JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

DOI: 10.1111/jedm.12408

*Journal of Educational Measurement*
*Winter 2024, Vol. 61, No. 4, pp. 740–762*

# Model Selection Posterior Predictive Model Checking via Limited-Information Indices for Bayesian Diagnostic Classification Modeling

**Jihong Zhang**
*University of Arkansas*
**Jonathan Templin**
*University of Iowa*
**Xinya Liang**
*University of Arkansas*

*Recently, Bayesian diagnostic classification modeling has been becoming popular in health psychology, education, and sociology. Typically information criteria are used for model selection when researchers want to choose the best model among alternative models. In Bayesian estimation, posterior predictive checking is a flexible Bayesian model evaluation tool, which allows researchers to detect Q-matrix misspecification. However, model selection methods using posterior predictive checking (PPC) for Bayesian DCM are not well investigated. Thus, this research aims to propose a novel model selection approach using posterior predictive checking with limited-information statistics for selecting the correct Q-matrix. A simulation study was conducted to examine the performance of the proposed method. Furthermore, an empirical example was provided to illustrate how it can be used in real scenarios.*

## Introduction

Bayesian diagnostic classification models (BDCMs) have recently gained more attention across multiple disciplines (Hu & Templin, 2020; Thompson, 2020). Both model selection and model fit evaluation methods play important roles in a model-building sequence of psychometric models, but the link between these two components in an analysis has not been extensively examined in the literature on BDCMs. This paper seeks to fill the gap of model selection problems in BDCMs by proposing a novel model selection approach based on limited-information model fit methods, as implemented in Bayesian posterior predictive model checking. The novelty of the proposed approach stems from two aspects: first, few previous studies have utilized limited-information statistics as test statistics in model selection and model selection within Bayesian DCMs; second, to our knowledge, this is the first study employing Bayesian Networks (BN; Almond et al., 2007) as a fully Bayesian model selection method for DCMs.

The core component of model selection of diagnostic classification modeling is selecting an appropriate *Q-matrix*. A Q-matrix is an indicator matrix linking the items to the latent constructs they measure (i.e., attributes; e.g., Tatsuoka, 1983). The Q-matrix is usually established by expert judgment, leading to uncertainties

about some of its elements. To address this concern, two primary strategies are commonly used to identify the best-fitting model: Q-matrix validation approaches and model selection approaches. Many Q-matrix validation methods attempt to reconstruct the Q-matrix by specifying certain elements of Q-matrix as random variables and subsequently penalize them using a shrinkage approach (e.g., DeCarlo, 2012). The alternative model selection methods select a best-fitting Q-matrix by comparing multiple Q-matrices using model fit measures. Typical model fit indices include information criteria, such as Akaike's information coefficient (AIC; Akaike, 1974), Bayesian information coefficient (BIC; Schwarz, 1978), and the Watanabe-Akaike information criterion (WAIC; Watanabe, 2010). However, the effectiveness of these information indices in selecting the Bayesian diagnostic classification model from item response theory (IRT) has not been demonstrated adequately (Sen & Bradshaw, 2017; Zhang et al., 2019). In addition, although previous studies have investigated the performance of several fit indices in choosing the correct DCM in the frequentist framework (Lei & Li, 2016), the comparative performance of these information indices in selecting Bayesian DCMs with various levels of Q-matrix misspecification has yet been well investigated. Thus, the main purpose of this study is to propose a fully Bayesian model selection approach based on limited-information statistics and Bayesian network for comparing multiple Bayesian DCMs with various types of Q-matrix misspecification.

Specifically, our proposed approach employs the log-linear cognitive diagnosis model (LCDM; Henson et al., 2008) as both the data-generation model and data-analysis model. The LCDM was chosen due to its status as the saturated version of many diagnostic classification model variants (Henson et al., 2008) and its robustness in cases where a hierarchical structure exists within the DCM (Templin and Bradshaw, 2014). The limited-information statistic, $M_2$, was used as the discrepancy measure of the posterior predictive modeling to evaluate the goodness-of-fit of models (Maydeu-Olivares & Joe, 2005). The $M_2$ statistic is a model-data fitting index making use of only up to second-order marginal probabilities of the data tables and has shown good performance in simulation studies (Maydeu-Olivares, 2013). Instead of using the point estimate of $M_2$ in the posterior predictive checking process for model fit evaluation, the Kolmogorov-Smirnov (KS) statistic—a nonparametric method of describing the highest distances between two samples—was used to evaluate the entire space of posterior predictive distributions between the proposed model and the saturated model.

The rest of the paper is organized as follows. First, "Background" section reviews the background of key components of the proposed method, including diagnostic classification modeling, Bayesian Network modeling, and the general form of the posterior predictive model checking method. Second, the proposed KS statistics-based posterior predictive model checking, utilizing limited-information $M_2$ statistics as summary statistics (hereafter referred to as KS-PP-M2), is introduced in "KS-PP-M2" section. Third, a Monte Carlo simulation study was performed to examine the performance of the proposed approach in model selection with various types of Q-matrix misspecification in "Simulation Study" section. Fourth, an empirical study was conducted to illustrate how to apply the proposed method in real scenarios in

"Empirical Study" section. Finally, "Discussion" section provides a discussion about the advantages and limitations of the proposed method.

To comprehensively examine the proposed approach for model selection, the following main research questions are investigated:

**RQ1** Is KS-PP-M2 sensitive to the model-data misfit with various degrees of Q-matrix misspecification?

**RQ2** Compared to conventional information criteria, does the proposed KS-PP-M2 approach yield higher accuracy in selecting the correct model?

**RQ3** How does the overall discrimination power affect the performance of the fit indices in choosing the correct model?

## Background

### Diagnostic Classification Modeling

Diagnostic Classification Models seek to provide each individual's skill mastery profile (e.g., whether or not students have mastered a subtraction skill in a mathematical assessment), which can be further used for developing targeted interventions. From the statistical perspective, DCMs are a family of restricted latent class models, which classify samples into attribute profiles (also known as attribute patterns) based on observed responses (typically assessment item responses). Numerous types of DCMs have been proposed based on various research questions, each with a different set of assumptions regarding how latent attributes interact to produce item responses. Among DCMs, LCDM is one of a set of general diagnostic classification models (Rupp et al., 2010). Many other constrained DCMs are special cases of the LCDM obtained by imposing different constraints on item parameters (Henson et al., 2008).

Throughout this study, we denote the index of an item as $j$, the index of a person as $i$, the index of a latent profile as $c$, and $K$ as the number of attributes. The latent profile $\alpha_c$ is constituted as a vector of attribute mastery status $\alpha_c = \{\alpha_{1,c}, \ldots, \alpha_{k,c}, \ldots, \alpha_{K,c}\}$. Thus, the LCDM yields a conditional probability of a correct response of person $i$ for item $j$ with an attribute profile $\alpha_c$:

$$P\left(X_{jc} = 1 \mid \alpha_c\right) = \frac{\exp\left[\lambda_{j,0} + \lambda_j \mathbf{h}\left(\alpha_c, \mathbf{q}_j\right)\right]}{1 + \exp\left[\lambda_{j,0} + \lambda_j \mathbf{h}\left(\alpha_c, \mathbf{q}_j\right)\right]}, \tag{1}$$

where $\alpha_c$ denotes the attribute mastery pattern of person $i$. Similar to linear logistic regression, $\lambda_{j,0}$ represents the item intercept parameter for item $j$, and $\lambda_j$ represents all main and interaction effects for item $j$. The Q-matrix is a $J$ by $K$ matrix with the row vector $\mathbf{q}_j = \left(q_{j1}, \ldots, q_{jk}, \ldots, q_{jK}\right)^T$ that contains the required attributes to answer item $j$ correctly (e.g., if attribute $k$ is relevant for item $j$, $q_{jk} = 1$). The mapping function $\mathbf{h}$ is used to specify the linear combination of attribute patterns $\alpha_c$ and Q-matrix entries $\mathbf{q}_j$.

$$\begin{aligned}\lambda_j \mathbf{h}\left(\alpha_c, \mathbf{q}_j\right) = &\sum_{k=1}^{K} \lambda_{j,1,(k)} \alpha_{k,c} q_{jk} \\ &+ \sum_{k=1}^{K-1} \sum_{k'=k+1}^{K} \lambda_{j,2,(k,k')} \alpha_{k,c} \alpha_{k',c} q_{jk} q_{jk'} + \cdots,\end{aligned} \tag{2}$$

where $k$ and $k'$ denote the index of interaction effects. For example, $\lambda_{j,2,(k,k')}$ denotes the regression coefficients of the second-order interaction between $k$th attribute and $k'$th attribute for item $j$.

In addition to the measurement model that connects observed item responses to the set of latent attributes, the structural part of the DCM models the dependencies between latent attributes, serving the role of the proficiency model within the BN framework. Hu and Templin (2020) demonstrated that a BN could be used for the purpose of model comparisons of nested DCMs. Thus, we use BNs as the fully saturated reference model in our model fit method. In the next section, BNs are introduced and compared to DCMs in terms of their parameterization and terminology.

## Bayesian Networks

To build the reference model of DCMs and the comparative goodness-of-fit statistic for BDCM, Bayesian Networks (BN; Almond et al., 2007, 2009; Pearl, 1988), also called Bayesian inference networks, was employed as a general version of DCM (Hu & Templin, 2020; Sinharay & Almond, 2007). BNs are a type of graphical model, whose *nodes* represent the variables and whose *edges* represent conditional dependencies between nodes.

Compared to DCMs or other latent class models, BNs allow any pattern of dependence consistent with an analyst-specified graph (Almond et al., 2009). A statistical likelihood of a directed acyclic graph could be represented as:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid pa(X_i)), \tag{3}$$

where $pa(X_i)$ denotes all parents of node $X_i$, and $P(X_i = x_i \mid pa(X_i))$ denotes the local probability distribution of variable $X_i$ conditional on the values of the node's parents, $pa(X_i)$. Considering *nodes* in BNs have the same statistical interpretation as observed or latent variables in other latent variable models, some latent variable models can be reparameterized as BNs (Hu & Templin, 2020), which we demonstrate later. In educational assessment, the use of such parametrization of BN allows researchers to estimate the direct dependency from node $X_2$ to node $X_1$, and thus reflects the dependency in the local distribution at attribute $X_1$.

In BN, conditional probability tables (CBT), $Pr(X_i \mid pa(X_i))$, can be constructed based on saturated multinomial logistic regression models (we will call it saturated BN in the rest of the paper). As Equation 3 shows, the probability distribution of a set of random variables $\mathbf{X} = (X_1, \ldots, X_n)$ can be recursively factorized as the conditional probability of each node conditioning on its parent nodes. Following the notations of Equation 3, let $X_j$, $j = \{1, \ldots, J\}$, denotes the observed categorical response of item $j$, and $\mathbf{X}_{-j}$ as a matrix with the size of N $\times$ $(J - 1)$, which represents the responses of $(J$ - 1) parent nodes. Then, for a fully connected BN, the conditional probability of node $j$ can be reparameterized as a logistic regression:

$$\Pr(X_j \mid \mathbf{X}_{-j}, \beta_{-j}) = \frac{\exp(\mathbf{X}_{-j}\beta_{-j})}{1 + \exp(\mathbf{X}_{-j}\beta_{-j})}, \tag{4}$$

Table 1
*Conversion between BN with LCDM*

| Marginal Probability | LCDM | BN |
|---|---|---|
| $P(X_1 = 1)$ | $\sum_{c=1}^{2} P(X_1 = 1 \mid \alpha_c)\pi_c$ | $P(X_1 = 1)$ |
| $P(X_2 = 1)$ | $\sum_{c=1}^{2} P(X_2 = 1 \mid \alpha_c)\pi_c$ | $P(X_2 = 1 \mid X_1)P(X_1)$ |
| $P(X_3 = 1)$ | $\sum_{c=1}^{2} P(X_3 = 1 \mid \alpha_c)\pi_c$ | $P(X_3 = 1 \mid X_1, X_2)P(X_2 \mid X_1)P(X_1)$ |

*Note*: $c$ = index of latent class; $\alpha_c$ = attribute profile for latent class $c$; $\pi_c$ = proportion of the latent class $c$; $P(X_1 = 1)$ = marginal probability of item 1's value being correct.

where $\Pr(X_j \mid \mathbf{X}_{-j}, \beta_{-\mathbf{j}})$ represents the conditional probability table regarding node $j$. Following the logistic regression form, $\mathbf{X}_{-j}$ denotes the design matrix of $(J-1)$ parent nodes. $\beta_{-j}$ denotes a $(J-1)$-dimensional vector of regression coefficients of dependent variable node $X_j$. We note that the model with a logistic link function used here is a special case of a BN with multinomial logistic regression in which all variables contain dichotomous values ($X_j \in \{0, 1\}$). Please refer to Rijmen (2008) for a detailed explanation of logistic BNs.

In this study, we focus on logistic BNs with up-to second-order interactions because they are related statistically to the LCDM. Figure S1 shows a BN-based diagram and an LDCM-based diagram for a 3-item test, respectively. Both models have three main effects (arrows) to be estimated. Table 1 presents the conversion of those effects between the LCDM and the BN. Besides the main effects, an LCDM has the same number of parameters as a saturated BN. To be specific, the LCDM has one intercept and one main effect for each item and one attribute mastery probability parameter (see Equation 1), which gives rise to $2 \times 3(items) + 1 = 7$ parameters. Similarly, a 3-item BN with saturated logistic regression has one intercept, two main effects and one two-way interaction effect ($1 + 2 + 1 = 4$ parameters) for item 3, one intercept and one main effect $1 + 1 = 2$ parameters for item 2, and one marginal probability parameter for item 1, which leads to 7 parameters estimated as well. However, as the number of attributes increases, there are fewer parameters to be estimated in the LCDM with up-to two-way interactions than BN, and thus BN can be applied as the reference model of LCDM.

In addition to the number of parameters estimated, the key distinction between the LCDM and general BNs is that the LCDM contains *person parameters* (the latent attributes for each individual). The LCDM marginal likelihood function marginalizes across the person parameters using a class membership proportion parameter, $\pi_c$, indicating the proportion of individuals within a given class (e.g., with a given attribute profile). Saturated BNs do not include such parameters of latent variables because BNs are typically modeled as conditional probabilities of *children* items conditioning on their *parent* items, $P(X_i = 1|pa(X_i))$, which are all observed variables. It should be noted that it is possible to include latent variable(s) in BNs. However, comparisons between latent BNs with DCMs are outside of the scope of this study. Please refer to Romeijn and Williamson (2018) for the identifiability of BN with latent variables.

In sum, the logistic BN with up-to second-order interactions is statistically equivalent to the LCDM since one's parameters can be transformed into the other's. In the proposed model selection approach, the BN with a saturated logistic regression is considered the reference model for the specified constrained LCDM. The comparison between these two models facilitates the construction of the proposed goodness-of-fit statistic. Essentially, this statistic quantifies the degree to which the posterior predictive distribution of the fit statistic of the specified model overlaps with the posterior predictive distribution of the fit statistic of the BN model, through a discrepancy measure. A larger discrepancy measure suggests a worse model fit.

### Limited-Information Statistics

In this study, limited-information statistics for both DCMs and BN are estimated to construct the discrepancy measure, which can further be used for model comparison between DCMs. Goodness-of-fit statistics for categorical response models can be categorized into two types in terms of the dimensions of contingency tables they use. The first type is called *full-information* statistics, which are the most commonly used statistics for model evaluation. Some examples of this type include Pearson's test statistics ($X^2$) and likelihood ratio test statistics ($G^2$). The second type is called *limited-information* indices, originally proposed by Maydeu-Olivares (2006), which make use of lower-order information from observed data contingency tables. Previous studies suggest that in sparse contingency tables, the empirical Type I error rates of the $X^2$ and $G^2$ test statistics do not match their expected rates under their asymptotic distributions (e.g., Maydeu-Olivares et al., 2018; Maydeu-Olivares, 2013; Ma, 2020). As the number of items $J$ increases, the contingency tables become more sparse (more empty cells), with many response patterns having no observations. A survey with 10 dichotomous items has $2^{10} = 1.024 \times 10^3$ response patterns, which exponentially increases to $2^{20} = 1.0480576 \times 10^7$ for 20 items. For most studies with reasonable sample sizes, the $p$-value of $X^2$ and $G^2$ of this 20-item survey is likely incorrect due to zero observations for most of the item response patterns. To address this issue, parameter bootstrapping or limited-information indices are used. However, compared to limited-information statistics, the parameter bootstrapping approach may be time-consuming and computationally burdensome when the models are complicated. By using only a small part of the information at hand, researchers can obtain a limited-information statistic that produces asymptotic $p$-values that are accurate even in large models and small samples. For example, Maydeu-Olivares and Joe (2005) suggested using $M_r$ statistics as in multidimensional IRT models, where $r$ denotes up to order $r$th marginal probabilities of the data tables. Previous research on limited-information statistics has focused on employing limited-information statistics in models estimated by maximum-likelihood methods. This study extends limited-information statistics in model checking of Bayesian analysis.

### Posterior Predictive Model Checking

In Bayesian analysis, the uncertainty of limited-information fit statistics can be obtained by summarizing the posterior predictive distribution. The model-checking procedures using posterior predictive distribution is called posterior predictive model

checking (PPMC; Gelman & Rubin, 1992). It allows researchers to check local or global model-data misfit for some aspects of an estimated model. However, one limitation of PPMC approaches is the potential uncertainty of the reference points of PPMC test statistics (i.e., item means, item pairwise correlations). For example, posterior predictive p-values (PPP-values), one popular method of checking model fit, can be interpreted as the likelihood of the statistics among potential predictive data sets implied by the hypothesized model with cutoffs set at.05 and.95. Extreme PPP-values suggest poor model fit. PPP-values reply on summary statistics obtained from the observed data as the reference but ignore the uncertainty of the statistics coming from sampling or measurement errors. Consequently, the precision of PPP-values could be worse with smaller sample sizes or higher missingness when the observed statistics do not accurately represent the population parameters.

To overcome this drawback, the empirical distributions instead of summary statistics (e.g., point estimates) can been applied as a reference (Matteucci & Mignani, 2020; Wu et al., 2014). The *empirical distribution*-based posterior predictive checking method seeks to quantify the distance between the realized and predictive distributions. However, a challenge arises when applying distance-based PPMC: Since this approach accounts for the uncertainty of observed data for each specified model, how does such uncertainty affect the comparison of alternative models? To answer this question, it is essential to employ the distance-based discrepancy measure to quantify the degree of overlap between posterior predictive distributions between alternative models. A variety of distance measures have been proposed in the prior studies, each with specific strengths and limitations. For example, Wu et al. (2014) proposed the relative entropy of PPMC (RE-PPMC) using Kullback-Leibler (KL) divergence. Matteucci and Mignani (2020) employed the Hollinger distance with PPMC to evaluate model fit in IRT. In a recent paper, Zhang et al. (2022) suggested that the Kolmogorov-Smirnov test (KS-test; Goodman, 1954) has advantages over other measures regarding accuracy and sensitivity, when used as the distance measure in model checking for factor analysis. The KS-test method for posterior predictive distributions demonstrated overall lower Type I error rates in Bayesian confirmatory factor analysis when detecting local misfit (Zhang et al., 2022), and thus it is a natural extension for DCMs.

In this study, we focus on applying the KS-test for our proposed model selection methods. The model-data fitting is defined as the divergence between the posterior predictive distributions of $M_2$ statistics from an alternative model (also called Model $H_0$) with those from a referenced BN model (also called Model $H_1$). The proposed method is detailed in the next section.

## KS-PP-M2

The proposed approach, KS-test-based Posterior Predictive Model Checking with M2 as the summary statistic (KS-PP-M2), employs a BN model as the reference model and calculates the distance between its goodness-of-fit distribution (e.g., M2 statistics) and the fit distributions for alternative DCMs. Furthermore, the distance measures between alternative DCMs will be compared, with lower values of distance measures suggesting better model fit. In this procedure, both goodness-of-fit

summary statistics and their uncertainty are considered in the model comparison. Specifically, the degree of distance between distributions of goodness-of-fit statistics for alternative DCMs is quantified using distance-based discrepancy measure—KS-test. The detailed procedure for calculating KS-PP-M2 is as follows.

First, the posterior distributions of parameters for the hypothesized DCM ($H_0$) and the referenced BN model ($H_1$) were calculated, respectively. The BN model mirrors the results of all two-way contingency tables with an identical number of parameters ($J + J(J-1)$), where $J$ denotes the number of items.

Second, posterior values of parameters are randomly sampled from the posterior distribution of parameters of the hypothesized model ($H_0$) as well as the saturated BN model ($H_1$) to generate predictive data sets: $\mathbf{Y}^{rep}|\theta_{H_0}$ and $\mathbf{Y}^{rep}|\theta_{H_1}$. $\mathbf{Y}^{rep}$ denotes a set of predictive data sets generated by posteriors draws of parameters $\theta$.

Third, according to the formula from Maydeu-Olivares and Joe (2005), each draw of the posterior predictive distribution of the $\boldsymbol{M}_2$ statistic with the hypothesized model ($H_0$) is derived using $\hat{\mathbf{r}}_2^{H0}$ and a weight matrix $\mathbf{W}_2$, as follows:

$$M_{2,i} = N\left(\hat{\mathbf{r}}_2^{H0}\right)' \mathbf{W}_2 \left(\hat{\mathbf{r}}_2^{H0}\right), \tag{5}$$

where $\hat{\mathbf{r}}_2^{H0} = \hat{\pi}_2^{H0} - \pi_2$ denotes up-to bivariate residuals between the posterior predictive data set ($\tilde{X}$) under model $H_0$ and the observed data set ($X$), $M_{2,i}$ is $i$th draw of posterior predictive limited-information statistic for the model $H_0$, and $\mathbf{W}_2 = \mathbf{I}$ is an identity matrix. Thus, by iteratively sampling thousands of draws, we can obtain a posterior predictive distribution of $M_2$ for Model $H_0$, which quantifies the goodness-of-fit of Model $H_0$ to the observed data and its variation. Similarly, the posterior predictive distribution of $M_2$ between the saturated model $H_1$ and observed data could be computed, which is denoted as $\mathbf{M}_2^{H1}$. It should be noted that in the original $M_2$ formula (Maydeu-Olivares & Joe, 2005), $\mathbf{W}$ has a more complicated statistical form for normalization.[1] However, in Bayesian PPMC, normalization is not important since the goal of the proposed method is not to derive an asymptotic distribution of $M_2$ but to compare models. Thus, the weight matrix $\mathbf{W}$ is fixed to the identity matrix $\mathbf{I}$.

Given posterior distributions of parameters and the $M_2$ formula, the posterior predictive distribution of $M_2^{H0}$ and $M_2^{H1}$, $p(\mathbf{M_2^{H0}}|\theta^{H0})$ and $p(\mathbf{M_2^{H1}}|\theta^{H1})$, are calculated using all values of parameters in the posterior distributions, respectively. $\theta^{H0}$ and $\theta^{H1}$ indicate the draws of posterior parameters of the hypothesized model and the reference model, respectively.

Finally, to quantify model-data fit, the misfit can be measured by the distance between these two distributions of $M_2$, $p(\mathbf{M_2^{H0}}|\theta^{H0})$ and $p(\mathbf{M_2^{H1}}|\theta^{H1})$, using Kolmogorov-Smirnov statistics. The KS statistics measure the distance between the posterior predictive distribution of $M_2$ from the hypothesized model and the posterior predictive distribution of $M_2$ from the saturated model as follows: $\mathbf{KS}(p(\mathbf{M_2^{H0}}|\theta^{H0}), p(\mathbf{M_2^{H1}}|\theta^{H1}))$, which can be denoted as KS-PP-M2. A higher value of KS-PP-M2 indicates a larger distance of $M_2$ estimates between the hypothesized model and the saturated model, which suggests a worse model fit. A significant KS-PP-M2 value indicates there is a significant difference between the hypothesized models with the saturated model in terms of posterior predictive M2 statistics.

Table 2
*Simulation Settings for the Monte Carlo Simulation*

| Factors<br>Structure | Data Generation | Data Analysis |
|---|---|---|
| Sample size ($N$)* | Unif(1,000, 2,000) | |
| Correlation of attributes ($\rho$) | .25,.5 | |
| Cutting scores ($\boldsymbol{\tau}$)* | Unif(–.5,.5) | |
| Test length ($J$) | 30 | 30 |
| Number of attributes ($K$) | 5 | 5 |
| **Item parameters** | | |
| Intercepts* | Unif(–3, 0) | |
| Main effects* | Unif(1.5, 2.5) | |
| Interaction effects* | Unif(.5, 1.5) | |
| **Models** | | |
| Fitted model | LCDM (Model 1) | BN and LCDM (Models 2, 3, 4, 5) |
| Q-matrix types | data-generation Q-matrix | 10%, 20% items underspecified<br>10%, 20% items incorrectly specified |

*Note*: Factors with asterisks are random factors, among which sample size ($N$) and correlation of attributes ($\rho$) are factors of interest, used for generating different conditions of simulated data. In total, there are 100 ($\mathbf{N}$) $\times$ 2 ($\boldsymbol{\rho}$) = 200 conditions in simulation study. For each condition, item parameters and cutting scores for skill scores are randomly drawn from the distributions above. The cognitive diagnosis index (CDI) will be computed for each condition to evaluate the measurement quality.

In addition, to evaluate the performance of the proposed method, true-positive rates (sensitivity) of the proposed method and information criterion were calculated:

$$\text{TPR} = \frac{\text{Number of selected model being true model}}{\text{Total number of model selection}} \tag{6}$$

## Simulation Study

To investigate the performance of the proposed KS-PP-M2 method, a Monte Carlo simulation study was conducted with various design factors. The simulation design in this study is partially borrowed from Ma (2020) and Liang et al. (2014).

### Design

**Data generation.** Table 2 presents the Monte Carlo simulation settings for data generation. Simulated data sets were generated based on the LCDM with four factors manipulated: (1) sample size ($N$); (2) attribute correlation ($\rho$); (3) cut scores for categorizing continuous attribute scores ($\tau$); and (4) item parameters including intercepts, main effects, and interaction effects ($\boldsymbol{\lambda}$s). The specific data-generation process is as follows. First, 200 conditions with different sample size ($N$) were randomly sampled from $U(1, 000, 2, 000)$. The random sampling of sample size makes the generated data set more realistic and allows us to examine the tendency of the proposed indices as sample size increases. Both test length ($J$) and the number of attributes ($K$) are fixed to 30 and 5, respectively, to represent a middle-size data set.

Regarding the attribute correlations, low ($\rho = .25$) and moderate ($\rho = .5$) attribute correlation coefficients were used to represent low attribute relation and moderate attribute relation (Kunina-Habenicht et al., 2012). In other words, a correlation matrix ($\mathbf{\Phi}$) of five skills is a matrix with its diagonal elements as "1"s, and nondiagonal elements as.25 or.5, respectively. Given the attribute correlation matrix $\mathbf{\Phi}$, sample size $N$, and number of attributes $K$, then a latent score matrix $\Theta$ could be generated from a multivariate normal distribution $MVN(\mathbf{0}, \mathbf{\Phi})$. To create categorical latent scores, a vector of attribute thresholds $\boldsymbol{\tau}$ was used to categorize continuous latent scores into binary responses. Each element of $\boldsymbol{\tau}$ was randomly sampled from a uniform distribution $U(-.5, .5)$. This generation process is consistent with the practical scenario that each measured skill has its own thresholds in real-world settings. If the latent score of attribute $x_k$ for person $i$ was larger than the threshold $\tau_k$ ($x_{k,i} > \tau_k$), the mastery status of person $i$ would be 1, $\alpha_{k,i} = 1$; otherwise, $\alpha_{k,i} = 0$. Thus, mastery profiles for all samples $\alpha$ with size $N \times K$ were generated. In total, 200 conditions (100 Ns $\times$ 2 $\rho$s) were generated for the simulation study. Given the estimation complexity of analysis models, each condition was replicated only once. However, since we randomly sampled levels of sample size and item parameters, the simulation study could still be generalized to varied real scenarios to some degree.

Finally, item parameters were randomly generated similar to Templin and Hoffman (2013)'s LCDM item parameter estimates for the Examination for the Certificate of Proficiency in English (ECPE) data. To be specific, the item intercept of item $i$, $\lambda_{i,0}$ was sampled from a uniform distribution $U(-1, 1)$. The main effects of attribute $k$ on item $i$, $\boldsymbol{\lambda}_{i,1,(k)}$, were randomly sampled from a uniform distribution $U(1.5, 2.5)$. The values of two-way interaction effects of attribute $k$ and $k'$, $\lambda_{i,2,(k,k')}$, were sampled from a uniform distribution $U(.5, 1.5)$, and the three-way interaction effect of three attributes, $\lambda_{i,3,(k,k',k'')}$, was randomly sample from a uniform distribution $U(.5, 1.5)$. All data sets were simulated based on parameters $\lambda$ and mastery status $\alpha$ using R ver. 4.2.1 (R Core Team, 2013).

**Cognitive diagnostic index.** To examine how the uncertainty of the data-generation process and sampling error influence the performance of the KS-PP-M2 method, the *Cognitive Diagnostic Index* (CDI Henson & Douglas, 2005) was calculated for each condition. The CDI is an alternative to Fisher information in diagnostic classification models. It can be defined as the amount of information the observed data contain regarding the distributions of latent variables. Specifically, in DCMs, CDI measures an item's overall discrimination power and serves as a measure of an item's capacity to accurately classify the examinees' true status (Kuo et al., 2016; Rupp et al., 2010). Item-level $CDI_j$ can be summed over $J$ items to form a test-level CDI, formulated as $CDI = \sum_{j=1}^{J} CDI_j$. Higher test-level CDI suggests a test possesses more discrimination power to identify the examinees' unobserved skill patterns, and vice versa.

## Analysis

As shown in Table 3, to investigate the performance of the proposed model selection approach in choosing the correct Q-matrix that was used for data generation, six models were analyzed in total for each condition including one reference model (sat-

Table 3
*Analysis Models in the Simulation Study*

| Model | Type | Q-Matrix Design |
|---|---|---|
| Model 0 | Saturated BN | Each item measured by all previous items |
| Model 1 | Data generation | true Q-matrix |
| Model 2 | Underspecified | 10% items underspecified |
| Model 3 | Underspecified | 20% items underspecified |
| Model 4 | Incorrectly specified | 10% items misspecified |
| Model 5 | Incorrectly specified | 20% items misspecified |

urated BN), one data-generation model, and four alternative models (Models 2-5). These alternatives consisted of four permutations of the Q-matrix, with misspecification of 10% and 20% of items, categorized into two types: (1) 10% and 20% items incorrectly specifying attributes and (2) 10% and 20% items specifying fewer attributes than the correct Q-matrix. Thus, in total 200 (conditions) $\times$ 6 (models) = 1,200 Bayesian models were estimated in this simulation study.

For each condition, the saturated BN (Model 0) was employed as the reference model to compute the proposed statistic, KS-PP-M2. Figure S2 presented the directed acyclic graph (DAG) for the 30-item saturated BN model (arrows are hidden for clarity). That said, item responses in saturated BN were predicted by other items with the link function as a logistic regression. The sign of the regression coefficients in the logistic regression represented whether there were positive or negative correlations between the target item (dependent variable) and its previous items (independent variables). For example, item 3's responses were regressed on item 1's and item 2's responses.

Figure S3 shows Q-matrices for Models 1-5 with white or grey cells as attributes required for each item. Items with misspecified entries were colored with grey cells and dots in the Q-matrix. To be specific, Model 1 is the data-generation model (see the first plot of Figure S3). Model 2 is an incorrect model with 10% of items underspecifying attribute 1 (see the second plot of Figure S3). Model 3 is an underspecifying model with 20% of items underspecifying attribute 1 (see the third plot in Figure S3). Model 4 is an incorrectly specified model with 10% of items misspecifying their attributes (see the forth plot in Figure S3). Model 5 is an incorrectly specified model with 20% of items misspecifying attributes (see the last plot in Figure S3). All models were estimated using *blatent* package (Templin, 2023) in R.

To examine the sensitivity of KS-PP-M2, the average posterior predictive M2 among five analysis models and the reference model (BN model) across all conditions was reported with the hypotheses that (1) the reference model (Model 0) and the data-generation model (Model 1) have the lowest average posterior predictive M2 across all conditions; (2) the analysis models with less misspecified entries in Q-matrix (Models 2 and 4) have lower average posterior predictive M2 than those with more misspecified entries (Models 3 and 5) across all conditions. Furthermore, to compare the accuracy of the proposed model-selection approach, summary statistics and power of KS-PP-M2 and other information criteria (DIC, WAIC, AIC, and BIC)

were computed for all models to examine whether the data-generation model was selected based on the lowest values. Finally, the relationship between KS-PP-M2 and CDI was examined to see if KS-PP-M2 would not be influenced by test information (or data-generation uncertainty).

## Results

**Posterior predictive M2.** The preliminary analysis suggested that 4,000 iterations in total with the first 1,000 iterations discarded as a burn-in phase were sufficient to achieve model convergence of the BN model and each LCDM. Across all conditions, the BN model and five LCDMs converged when the Markov chain Monte Carlo (MCMC) algorithm sampled 4,000 iterations in total including the first 1,000 iterations discarded as burn-ins (Gelman-Rubin convergence diagnostic PSRF $< 1.1$).

Figure S4 shows the trend of average posterior predictive M2 (*PP-M2*) over 100 conditions along with sample size (*N*) for all models. The results suggested that for both attribute correlations, average posterior predictive M2 monotonically increased as sample sizes increased in general (correlation between *N* and PP-M2 ranged from.951 to.988 across all conditions). Furthermore, the BN models had the overall lowest average posterior predictive M2 followed by Model 1 (the data-generation model). In contrast, Model 3 (20% items specify fewer attributes) had the highest average posterior predictive M2 than other models in all conditions. Model 2 (10% items underspecified) had lower average PP-M2 than Model 3 (20% items underspecified), and Model 4 (10% items incorrectly specified attributes) tended to have lower PP-M2 than Model 5 (20% items incorrectly specified attributes). Even though the results of average PP-M2 values offered insights into Q-matrix misspecification, they were easily influenced by sample size and failed to account for the uncertainty of M2. Thus, to enhance the accuracy in model selection, we examined KS-PP-M2 as a goodness-of-fit index for a more accurate comparison of models.

**KS-PP-M2.** Comparing KS-PP-M2 to other information criteria, the true-positive rates (TPR) for all fit measures were 100% (in other words, Type I error rates were 0%). Figure S6 presents the means and ranges among KS-PP-M2 and other normalized information criteria for Models 1-5 across all replications. It suggested that the distance of model fitting between Model 1 (the data-generation model) and other models (models with misspecified Q-matrices) was larger for KS-PP-M2 than for information criteria. In addition, the variation of KS-PP-M2 across simulated data sets for all analysis models was much smaller than that for information criteria.

Table 4 reported summary statistics of the KS-PP-M2 and information criteria (IC) for analysis models by various factor correlations. Consistent with the results of information criteria, the data-generation models (Model 1) for both levels of factor correlations had average lower posterior predictive M2 with $M_{\rho=.25}(SD)$ as.18(.04) and $M_{\rho=.50}(SD)$ as.33(.11), which suggested Model 1 having the best model fit than other models. Comparing average values of fit indices across models, all fit indices were able to detect the misspecification of Q-matrix indicated by higher values of Models 2-5. For the relationship between factor correlation and model fitting, all models have worse model fits suggested by larger values of fit indices as the fac-

Table 4
*Summary Statistics of KS-PP-M2 and Information Criteria*

| Models | ρ | KS-PP-M2 | DIC | WAIC | AIC | BIC |
|---|---|---|---|---|---|---|
| Model 1 | .25 | .18(.04) | 44300.74(8815.56) | 44302.61(8816.89) | 44387.28(8812.27) | 45291.19(8845.77) |
| Model 2 | .25 | .66(.14) | 44868.57(8930.74) | 44871.57(8932.19) | 44960.81(8927.29) | 45864.73(8960.78) |
| Model 3 | .25 | .83(.1) | 45399.78(9027.92) | 45405.79(9029.71) | 45500.1(9024.41) | 46404.01(9057.92) |
| Model 4 | .25 | .6(.13) | 44842.11(8905.15) | 44844.96(8906.51) | 44932.18(8901.61) | 45804.56(8933.96) |
| Model 5 | .25 | .7(.12) | 45109.82(8952.04) | 45112.93(8953.04) | 45188(8948.59) | 45997.32(8978.59) |
| Model 1 | .50 | .33(.11) | 45606.57(9219.33) | 45612.02(9220.66) | 45699.17(9215.02) | 46613.61(9249.99) |
| Model 2 | .50 | .75(.13) | 46152.25(9340.82) | 46159.1(9342.41) | 46250.13(9336.67) | 47164.57(9371.63) |
| Model 3 | .50 | .86(.1) | 46658.77(9448.03) | 46668.42(9449.84) | 46762.69(9444.13) | 47677.12(9479.09) |
| Model 4 | .50 | .75(.13) | 46172.31(9325.58) | 46178.05(9326.99) | 46267.74(9322.09) | 47150.28(9355.84) |
| Model 5 | .50 | .8(.1) | 46402.79(9372.64) | 46408.55(9373.84) | 46485.35(9369.41) | 47304.08(9400.72) |

*Note*: The model selection indices which had best model fit. Model 1: the data-generation model; Model 2: 10% underspecified items; Model 3: 20% underspecified items; Model 4: 10% items incorrectly specified attributes; Model 5: 20% items incorrectly specified attributes.

tor correlation increased. However, controlling the factor correlation, KS-PP-M2 favored models with incorrectly specified attributes (Models 4 and 5) more than those with underspecified attributes (Models 2 and 3), while other fit indices favored models with underspecified attributes (Models 2 and 3) more than those with incorrectly specified attributes (Models 4 and 5).

**The relation between CDI with KS-PP-M2.** To examine whether test information affects the proposed KS-PP-M2, test-level CDIs were calculated for all simulated data across all conditions with Mean(SD) as 85.3(4.79) for $\rho = .25$ and 85.5(4.35) for $\rho = .50$. Figure S5 displays the fitted regression lines of KS-PP-M2 for different models across various CDIs. It shows a weak association between CDI and KS-PP-M2 since KS-PP-M2 for each model fluctuated randomly along different levels of CDI. Moreover, Table S1 shows the results of the regression analysis on the KS-PP-M2 with CDI as a predictor, conditional on the model types, in which *CDIc* denotes the mean-centered CDI values. The results suggested that the main effect of *CDIc* on PP-KS-M2 was.00 with $p = .82$, implying that the variation of test information did not have a significant effect on KS-PP-M2 values ($\beta^{CDIc} = .00$, $p = .82$) controlling for model types. The interaction effects of *CDIc* and models (CDIc: Model 2, CDIc: Model 3, CDIc: Model 4, and CDIc: Model 5) exhibited nonsignificant effects ($\beta^{CDIc:Model2} = .00$, $p = .93$; $\beta^{CDIc:Model3} = .00$, $p = .59$; $\beta^{CDIc:Model4} = .00$, $p = .97$; $\beta^{CDIc:Model4} = .00$, $p = .71$), which suggested that the slopes of test information were not significantly different between data-generation model (Model 1) and incorrect models (Models 2-5). Since the values of KS-PP-M2 represented the goodness-of-fit of different models, the regression analysis revealed that CDI did not significantly affect the performance of KS-PP-M2 in choosing the correct Q-matrix in the simulation settings.

## Empirical Study

### Design

In the empirical study, the KS-PP-M2 approach was used for the model selection with the *Examination for Certificate of Proficiency in English* (ECPE) data (Templin and Bradshaw, 2014). The main purpose of this empirical study is to demonstrate an application of employing KS-PP-M2 for model comparison in a real-world scenario. As characteristics of data sets differ across analyses, this study should not be considered a comprehensive study of the usefulness of the KS-PP-M2 method for various conditions, but as an illustration of the procedures for assessing model fit based on KS-PP-M2. The ECPE data has been well investigated with various model structures by prior studies (e.g., Templin & Hoffman, 2013; Templin and Bradshaw, 2014). For instance, Templin and Hoffman (2013) estimated the LCDM with the data and found that the three-dimensional model fitted better than other models according to AIC. The correlation among the three attributes ranges from.79 to.81, suggesting that the skills are, though highly correlated, differentiable from each other to some degree. Templin & Bradshaw (2014) further found that an attribute hierarchical structure was present: Examinees must master lexical rules before mastering cohesive rules, and must master cohesive rules before mastering morphosyntactic rules.

## Data

The ECPE is a test developed and scored by the English Language Institute at the University of Michigan. The test measures advanced English skills in examinees whose primary language is not English and is administered internationally once a year between November and April, depending on the location (Templin & Hoffman, 2013).

Compared to the original data, the example data is restricted to the grammar section, including 28 multiple-choice questions, and 2,922 test takers from a single year's administration. According to prior literature (e.g., Templin & Hoffman, 2013), the ECPE contains three skills: (1) morphosyntactic rules, (2) cohesive rules, and (3) lexical rules (see Buck & Tatsuoka, 1998; Henson et al., 2008). Please refer to Templin and Hoffman (2013) for more detailed information on the example data, such as the observed scores and example items. The ECPE data is publicly available in the *CDM* package in R (Ravand & Robitzsch, 2015).

## Analysis

In the empirical study, three analysis models and one saturated model (BN) will be presented: (1) *Model 1*: the three-dimensional model used by Templin and Hoffman (2013); (2) *Model 2*: a two-dimensional model with an arbitrary Q-matrix, (3) *Model 3*: a unidimensional model with a one-column Q-matrix, and (4) the saturated BN model. The aim of applying Models 1, 2, and 3 to ECPE is to examine whether dimensionality affects the performance of model fit indices and to check whether the results of KS-PP-M2 were consistent with the previous study (Templin & Hoffman, 2013).

Please refer to Templin and Hoffman (2013) for the specification of Model 1. Model 2 was specified as follows: the first 14 items measure attribute 1 ($\alpha_1$) and the last 14 items measure attribute 2 ($\alpha_2$). Model 3 was specified as all 28 items measuring attribute 1 ($\alpha_1$). To examine the proposed method, we performed graphical checking of the posterior predictive distributions of $M_2$, and calculated the KS-PP-M2, WAIC, and DIC for the analysis models. Specifically, we can examine model fitting by visually inspecting the posterior predictive distribution of $M_2$ for the Bayesian Network model and the specified models. The one closer to zero suggests a better model fit. Alternatively, lower values of KS-PP-M2, WAIC, and DIC suggest better model fit. R codes for data analysis have been shared on the Open Science Framework (OSF) and can be accessed via https://osf.io/8HAVF/.

## Results

According to convergence diagnostics, all models converged after 2,000 iterations with the first 1,000 iterations treated as burn-ins. The maximum PSRF across three chains $\leq 1.01$ suggested that all models converged.

Figure 1 shows the posterior predictive distribution of $M_2$ values for the saturated model (Bayesian Network) and three analysis models (Models 1-3). The results suggested that the Bayesian Network model (solid line; $M = 1.31$, $SD = .315$) had the lowest average posterior predictive $M_2$ values followed by Model 3—unidimensional model (dotdash line; $M = 2.22$, $SD = .355$), and Model 1—three-

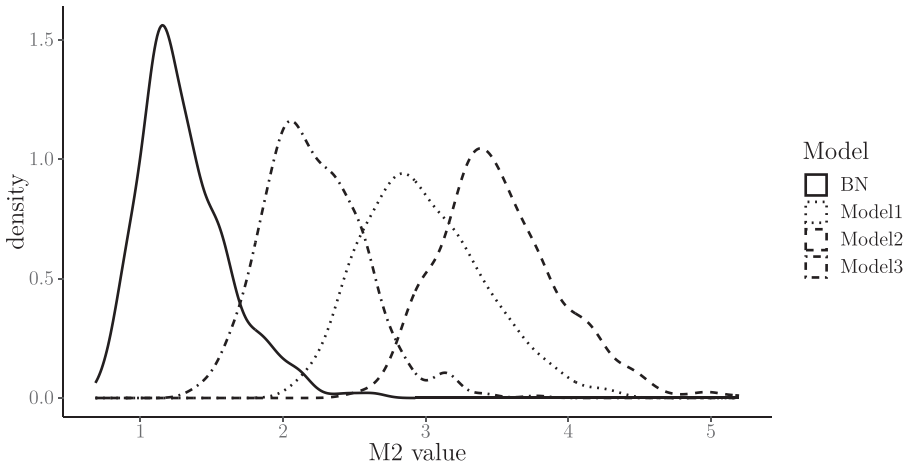


*Figure 1.* Posterior predictive distribution of $M_2$ for three analysis models for the ECPE data in the empirical study.
*Note*: The solid line is the Bayesian Network model; the dotted line is Model 1 (three-dimensional model); the dashed line is Model 2 (two-dimensional model); the dotdashed line is Model 3 (unidimensional model). 500 posterior predictive values are sampled for each model.

Table 5
*KS-PP-M2 and Information Criteria for Analysis Models*

| | KS-PP-M2 | DIC | WAIC | AIC | BIC |
|---|---|---|---|---|---|
| Model 1 | .980*** | 86144.46 | 86162.92 | 86173 | 86663.37 |
| Model 2 | .994*** | 86485.48 | 86486.92 | 86486.49 | 86845.29 |
| Model 3 | .840*** | 86060.35 | 86061.67 | 86063.37 | 86410.21 |

*Note*: ***$p < .001$.

dimensional model (dotted line; $M = 2.99$, $SD = .424$). The worst fitted model was Model 2—two-dimensional model (dashed line; $M = 3.51$, $SD = .425$). Although none of the three analysis models highly overlaps with the Bayesian Network model regarding posterior predictive $M_2$ values, Model 3 shows a shorter distance with the Bayesian Network model than other models.

Table 5 presents the results of KS-PP-M2, DIC, WAIC, AIC, and BIC for Models 1-3. Comparing the fit among these models, Models 1 and 2 yielded higher fit values indicating less representative of the "true" structure of the ECPE data, but the unidimensional attribute structure measured by Model 3 fitted better to the ECPE data, suggesting that the model fit indices were not inflated by model complexity with additional dimensions. Inconsistent with the findings from Templin and Hoffman (2013), all model selection methods suggested that Model 3 (unidimensional model) exhibited the best model fit than other models, followed by Model 1 (three-dimensional model). The results of KS-PP-M2 aligned with the results of information criteria (AIC, BIC, WAIC, and DIC).

It is worth noting that DIC, WAIC, AIC, and BIC as shown in Table 5 did not provide any information regarding the model goodness-of-fit but for model comparison. Instead, KS-PP-M2 of Models 1-3 rejected the null hypothesis that the specified model had an equivalent fit to the saturated BN model, suggesting that these analysis models and the BN model were not likely to have the same shape of distributions of posterior predictive $M_2$ statistics.

## Discussion

This study proposes a novel model fit statistic for Bayesian diagnostic classification modeling, namely KS-PP-MC, based on Bayesian Network, posterior predictive model checking, KS-test, and the $M_2$ statistic. This fully Bayesian statistic facilitates visual inspection for model comparison. Moreover, it accounts for the uncertainty information inherent in the posterior predictive distribution by quantifying the distance between the posterior predictive $M_2$ of the analysis model and that of the saturated model. A Monte Carlo simulation study was conducted to explore the impact of various design factors (factor correlation, sample size, and Q-matrix misspecification) on the performance of KS-PP-M2. Additionally, this study examined the sensitivity of KS-PP-M2 in detecting different levels of Q-matrix misspecification compared to commonly used information criteria. An empirical study utilizing ECPE data demonstrated the application of KS-PP-M2 to real data and its ability to detect misspecifications in dimensionality. Overall, KS-PP-M2 exhibited stable power in detecting model misspecifications arising from the Q-matrix and dimensionality under diverse conditions. The interpretations of results, practical implications, and strengths and limitations of KS-PP-M2 are discussed below according to findings from both the simulation and empirical studies.

### RQ1: Is KS-PP-M2 an Appropriate Approach for Detecting the Model-Data Misfit with Various Degrees of Q-Matrix Misspecification?

As depicted in Figure S4, the saturated model (BN) and the data generation model (Model 1) exhibited lower posterior predictive $M_2$ values than the LCDMs with incorrectly specified Q-matrices (Models 2-5). This indicates that the distributions of posterior predictive $M_2$ could serve as a criterion for selecting the correct Q-matrix. Further analysis, as shown in Table 4 and Figure S6, revealed that models with misspecified Q-matrices consistently had higher average KS-PP-M2 values. Moreover, increased levels of Q-matrix misspecification were associated with higher KS-PP-M2 values. For instance, for the factor correlation as.25, the average KS-PP-M2 value for Model 2, with 10% of items underspecified, was.66, whereas for Model 3, with 20% of items underspecified, it was.83. Similarly, the average KS-PP-M2 value for Model 3, with 10% of items incorrectly specified, was.60, whereas for Model 3, with 20% of items incorrectly misspecified, it was.70. Compared to weak factor correlation, medium factor correlation ($\rho = .50$) yielded a worse model fit for all analysis models.

Similarly, evidence from the empirical study, as illustrated in Figure 1, supported the proposed method's ability to accurately identify the unidimensional model as the

best-fitting model. This finding is aligned with the findings from information criteria but deviates from previous literature (Templin & Hoffman, 2013).

In summary, the PP-M2 method effectively distinguished between correct and incorrect models across varying levels of Q-matrix misspecification. Building upon PP-M2, the KS-PP-M2 method refines the approach by taking into account the uncertainty of posterior distributions, offering a more effective means for selecting the appropriate Q-matrix.

### RQ2: Compared to Other Information Criteria, Does the Proposed Approach Have Better Performance When Selecting the Correct Model?

PP-M2 can be used to detect model misfit but was strongly influenced by sample size. By performing the KS-test on PP-M2, the KS-PP-M2 approach demonstrated at least equivalent power of choosing the correct Q-matrix compared to information criteria according to true-positive rates. Moreover, for both levels of factor correlations, as shown in Figure S6, KS-PP-M2 has relatively lower variation and higher sensitivity to model misspecification than other normalized information criteria across various replications. In the empirical study, both KS-PP-M2 and information criteria selected the unidimensional model as the best-fitting model, suggesting that KS-PP-M2 did not overestimate the model fit due to the increased number of parameters associated with multidimensionality.

### RQ3: How Does the Information of Observed Data Affect the Performance of the Proposed Method?

As shown in Table S1 and Figure S5, test information indices by the cognitive diagnostic index have no statistically significant effects on either KS-PP-M2 of the data-generation models or the difference of KS-PP-M2 between the data-generation model and the model with a misspecified Q-matrix. That is, in all simulated conditions, the uncertainty of classifying respondents or the information contained in data did not affect the power of the proposed method in selecting the correct model.

### Advantages

In summary, the proposed KS-PP-M2 method has the following advantages. First, the KS-PP-M2 method can compare alternative models to select the best-fitting model among competitors taking the uncertainty of posterior prediction distributions into account. The saturated BN model as the reference model contains the uncertainty coming from data (i.e., sampling error) and may be more theoretically valid than a point estimate derived from observed statistics that is assumed to be true. In the future, it would be helpful to derive cut-off scores for KS-PP-M2 as an absolute model fit method for the model goodness-of-fit evaluation. A comprehensive power analysis of KS-PP-M2 aiming to determine cutoffs under various conditions may be a valuable direction in the future.

Second, the KS-PP-M2 approach takes the uncertainty of observed data into account when quantifying model-data fit, which increases the precision of the model-data checking process. Ignoring the uncertainty of fit statistics potentially leads to the low accuracy problem of model-data fit indices under the conditions of small

sample sizes or high missing rate (e.g., Asparouhov & Muthén, 2020; Winter & Depaoli, 2022). For example, Asparouhov and Muthén (2020) found that Bayesian approximate fit indices failed for small sample sizes in some situations. The previous study relies on rules of thumb or confidence limits of the fit indices of the maximum-likelihood estimator. In this study, similar to Asparouhov and Muthén (2020)'s model fit evaluation in structural equation modeling, the proposed approach using the Bayesian Network model can be easily generalized to missing data situations by generating replicated data under the Bayesian Network model so that they have the same amount of missing data as the real data set.

Third, the KS-PP-M2 does not require likelihood-based statistics for computation, thus it could be generalized to approximation modeling such as *variational Bayes (VB)* with intractable likelihoods (e.g., Tran et al., 2016) or neural network models (e.g., Lenzi et al., 2021). This feature could be very useful when conducting large-scale data analyses and complex modeling, in which the likelihood function is either difficult to calculate or intractable. Information criteria such as AIC, BIC, DIC, and related criteria have different target quantities. For example, the motivation for the use of the BIC is comparing probabilities for each of the models under consideration. Their performance varies depending on their assumptions, analysis models, and observed data sets. In comparison with information criteria, the KS-PP-M2 method does not assume each model under consideration is the true model. Instead, its motivation is to quantify the degree to which the prediction accuracy of each model under consideration matches a saturated model, whatever the true data-generation process is.

Last, the proposed approach takes advantage of limited-information statistics, which have been shown to perform better in highly sparse data using both frequentist estimation (e.g., Maydeu-Olivares & Joe, 2008; Ma, 2020; Maydeu-Olivares & Joe, 2005; Ranger & Kuhn, 2012) and Bayesian estimation (e.g., Maydeu-Olivares et al., 2018; Sinharay, 2005; Stone & Zhang, 2003). Further research is needed to test the performance of KS-PP-M2 under high amounts of sparseness.

## Future Directions and Limitations

Some disadvantages exist when applying the proposed KS-PP-M2 method in real settings. First, the limited-information $M_2$ statistics used in this study do not contain a weight matrix for residuals, which makes the values of $M_2$ not follow an asymptotic distribution. Compared to the original $M_2$ statistics developed by Maydeu-Olivares and Joe (2005), posterior predictive $M_2$ values used in this study are not easy to interpret without the KS test. In the future, more investigation is needed to compare the performance of various limited-information discrepancy measures in Bayesian estimation.

Second, the KS-PP-M2 method requires estimating a saturated BN model beforehand, which could be computationally burdensome when the analyzed model contains a large number of parameters or when the data have a large sample size. One potential solution is replacing Bayesian estimation with variational Bayes (VB) estimation for the Bayesian Network model. For instance, Yamaguchi and Okada (2020) demonstrated that the computational time required for the VB inference for the DINA

model was about 1.28 seconds using a laptop with a 3.1 GHz processor and 16 GB of memory. In contrast, MCMC estimation required about 600 seconds for one chain to converge in the same computational environment.

Third, recently the leave-one-out cross-validation (LOO-CV) based model fitting technique is getting more attention (Kuh et al., 2022), but the present study does not include these relative fit indices for model comparison except for WAIC. This is because, to our knowledge, the performance of LOO with Pareto smoothed importance sampling (LOO-PSIS; Vehtari et al., 2017) in Bayesian diagnostic classification modeling has not been well investigated yet. In addition, currently only *Stan* (Stan Development Team, 2020) supports this kind of fit indices. Research on comparing the performance of LOO-PSIS, IC and KS-PP-M2 in BDCM selection is needed in the future.

Last, although the proposed KS-PP-M2 method is relatively more stable across various conditions than other model fit indices, the proposed method is not different from other methods regarding true-positive rates. The possible explanation is the model checking methods have the ceiling effect with the current relatively straightforward simulation setting (i.e., 30-item test with five attributes measured). One possible direction could be testing more complicated diagnostic classification models with more design factors, such as higher attribute correlations, more latent attributes, and longer test lengths.

## Note

[1]$\mathbf{W} = \Xi_2^{-1} - \Xi_2^{-1}\Delta_2\left(\Delta_2'\boldsymbol{\Xi}_2^{-1}\Delta_2\right)^{-1}\boldsymbol{\Delta}_2'\boldsymbol{\Xi}_2^{-1}$. $\Xi_2$ is the sample varaince-covaraince matrix of up-to second-order marginal probabilities. $\Delta$ denotes derivation of likelihood function along parameters.

## References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement*, *44*(4), 341–359.

Almond, R. G., Mulder, J., Hemat, L. A., & Yan, D. (2009). Bayesian network models for local dependence among observable outcome variables. *Journal of Educational and Behavioral Statistics*, *34*(4), 491–521.

Asparouhov, T., & Muthén, B. (2020). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *0*(0), 1–14.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119–157.

DeCarlo, L. T. (2012). Recognizing uncertainty in the q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, *36*(6), 447–468.

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

Goodman, L. A. (1954). Kolmogorov-Smirnov tests for psychological research. *Psychological Bulletin*, *51*, 160–168.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*(4), 262–277.

Henson, R., Templin, J., & Willse, J. (2008). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191.

Hu, B., & Templin, J. (2020) Using diagnostic classification models to validate attribute hierarchies and evaluate model fit in Bayesian networks. *Multivariate Behavioral Research*, *55*(2), 300–311.

Kuh, S., Kennedy, L., Chen, Q., & Gelman, A. (2022). Using leave-one-out cross-validation (LOO) in a multilevel regression and poststratification (MRP) workflow: A cautionary tale. *Statistics in Medicine*, *43*(5), 953–982.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59–81.

Kuo, B.-C., Pai, H.-S., & de la Torre, J. (2016). Modified cognitive diagnostic index and modified attribute-level discrimination index for test construction. *Applied Psychological Measurement*, *40*(5), 315–330.

Lei, P.-W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and q-matrices. *Applied Psychological Measurement*, *40*(6), 405–417.

Lenzi, A., Bessac, J., Rudi, J., & Stein, M. L. (2021). Neural networks for parameter estimation in intractable models. arXiv:2107.14346.

Liang, T., Wells, C. S., & Hambleton, R. K. (2014). An assessment of the nonparametric approach for evaluating the fit of item response models. *Journal of Educational Measurement*, *51*(1), 1–17.

Ma, W. (2020). Evaluating the fit of sequential g-DINA model using limited-information measures. *Applied Psychological Measurement*, *44*(3), 167–181.

Mark, H., Cai, L., Scott, M., & Zhen, L. (2014). Limited-information goodness-of-ft testing of diagnostic classification item response theory models. *The British Journal of Mathematical and Statistical Psychology*, *69*(3), 225–252.

Matteucci, M., & Mignani, S. (2020). The hellinger distance within posterior predictive assessment for investigating multidimensionality in IRT models. *Multivariate Behavioral Research*, *0*(0), 1–22.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research & Perspective*, *11*(3), 71–101.

Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika*, *71*(1), 57–77.

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in $2^n$ contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2008). An overview of limited information goodness-of-fit testing in multidimensional contingency tables. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (253–262). Universal Academy Press.

Maydeu-Olivares, A., Shi, D., & Rosseel, Y. (2018). Assessing fit in structural equation models: A Monte-Carlo evaluation of RMSEA versus SRMR confidence intervals and tests of close fit. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 389–402.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ranger, J., & Kuhn, J.-T. (2012). Assessing fit of item response models using the information matrix test. *Journal of Educational Measurement*, *49*(3), 247–268.

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using r. University of Massachusetts Amherst.

Rijmen, F. (2008). Bayesian networks with a logistic regression model for the conditional probabilities. *International Journal of Approximate Reasoning*, *48*(2), 659–666.

Romeijn, J.-W., & Williamson, J. (2018). Intervention and identifiability in latent variable modelling. *Minds and Machines*, *28*(2), 243–264.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Sen, S., & Bradshaw, L. (2017). Comparison of relative fit indices for diagnostic model selection. *Applied Psychological Measurement*, *41*(6), 422–438.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*(4), 375–394.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement*, *67*(2), 239–257.

Stan Development Team (2020). RStan: The R interface to Stan. R package version 2.21.2.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*(4), 331–352.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354.

Templin, J. (2023). blatent: Bayesian Latent Variable Models in R.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317–339.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37–50. https://onlinelibrary.wiley.com/doi/pdf/10.1111/emip.12010.

Thompson, W. J. (2020). Bayesian psychometrics for diagnostic assessments: A proof of concept. Technical Report.

Tran, M.-N., Nott, D. J., & Kohn, R. (2016). Variational Bayes with intractable likelihood. arXiv:1503.08621.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

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.

Winter, S. D., & Depaoli, S. (2022). Performance of model fit and selection indices for Bayesian structural equation modeling with missing data. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(4), 531–549.

Wu, H., Yuen, K.-V., & Leung, S.-O. (2014). A novel relative entropy-posterior predictive model checking approach with limited information statistics for latent trait models in sparse 2k contingency tables. *Computational Statistics & Data Analysis*, *79*, 261–276.

Yamaguchi, K., & Okada, K. (2020). Variational Bayes inference for the DINA model. *Journal of Educational and Behavioral Statistics*, *45*(5), 569–597.

Zhang, J., Templin, J., & Mintz, C. E. (2022). A model comparison approach to posterior predictive model checks in Bayesian confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 339–349.

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

## Authors

JIHONG ZHANG is an Assistant Professor of Educational Statistics and Research Methods at the University of Arkansas, 751 W Maple Street, Fayetteville, AR 72701; jzhang@uark.edu. His primary research interests include psychometric networks, diagnostic classification models, and Bayesian psychometrics.

JONATHAN TEMPLIN is E.F. Lindquist Chaired Professor of Measurement and Statistics at the University of Iowa, S300A Lindquist Center, Iowa City, IA 52241; jonathan-templin@uiowa.edu. His primary research interests include multidimensional psychometric methods and Bayesian statistics.

XINYA LIANG is an Associate Professor of Educational Statistics and Research Methods at the University of Arkansas, 751 W Maple Street, Fayetteville, AR 72701; xl014@uark.edu. Her primary research interests include structural equation modeling and Bayesian statistics.

## Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1