1ij} + e_{m,ij}$ and $\beta_{m,0j} = \beta_{m,00} + \beta_{m,01}x_{2j} + u_{m,0j}$. The model for x_{1ij} and x_{2j} is $x_{1ij} = \beta_{x1,0j} + e_{x1,ij}$, $\beta_{x1,0j} = \beta_{x1,00} + \beta_{x1,01}x_{2j} + u_{x1,0j}$, and $x_{2j} = \beta_{x2,0} + u_{x2,0j}$.

We fitted data with the true model and four misspecified models (one more fixed effect, one less fixed effect, one more random effect, and one less random effect). In the "one more fixed effect" model, we wrongly assumed that y is also conditional on x_1.

$$
y_{ij} = \beta_{y,0j} + \beta_{y,1j}m_{ij} + \beta_{y,2}x_{1ij} + e_{y,ij} \tag{20}
$$

In the "one less fixed effect" model, we wrongly assumed that m is not conditional on x_2.

$$m_{ij} = \beta_{m,0j} + \beta_{m,1}x_{1ij} + e_{m,ij} \tag{21}$$

$$\beta_{m,0j} = \beta_{m,00} + u_{m,0j}$$

In the "one more random effect" model, we wrongly assumed a random slope in m_{ij}'s regression model.

$$m_{ij} = \beta_{m,0j} + \beta_{m,1j}x_{1ij} + e_{m,ij} \tag{22}$$

$$\beta_{m,0j} = \beta_{m,00} + \beta_{m,01}x_{2j} + u_{m,0j}$$

$$\beta_{m,1j} = \beta_{m,10} + u_{m,1j}$$

In the "one less random effect" model, we wrongly deleted a random slope in y_{ij}'s regression model.

$$y_{ij} = \beta_{y,0j} + \beta_{y,1}m_{ij} + e_{y,ij} \tag{23}$$

$$\beta_{y,0j} = \beta_{y,0} + u_{y,0j}$$

The fixed effects misspecified models led to one degree of freedom discrepancy, and the random effects misspecified models led to two degrees of freedom discrepancy because there would be one more/less variance and one more/less covariance.

We varied the values of four factors: the intraclass correlation coefficient of x_1, m, and y ($ICC = 0.1$ and 0.5), the proportion of missingness ($p_{miss} = 0$, 0.2, and 0.4), the sample size per cluster ($SZ = 5$, 15 and 30), and number of clusters ($J = 25$, 50, and 200). More specifically, we set the ICCs to be the same for x_1, m, and y (i.e., all of them are 0.1 or 0.5). We do not directly specify fixed and random effects in the model to simulate data; instead, we specify the proportion of explained variance and variance/covariance of x_1, x_2, m, and y based on real data sets to simulate the data (Enders et al., 2023). x_1

had a mean of 0 and a within-cluster variance of 1 (the between-cluster variance would be determined by ICC). x_2 had a mean of 0, a variance of 1, and a correlation with x_1's cluster mean of 0.3. m had a mean of 0, a total variance of 100, and the proportion of within-cluster variance explained by $\beta_{m,1}$ was 0.1 and the incremental proportion of between-cluster variance explained by $\beta_{m,01}$ was 0.1. These proportions are defined in Rights & Sterba (2019). y had a mean of 50, a total variance of 100, and the proportion of within-cluster variance explained by $\beta_{y,1}$ was 0.1, the proportion of within-cluster variance explained by the random slope ($u_{y,1j}$) was 0.1, and the correlation between the random intercept and slope was 0.3. We generated missing observations of m_{ij} and x_{1ij} based on y_{ij}. When one participant's y_{ij} is larger than the sample grand mean, this participant has a probability of p_{miss} to miss their m_{ij} and x_{1ij}. There are $2 \times 2 \times 3 = 12$ types of information criteria calculated in each condition: conditional versus marginal likelihood, the joint likelihood of y, m, x_1, and x_2 based on the sequential specification (denoted as xmy model) versus the joint likelihood of y and m based on the separate specification (i.e., excluding the likelihood contributions of x_1 and x_2, such that the information criteria were computed using only the y and m models; denoted as my model), and three criteria (DIC_1, DIC_2, and $WAIC$). We used the following notations to distinguish different likelihoods. We denote the marginal likelihood based information criteria with including covariate models (xmy model) as $m.xmy$ where m indicates the marginal likelihood, denote the marginal likelihood based information criteria with excluding covariate models (my model) as $m.my$ where m indicates the marginal likelihood, denote the conditional likelihood based information criteria with including covariate models as $c.xmy$ where c indicates the conditional likelihood, and denote conditional likelihood based information criteria with excluding covariate models as $c.my$ where c indicates the conditional likelihood. We conducted 2000 replications for each condition. The burn-in period is 30,000 and the post burn-in period is 20,000. We used version 3.1.28 of the Blimp application (Keller & Enders, 2021) to implement the fully Bayesian estimation and obtain the likelihoods. Then, we used R to compute DIC_1, DIC_2, and $WAIC$ based on the posterior samples and likelihoods from Blimp.

We compared DIC_1, DIC_2, and $WAIC$ with $m.xmy$, $m.my$, $c.xmy$, and $c.my$ likelihoods in terms of their proportions of selecting the true model. More specifically, we count the number of times each information criterion yields the lowest value for the true model compared to the four misspecified models because the true model is supposed to have the best model fit and thus lowest information criterion value. The best information criterion should have the highest detection rate of the true model (highest proportions of yielding the lowest value for the true model).

Simulation Results

We checked the convergence of each condition. Although generally the convergence criterion is $PSRF \leq 1.1$ (Gelman et al., 2014), we use a more strict convergence criterion, $PSRF \leq 1.05$ (Asparouhov & Muthén, 2010). That is, when the largest potential scale reduction factor ($PSRF$) among all parameters was larger than 1.05, we concluded that the model failed to converge. Only converged replications were kept and the convergence rates were higher than 92.7%.

Effects of ICC, p_{miss}, J, and SZ. We plot the proportions of selecting the true model from different information criteria with $ICC = 0.1$ and $p_{miss} = 0$ in Figure 1, $ICC = 0.1$ and $p_{miss} = 0.2$ in Figure 2, $ICC = 0.1$ and $p_{miss} = 0.4$ in Figure 3, $ICC = 0.5$ and $p_{miss} = 0$ in Figure 4, $ICC = 0.5$ and $p_{miss} = 0.2$ in Figure 5, and $ICC = 0.5$ and $p_{miss} = 0.4$ in Figure 6. In each figure, the row panel effects reflect the influence from the number of clusters (J), the column panel effects reflect the difference between likelihoods ($m.xmy$, $m.my$, $c.xmy$, or $c.my$), and the changes along the x-axis reflects the influence from the sample size per cluster (SZ).

Figure 1 contains no missing data ($p_{miss} = 0$). Based on Figure 1, more clusters (J) and a larger sample size per cluster (SZ) generally could increase the correct detection rates of information criteria based on marginal likelihood ($m.xmy$ and $m.my$) and conditional likelihood excluding covariate models ($c.my$). The increase perhaps does not only due to more information in the data but also due to more

accurate parameter estimation. For conditional likelihood including covariate models ($c.xmy$), more clusters and a larger sample size fail to increase the proportion of correct detection rates in DIC_1, DIC_2, and $WAIC$. The effect of the number of clusters and sample size did not have a noticeable difference on DIC_1, DIC_2, and $WAIC$.

Comparing Figures 1-3 with different proportions of missing data ($p_{miss} = 0$ vs. $p_{miss} = 0.2$ vs. $p_{miss} = 0.4$), we found that as the amount of missing data increased, this decreased the probabilities of all information criteria detecting the true model. Additionally, missing data had a larger influence on conditional likelihood information criteria ($c.my$ and $c.xmy$) than marginal likelihood information criteria ($m.my$ and $m.xmy$). Comparing Figures 4-6, we could see the same influence pattern of missing data.

Figures 1-3 and Figures 4-6 have different ICC. Comparing Figures 1 ($ICC = 0.1$) versus 4 ($ICC = 0.5$), Figures 2 ($ICC = 0.1$) versus 5 ($ICC = 0.5$), and Figures 3 ($ICC = 0.1$) versus 6 ($ICC = 0.5$), we found that higher ICC decreased the probabilities of all information criteria detecting the true model.

Marginal Likelihood vs. Conditional Likelihood . Within each figure, the marginal likelihood information criteria generally had much higher detection rates than conditional likelihood information criteria, especially with more clusters and/or a larger sample size per cluster. But when there were no missing data, conditional likelihood information criteria could have higher detection rates than marginal likelihood information criteria. With more missing data, the superiority of marginal likelihood over conditional likelihood information criteria became more obvious. Furthermore, we computed Monte Carlo errors of DIC_1, DIC_2, and $WAIC$ values (the standard deviation over converged replications) within each condition. Given the same likelihood, DIC_1, DIC_2, and $WAIC$ exhibited substantially similar Monte Carlo errors. But $c.xmy$ based information criteria had larger Monte Carlo errors compared to those based on $m.xmy$ (e.g., 212 vs. 30), while $c.my$ and $x.my$ based information criteria did not have noticeable differences in their Monte Carlo errors. Interesting readers can access the detailed results at

https://github.com/hduquant/dic.git.

Including vs. Excluding Covariate Models. Within conditional likelihood information criteria, there was a noticeable difference between including the likelihood of covariate models ($c.xmy$) and excluding the likelihood of covariate models ($c.my$). But based on the detection rates, it was unclear whether the covariate models as exogenous nuisance models need to be included in the model fit calculations. Including the covariate models in the information criteria sometimes provided higher detection rates but sometimes provided lower detection rates than the ones excluding the covariate models. We fail to find a clear pattern.

Within marginal likelihood information criteria, there were slightly higher detection rates when excluding the likelihood of covariate models ($m.my$) compared to including the likelihood of covariate models ($m.xmy$), however this difference was small.

DIC_1 vs. DIC_2 vs. $WAIC$. Across all of the 6 figures, within conditional likelihood information criteria, DIC_1 or $WAIC$ had the highest detection rates. Within marginal likelihood information criteria, DIC_2 usually had the highest detection rates, although DIC_1 also could have the highest detection rates when there were no missing data. Considering both conditional and marginal likelihoods, DIC_2 based on the marginal likelihoods generally provided higher detection rates than DIC_1 and $WAIC$ based on the conditional likelihoods except for a few cases with a small number of clusters ($J \leq 50$) and a small sample size per cluster ($SZ \leq 15$). If we combine the discussion of including vs. excluding covariate models in the previous section, DIC_2 based on the marginal likelihoods and excluding the likelihood of covariate models (denoted as $DIC_{2.m.my}$ where m indicates marginal likelihoods and my indicates the my model) was slightly better than the DIC_2 based on the marginal likelihoods and including the likelihood of covariate models (denoted as $DIC_{2.m.xmy}$ where xmy indicates the xmy model), and $DIC_{2.m.my}$ generally performed much better than $DIC_{1.m.my}$, $DIC_{1.c.my}$, $DIC_{1.m.xmy}$, $DIC_{1.c.xmy}$, $DIC_{2.c.my}$, $DIC_{2.c.xmy}$, $WAIC_{m.my}$, $WAIC_{c.my}$, $WAIC_{m.xmy}$, and $WAIC_{c.xmy}$ where m indicates marginal likelihoods, c

indicates conditional likelihoods, xmy indicates the xmy model, and my indicates the my model.

Selected Models . We conducted an additional analysis to explore the selection patterns using different information criteria, as the true model was not consistently chosen. The results showed that when we utilized the marginal likelihood ($m.my$ and $m.xmy$), the true model was predominantly selected. In cases where the true model was not chosen, there was a small proportion of selections favoring either the "one less fixed effect" or the "one less random effect" model. However, when we employed the conditional likelihood ($c.my$ and $c.xmy$), the "one more random effect" model was consistently favored, which implied that the conditional likelihood favored more complex models.

Cain and Zhang (2019) proposed to only select a winner when the difference of DIC is larger than 7. If the the difference of DIC is not larger than 7, we can choose the more substantively meaningful model or the simpler model. Therefore, we summarized the average proportion of DIC differences that were larger than 7 when the true model was selected in Table 2 (the difference is computed as DIC/WAIC of the misspecified model - DIC/WAIC of the true model, and the proportion is computed for each condition and averaged across conditions). Although from Table 1 we know that the conditional likelihood information criteria ($c.my$ and $c.xmy$) did not select the true model in most cases, from Table 2 we can see that once the conditional likelihood information criteria selected the true model, the DIC/WAIC of the true model generally was 7 points lower than the one from "one less random effect" model. The marginal likelihood information criteria ($m.my$ and $m.xmy$) were more likely to select the true model, and the proportions of DIC/WAIC difference larger than 7 were comparable to ones from $c.xmy$.

Overall, based on the simulation results, we suggest the marginal likelihood based DIC_2 that excludes the likelihood of covariate models ($DIC_{2.m.my}$).

Real Data Example

We use an employee dataset to illustrate the information criteria provided in Blimp (Keller & Enders, 2021). The data include several work-related variables (e.g., work satisfaction, turnover intention, employee–supervisor relationship quality) for a sample of 630 employees and 105 workgroups. We focus on three variables: employee empowerment composite (EMP), leader–member exchange (relationship quality with supervisor) composite (LME), and work satisfaction rating (WORK). We are interested in how LME affects WORK through the mediator EMP. 4.1% of the LME variable observations are missing, 16.2% of the EMP variable observations are missing, and 4.8% of the WORK variable observations are missing. We assume the missingness is MAR. We utilize group mean centering to isolate within-team variation in the regressors, but we are uncertain about whether allowing the within-team effects to vary across teams and including group means as predictors. As a result, we have developed three comparative models to investigate this further.

In the first model, we use group mean centering to isolate within-team variation in the regressors and we only consider random intercepts. Group means and missing data are estimated using MCMC sampling. The level-1 residuals ($e_{EMP,ij}$ and $e_{WORK,ij}$) and random intercepts ($\beta_{EMP,0j}$ and $\beta_{WORK,0j}$) follow normal distributions.

$$EMP_{ij} = \beta_{EMP,0j} + \beta_{EMP,1}\left(LMX_{ij} - \mu_{LMX,j}\right) + e_{EMP,ij},$$

$$WORK = \beta_{WORK,0j} + \beta_{WORK,1}\left(LMX_{ij} - \mu_{LMX,j}\right) + \beta_{WORK,2}\left(EMP_{ij} - \mu_{EMP,j}\right) + e_{WORK,ij}$$

In the second model, we add random slopes. The random intercepts ($\beta_{EMP,0j}$ and $\beta_{WORK,0j}$) and random

slopes ($\beta_{EMP,1j}$, $\beta_{WORK,1j}$, and $\beta_{WORK,2j}$) follow multivariate normal distributions.

$$EMP_{ij} = \beta_{EMP,0j} + \beta_{EMP,1j}\left(LMX_{ij} - \mu_{LMX,j}\right) + e_{EMP,ij},$$

$$WORK = \beta_{WORK,0j} + \beta_{WORK,1j}\left(LMX_{ij} - \mu_{LMX,j}\right) + \beta_{WORK,2j}\left(EMP_{ij} - \mu_{EMP,j}\right) + e_{WORK,ij}$$

In the third model, we add between cluster effects. $\beta_{EMP,2}$, $\beta_{WORK,3}$, and $\beta_{WORK,4}$ indicate the effect from the cluster means of regressors.

$$EMP_{ij} = \beta_{EMP,0j} + \beta_{EMP,1j}\left(LMX_{ij} - \mu_{LMX,j}\right) + \beta_{EMP,2}\mu_{LMX,j} + e_{EMP,ij},$$

$$WORK = \beta_{WORK,0j} + \beta_{WORK,1j}\left(LMX_{ij} - \mu_{LMX,j}\right) + \beta_{WORK,2j}\left(EMP_{ij} - \mu_{EMP,j}\right) + \beta_{WORK,3}\mu_{LMX,j} + \beta_{WORK,4}\mu_{EMP,j} + e_{WORK,ij}$$

We implemented the fully Bayesian estimation in version 3.1.28 of the Blimp application (Keller & Enders, 2021). We used R to compute DIC_1, DIC_2, and $WAIC$ based on the posterior samples and likelihoods output from Blimp. We excluded the likelihood of covariate models ($c.my$ and $m.my$). The data and code (both the Blimp and R code) are available at https://github.com/hduquant/dic.git. The burn-in period is 5,000 and the post burn-in period is 10,000. The marginal likelihood and conditional likelihood information criteria ($DIC_{1.m.my}$, $DIC_{2.m.my}$, $WAIC_{m.my}$, $DIC_{1.c.my}$, $DIC_{2.c.my}$, and $WAIC_{c.my}$) are presented in Table 1. The marginal likelihood based DIC_2 and $WAIC$ ($DIC_{2.m.my}$ and $WAIC_{m.my}$) and the conditional likelihood based DIC_1, DIC_2, and $WAIC$ ($DIC_{1.c.my}$, $DIC_{2.c.my}$, and $WAIC_{c.my}$) indicated that Model 2 had the best model fit, whereas the marginal likelihood based DIC_1 ($DIC_{1.m.my}$) indicated that Model 3 had the best model fit. Since compared to $DIC_{2.m.my}$, $DIC_{1.m.my}$ had a lower detection rate in the simulation, we concluded that Model 2 has the best model fit.

Discussion

It is common in applications of Bayesian estimation to compare different models. One may like to expand the model after successfully fitting a simple model, and one may compare non-nested models with different sets of predictors. In this case, we need to use model fit criteria to compare and select models. In this paper, we focus on Deviance Information Criterion (DIC) and Watanabe–Akaike Information Criterion ($WAIC$) in multilevel mediation models. The current study has four contributions. First, we propose a DIC_2 to compare to the traditional DIC_1 and $WAIC$ in terms of true model selection accuracy. Second, we explore whether these information criteria with conditional likelihood or marginal likelihood lead to higher accurate selection rates. Third, there is no literature about $WAIC$ with missing data to the best of our knowledge, and we extend $WAIC$ to incorporate missing data. Fourth, we explore whether it is necessary to include the nuisance models of exogenous variables in likelihood to compute information criteria. We summarize our findings as follows. Note that while we anticipate that the findings presented can be applied to other multilevel models (e.g., multilevel partial mediation and latent growth curve models), it would be prudent to conduct simulations in order to verify these conclusions.

First, comparing conditional and marginal likelihood information criteria, the marginal likelihood information criteria generally had higher detection rates than conditional likelihood information criteria. This is consistent with the findings in the previous studies by Merkle et al. (2019), Tong et al. (2022), and Zhang et al. (2019). We also find that more missing data increases the discrepancy between marginal likelihood and conditional likelihood information criteria.

Second, the performance of $WAIC$ with missing data depends on whether we use marginal likelihoods or conditional likelihoods. If we used conditional likelihoods, $WAIC$ could have a higher true model selection rate than DIC_1 and DIC_2. If we used marginal likelihoods, $WAIC$ failed to outperform DIC_2, but could be better or worse than DIC_1. We notice that the difference between $WAIC$ and DIC_2 was larger with missing data. It is possible that missing data can influence the posterior variance, which is

the basis for $WAIC$ calculation.

Third, comparing DIC_1, DIC_2, and $WAIC$, our proposed DIC_2 had the highest true model selection rates with marginal likelihoods but had the lowest rates with conditional likelihoods, while marginal likelihood based DIC_2 had a higher true model selection rate than conditional likelihood based DIC_1 and $WAIC$. Additionally, with marginal likelihoods, the difference between DIC_1, DIC_2, and $WAIC$ was small; but with conditional likelihoods, DIC_1, DIC_2, and $WAIC$ had more diverse performance. This is consistent with the conclusion of comparing DIC_1 and $WAIC$ in Tong et al. (2022).

Fourth, in terms of whether to include the nuisance models of exogenous variables in information criteria computation, the performance depended on the likelihood function. With marginal likelihoods, the marginal likelihood information criteria with excluding covariate models were slightly better than the ones including covariate models, while the difference was small. Within conditional likelihoods, including the covariate models in the information criteria could provide higher or lower detection rates than the ones excluding the covariate models. There was not a clear pattern.

The answer of whether DIC_2 was better than DIC_1 and $WAIC$ and whether we should include the nuisance models of exogenous variables in likelihood functions depended on whether we use marginal or conditional likelihoods. Overall, the marginal likelihood based DIC_2 that excludes the likelihood of covariate models ($DIC_{2.m.my}$) generally had the highest true model selection rates, and we recommend utilizing it in practical applications.

Open Practices Statements

The data and code (both the Blimp and R code) are available at https://github.com/hduquant/dic.git.

References

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis using mplus: Technical implementation (version 3).* Citeseer.

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2015). Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462–487.

Cain, M. K., & Zhang, Z. (2019). Fit for a bayesian: An evaluation of ppp and dic for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 39–50.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1), 1–32.

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

Du, H., Alacam, E., Mena, S., & Keller, B. T. (2022). Compatibility in imputation specification. *Behavior Research Methods*, 1–19.

Enders, C. K. (2022). *Applied missing data analysis (2nd ed.)*. New York: Guilford press.

Enders, C. K., Du, H., & Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and non-linear terms. *Psychological Methods*, *25*(1), 88–112.

Enders, C. K., Hayes, T., & Du, H. (2018). A comparison of multilevel imputation schemes for random coefficient models: Fully conditional specification and joint model imputation with random covariance matrices. *Multivariate Behavioral Research*, *53*(5), 695–713.

Enders, C. K., Keller, B. T., & Woller, M. T. (2023). A simple "monte carlo" method for estimating power in multilevel designs. *Revised manuscript submitted for publication to Psychological methods*.

Erler, N. S., Rizopoulos, D., Jaddoe, V. W., Franco, O. H., & Lesaffre, E. M. (2019). Bayesian imputation of time-varying covariates in linear mixed models. *Statistical Methods in Medical Research*, *28*(2), 555–568.

Erler, N. S., Rizopoulos, D., Rosmalen, J. v., Jaddoe, V. W., Franco, O. H., & Lesaffre, E. M. (2016). Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full bayesian approach. *Statistics in Medicine*, *35*(17), 2955–2974.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). London: Chapman & Hall.

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

Goldstein, H., Carpenter, J. R., & Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *177*(2), 553–564.

Gronau, Q. F., & Wagenmakers, E.-J. (2019). Limitations of bayesian leave-one-out cross-validation for model selection. *Computational brain & behavior*, *2*, 1–11.

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, *21*(1), 111–149.

Grund, S., Lüdtke, O., & Robitzsch, A. (2021). Multiple imputation of missing data in multilevel models with the r package mdmb: a flexible sequential modeling approach. *Behavior research methods*, *53*, 2631–2649.

Ibrahim, J. G., Chen, M.-H., & Lipsitz, S. R. (1999). Monte carlo em for missing covariates in parametric regression models. *Biometrics*, *55*(2), 591–596.

Keller, B. T., & Enders, C. K. (2021). Blimp userâs guide (version 3). Retrieved from www.appliedmissingdata.com/multilevel-imputation.html

Kim, S., Belin, T. R., & Sugar, C. A. (2018). Multiple imputation with non-additively related variables: Joint-modeling and approximations. *Statistical Methods in Medical Research*, *27*(6), 1683–1694.

Kim, S., Sugar, C. A., & Belin, T. R. (2015). Evaluating model-based imputation methods for missing covariates in regression models with interactions. *Statistics in Medicine*, *34*(11), 1876–1888.

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.

Lu, L., & Zhang, Z. (2022). How to select the best fit model among bayesian latent growth models for complex data. *Journal of Behavioral Data Science*, *2*(1), 35–58.

Lüdtke, O., Robitzsch, A., & West, S. G. (2020). Regression models involving nonlinear effects with missing data: A sequential modeling approach using bayesian estimation. *Psychological Methods*, *25*(2), 157–181.

Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *psychometrika*, *84*(3), 802–829.

Merkle, E. C., & Rosseel, Y. (2015). blavaan: Bayesian structural equation models via parameter expansion. *arXiv preprint arXiv:1511.05604*.

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.

Muthén, L., & Muthén, B. (1998–2017). *Mplus userâs guide. 8th edition*. Los Angeles, CA: Author.

Plummer, M., et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, pp. 1–10).

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining r-squared measures. *Psychological methods*, *24*(3), 309.

Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Medical Research Methodology*, *12*(1), 46.

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

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.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W., & Lunn, D. (1996). Bugs: Bayesian inference using gibbs sampling. *Version 0.5,(version ii) http://www. mrc-bsu. cam. ac. uk/bugs*, *19*.

Tong, X., Kim, S., & Ke, Z. (2022). Impact of likelihoods on class enumeration in bayesian growth mixture modeling. In M. Wiberg, D. Molenaar, J. González, J.-S. Kim, & H. Hwang (Eds.), *Quantitative psychology* (pp. 111–120). Cham: Springer International Publishing.

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., BÃŒrkner, P.-C., Paananen, T., & Gelman, A. (2023). *loo: Efficient leave-one-out cross-validation and waic for bayesian models.* Retrieved from `https://mc-stan.org/loo/` (R package version 2.6.0)

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, *27*, 1413–1432.

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., & Winther, O. (2016). Bayesian leave-one-out cross-validation approximations for gaussian latent variable models. *The Journal of Machine Learning Research*, *17*(1), 3581–3618.

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, *11*(12).

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.

Footnotes

[1] lavaan does not allow treating predictors/covariates as random when the model contains nonlinear covariate effects.

Table 1: The Proportion of Each Model Selected with Different Information Criteria

		True	One less fixed effect	One more fixed effect	One less random effect	One more random effect
$m.my$	DIC_1	0.70	0.11	0	0.19	0
	DIC_2	0.72	0.11	0	0.17	0
	$WAIC$	0.72	0.11	0	0.17	0
$c.my$	DIC_1	0.35	0.06	0	0.02	0.57
	DIC_2	0	0	0	0	1
	$WAIC$	0.11	0.07	0	0	0.81
$m.xmy$	DIC_1	0.72	0.11	0	0.17	0
	DIC_2	0.72	0.11	0	0.17	0
	$WAIC$	0.70	0.11	0	0.19	0
$c.xmy$	DIC_1	0.33	0	0	0	0.67
	DIC_2	0	0	0	0	1
	$WAIC$	0.33	0	0	0	0.67

Note: the lowest information criteria are bolded

Table 2: The proportion of DIC differences that were larger than 7 when the true model was selected

		One less fixed effect	One more fixed effect	One less random effect	One more random effect
$m.my$	DIC_1	0.42	0	0.73	0.09
	DIC_2	0.42	0	0.75	0.02
	$WAIC$	0.42	0.01	0.74	0.05
$c.my$	DIC_1	0.14	0	0.87	0
	DIC_2	0.17	0	0.92	0
	$WAIC$	0.12	0.02	0.88	0.01
$m.xmy$	DIC_1	0.41	0	0.73	0.1
	DIC_2	0.41	0	0.75	0.04
	$WAIC$	0.42	0.03	0.74	0.18
$c.xmy$	DIC_1	0.46	0.09	1	0.03
	DIC_2	0.03	0.1	1	0.02
	$WAIC$	0.41	0.28	1	0.17

Table 3: The Marginal Likelihood and Conditional Likelihood Information Criteria in the Real Data Example

	Model 1	Model 2	Model 3
Marginal Likelihood			
DIC_1	5612.406	5604.815	**5604.185**
DIC_2	5518.996	**5507.320**	5511.442
$WAIC$	5576.719	**5566.584**	5570.962
Conditional Likelihood			
DIC_1	5578.686	**5513.055**	5513.22
DIC_2	5448.113	**5358.068**	5366.736
$WAIC$	5541.298	**5478.100**	5488.689

Note: the lowest information criteria are bolded. We excluded the likelihood of covariate models.

Figure Captions

Figure 1. Plot of the proportions of selecting the true model when $ICC = 0.1$ and $p_{miss} = 0$

Figure 2. Plot of the proportions of selecting the true model when $ICC = 0.1$ and $p_{miss} = 0.2$

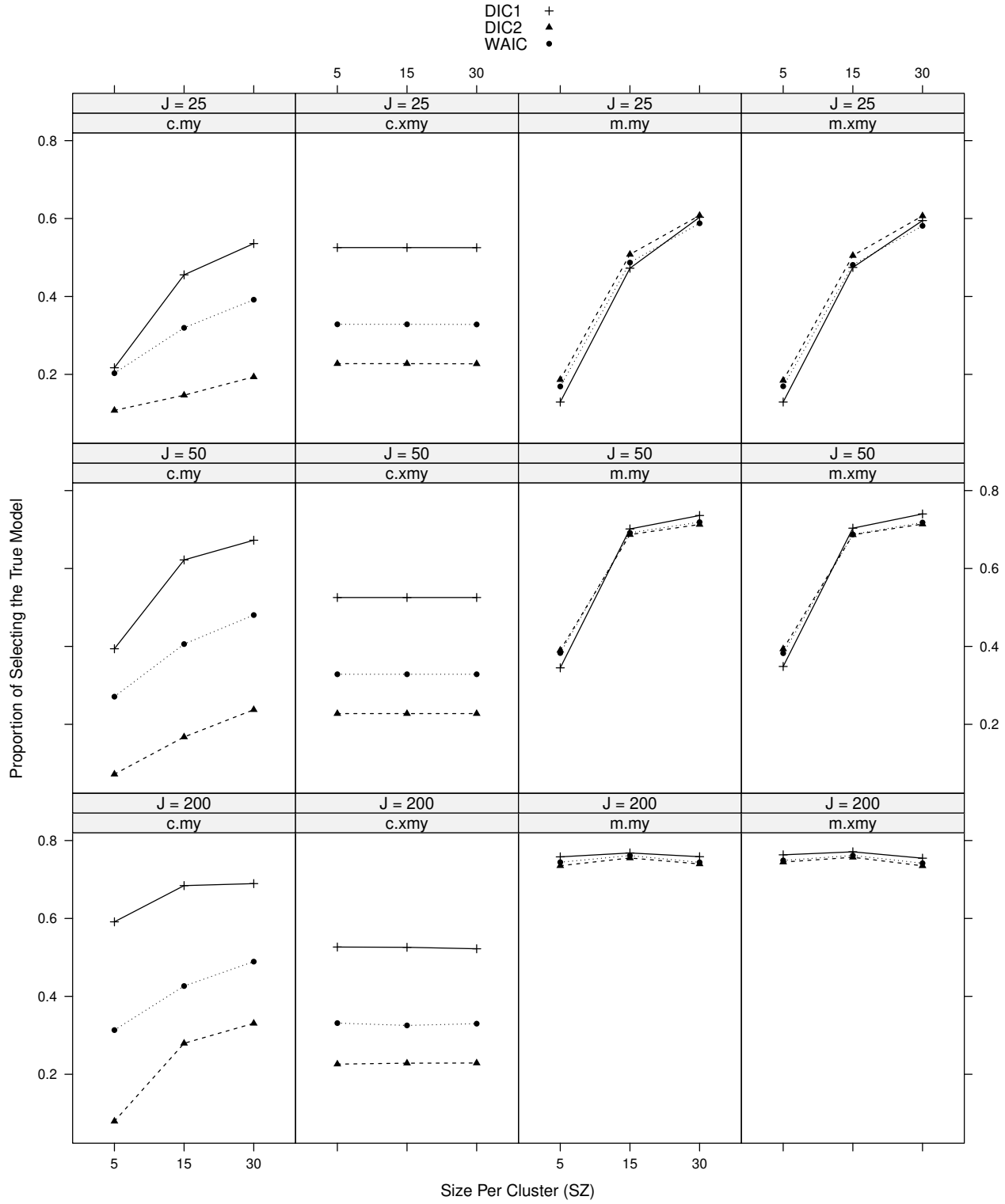
Figure 3. Plot of the proportions of selecting the true model when $ICC = 0.1$ and $p_{miss} = 0.4$

Figure 4. Plot of the proportions of selecting the true model when $ICC = 0.5$ and $p_{miss} = 0$

Figure 5. Plot of the proportions of selecting the true model when $ICC = 0.5$ and $p_{miss} = 0.2$

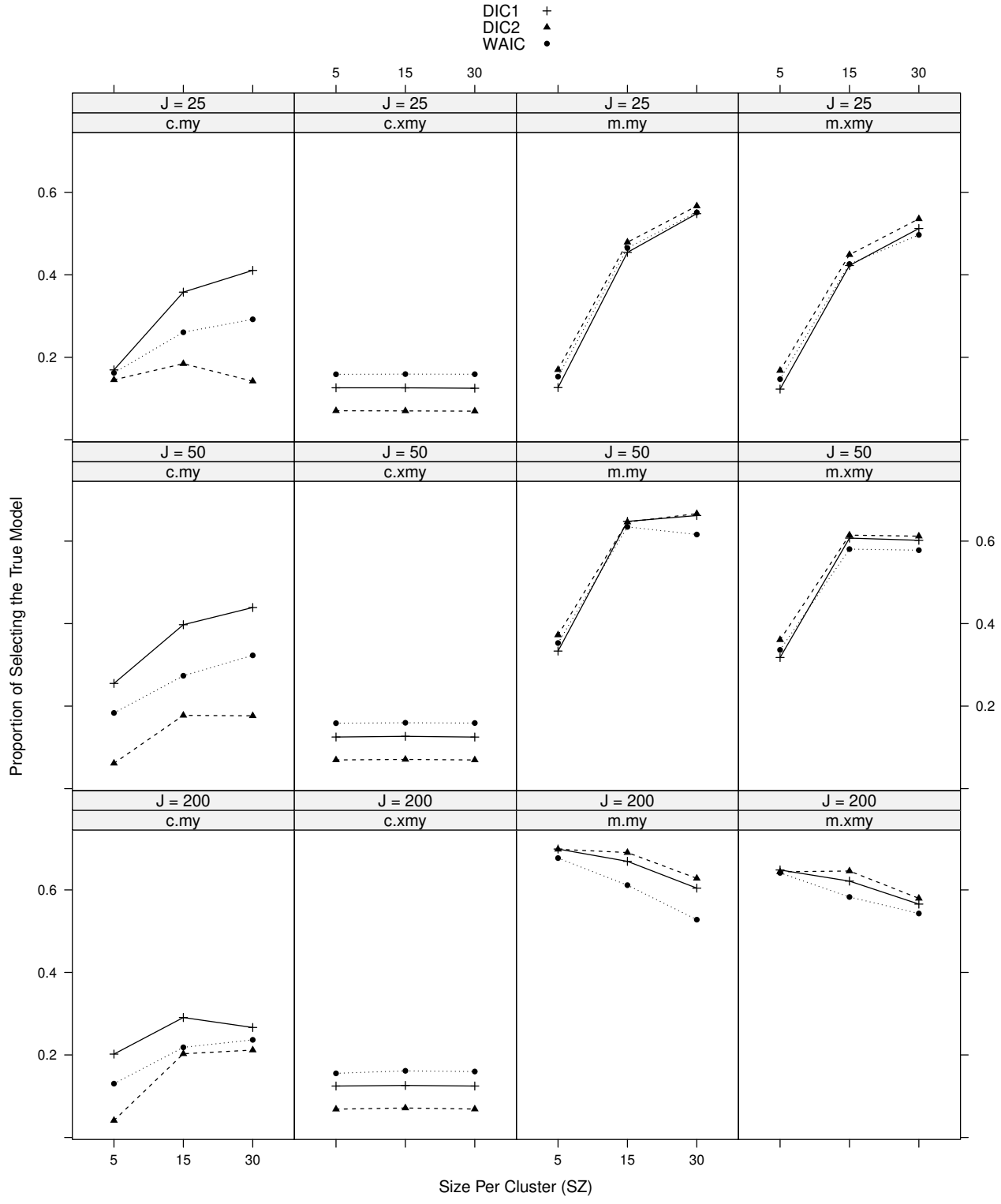
Figure 6. Plot of the proportions of selecting the true model when $ICC = 0.5$ and $p_{miss} = 0.4$

Figure 1: Plot of the proportions of selecting the true model when $ICC = 0.1$ and $p_{miss} = 0$



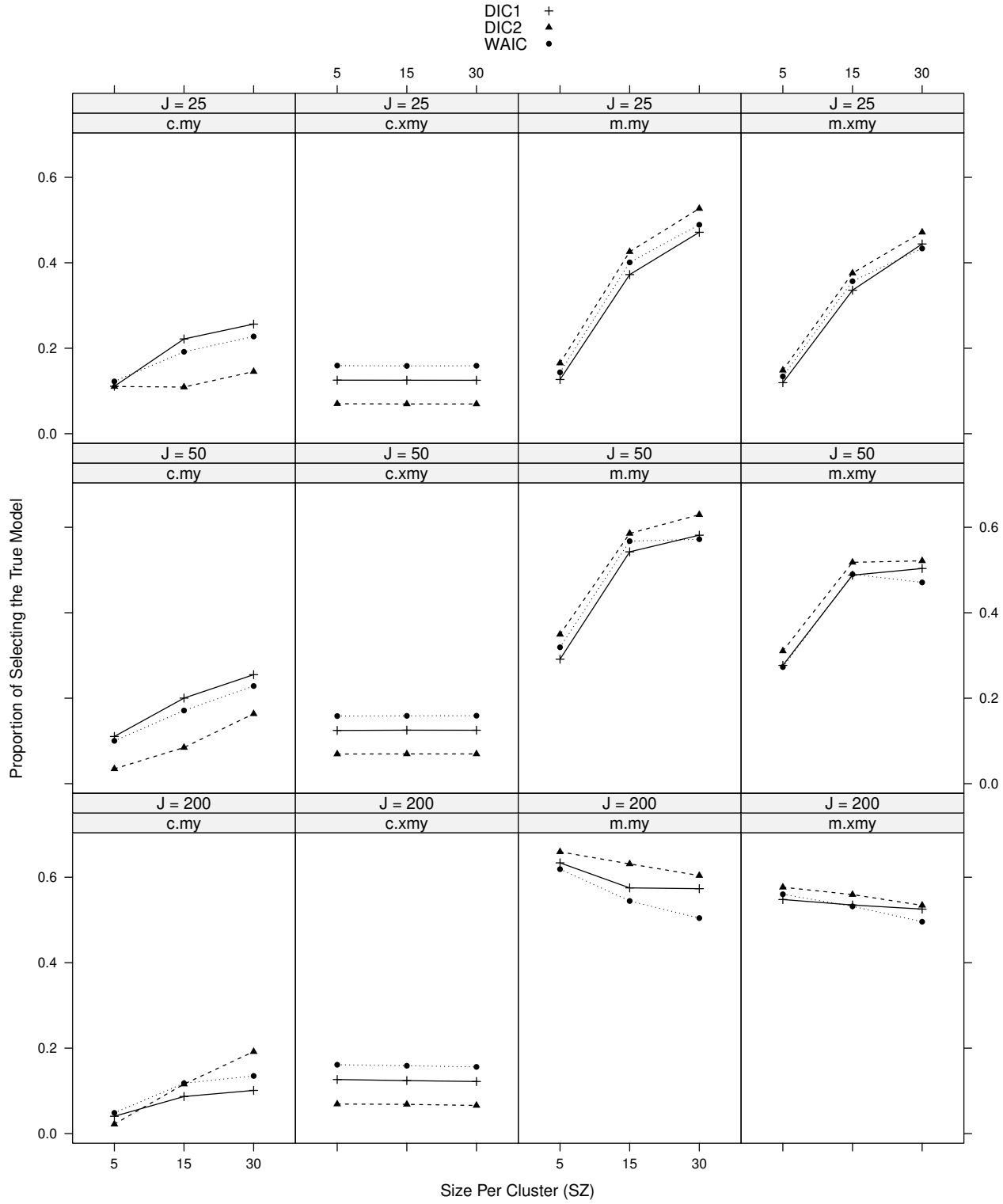
Note: ICC indicates the intraclass correlation coefficient of x_1, m, and y, p_{miss} indicates the proportion of missingness, J indicates the number of cluster, SZ indicates the sample size per cluster, $m.xmy$ indicates the marginal likelihood including covariate models, $c.xmy$ indicates the conditional likelihood including covariate models, $m.my$ indicates the marginal likelihood excluding covariate models, and $c.my$ indicates the conditional likelihood excluding covariate models.

Figure 2: Plot of the proportions of selecting the true model when $ICC = 0.1$ and $p_{miss} = 0.2$



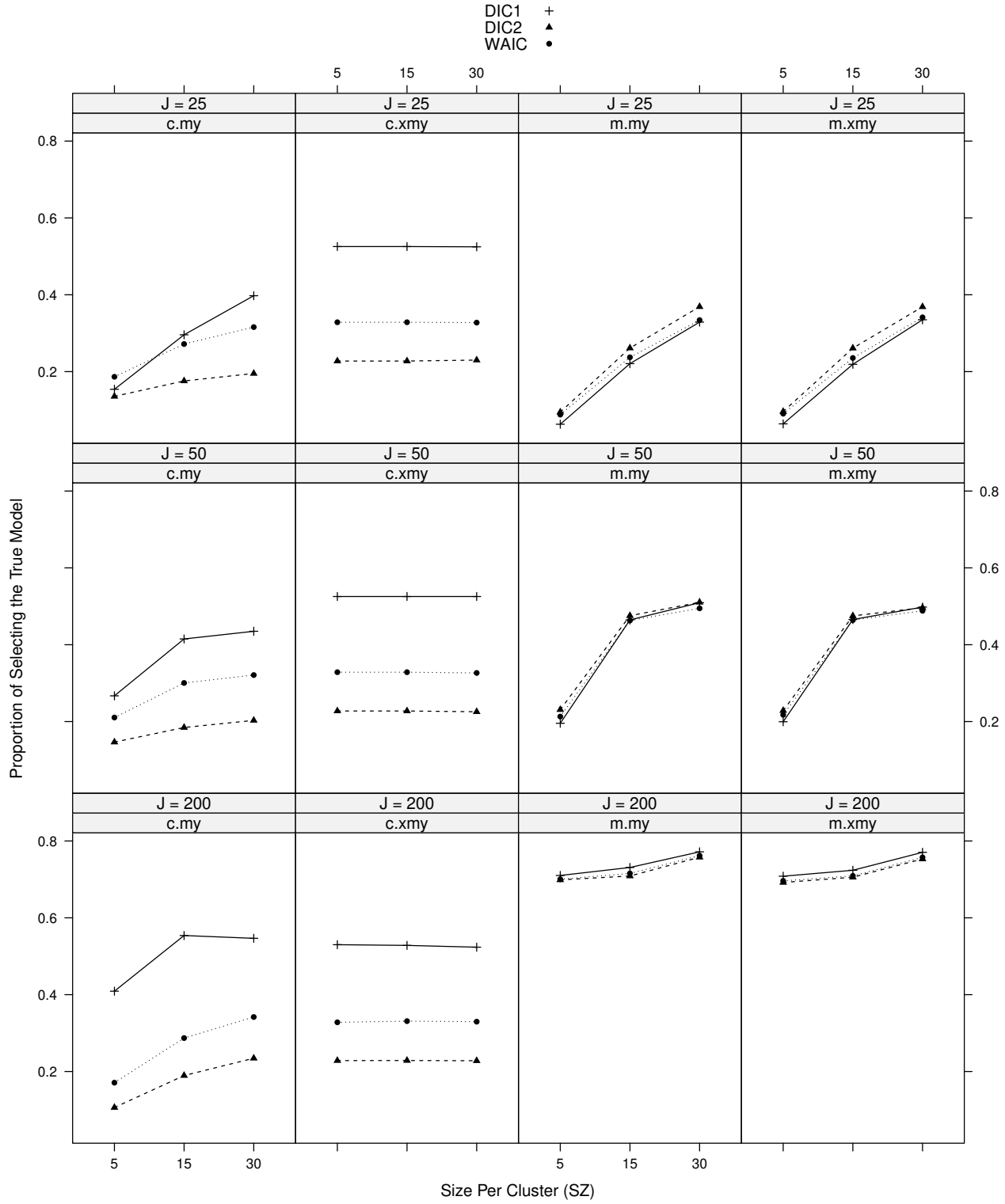
Note: ICC indicates the intraclass correlation coefficient of x_1, m, and y, p_{miss} indicates the proportion of missingness, J indicates the number of cluster, SZ indicates the sample size per cluster, $m.xmy$ indicates the marginal likelihood including covariate models, $c.xmy$ indicates the conditional likelihood including covariate models, $m.my$ indicates the marginal likelihood excluding covariate models, and $c.my$ indicates the conditional likelihood excluding covariate models.

Figure 3: Plot of the proportions of selecting the true model when $ICC = 0.1$ and $p_{miss} = 0.4$



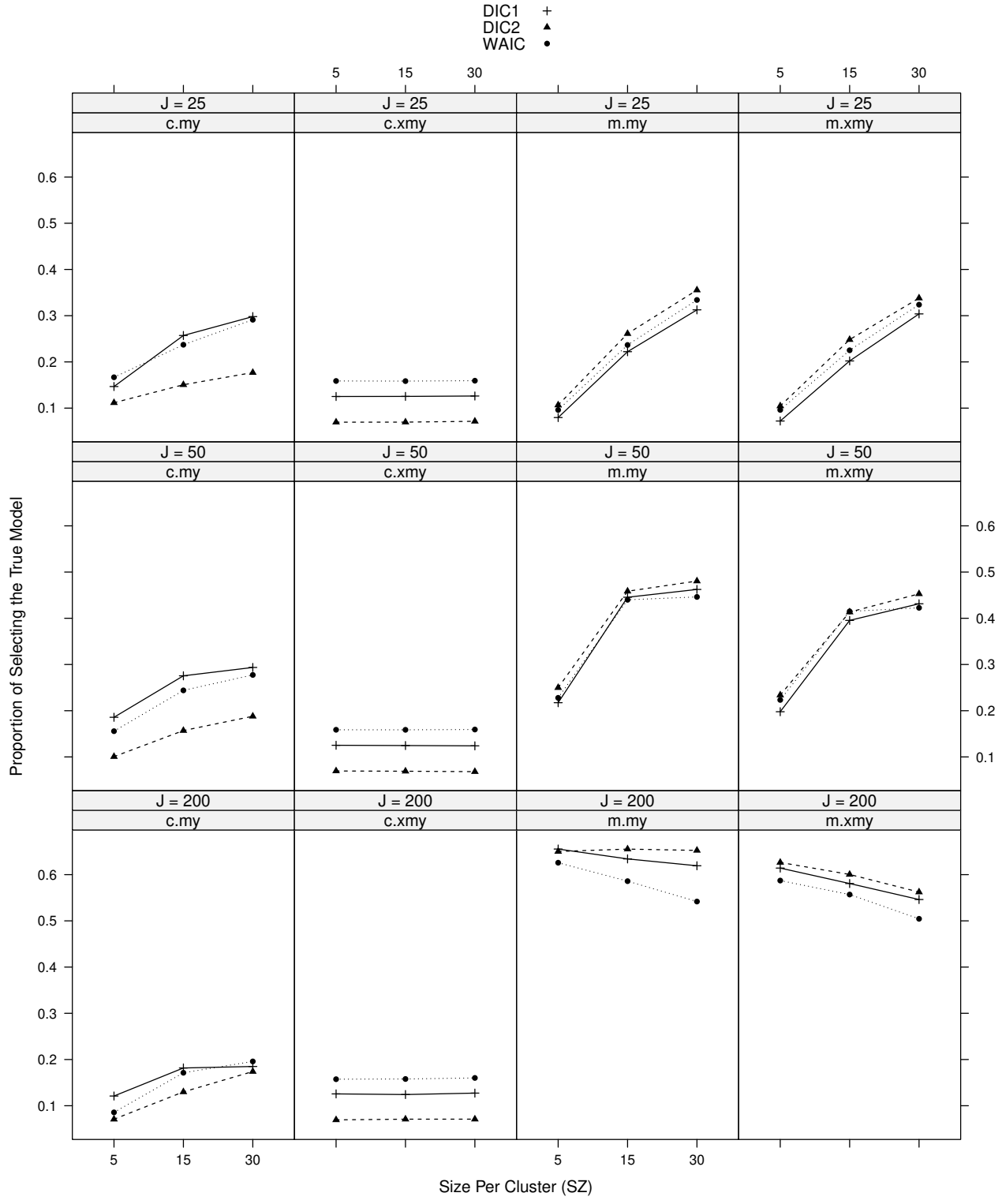
Note: ICC indicates the intraclass correlation coefficient of x_1, m, and y, p_{miss} indicates the proportion of missingness, J indicates the number of cluster, SZ indicates the sample size per cluster, $m.xmy$ indicates the marginal likelihood including covariate models, $c.xmy$ indicates the conditional likelihood including covariate models, $m.my$ indicates the marginal likelihood excluding covariate models, and $c.my$ indicates the conditional likelihood excluding covariate models.

Figure 4: Plot of the proportions of selecting the true model when $ICC = 0.5$ and $p_{miss} = 0$



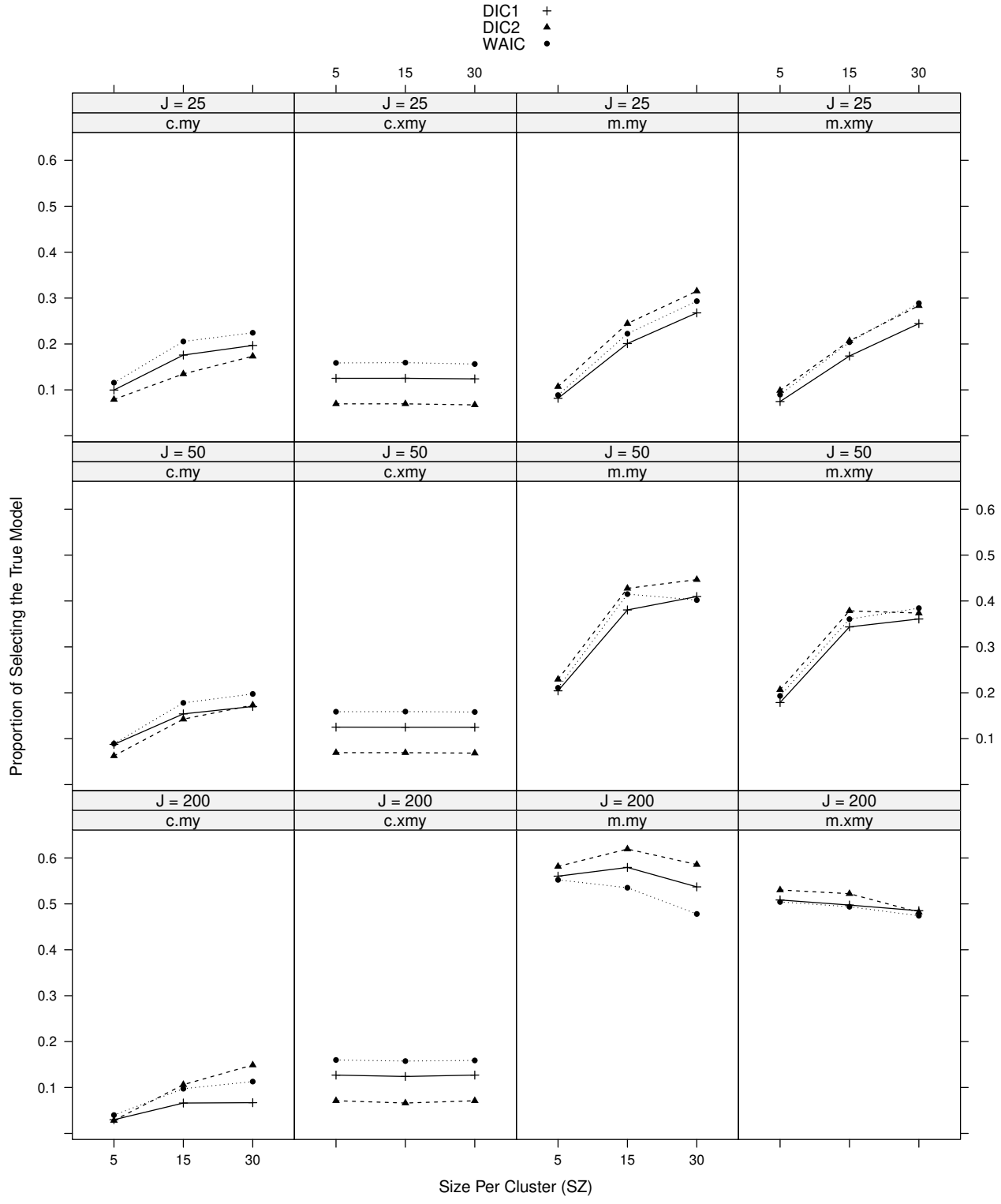
Note: ICC indicates the intraclass correlation coefficient of x_1, m, and y, p_{miss} indicates the proportion of missingness, J indicates the number of cluster, SZ indicates the sample size per cluster, $m.xmy$ indicates the marginal likelihood including covariate models, $c.xmy$ indicates the conditional likelihood including covariate models, $m.my$ indicates the marginal likelihood excluding covariate models, and $c.my$ indicates the conditional likelihood excluding covariate models.

Figure 5: Plot of the proportions of selecting the true model when $ICC = 0.5$ and $p_{miss} = 0.2$



Note: ICC indicates the intraclass correlation coefficient of x_1, m, and y, p_{miss} indicates the proportion of missingness, J indicates the number of cluster, SZ indicates the sample size per cluster, $m.xmy$ indicates the marginal likelihood including covariate models, $c.xmy$ indicates the conditional likelihood including covariate models, $m.my$ indicates the marginal likelihood excluding covariate models, and $c.my$ indicates the conditional likelihood excluding covariate models.

Figure 6: Plot of the proportions of selecting the true model when $ICC = 0.5$ and $p_{miss} = 0.4$



Note: ICC indicates the intraclass correlation coefficient of x_1, m, and y, p_{miss} indicates the proportion of missingness, J indicates the number of cluster, SZ indicates the sample size per cluster, $m.xmy$ indicates the marginal likelihood including covariate models, $c.xmy$ indicates the conditional likelihood including covariate models, $m.my$ indicates the marginal likelihood excluding covariate models, and $c.my$ indicates the conditional likelihood excluding covariate models.