Posterior Predictive Model Checks for Cognitive Diagnostic Models

Jung Yeon Park, Matthew S. Johnson, and Young-Sun Lee

Department of Human Development
Teachers College, Columbia University
Box 118, 525 W120th St., New York, NY 10027
E-mail: jyp2111@tc.columbia.edu
E-mail: johnson@tc.columbia.edu
E-mail: yslee@tc.columbia.edu

Abstract

Cognitive diagnostic models (CDMs; DiBello, Roussos, & Stout, 2007) have received increasing attention in educational measurement for the purpose of diagnosing examinees' strengths and weaknesses of their latent attributes. Despite the current popularity of a number of diagnostic models, research on assessing model-data fit has been limited. The current study applies one of the Bayesian model checking methods, namely the posterior predictive model check method (PPMC; Rubin, 1984) to investigate model misfit. Specifically, we aim to employ the technique to investigate model-data misfit from various diagnostic models using real data and a simulation study. An important issue with the application of PPMC is the choice of discrepancy measure. This study examines the performance of three discrepancy measures for assessing different aspects of model fit: observed total-scores distribution, association of item pairs, and correlation of pairs of attributes as adequate measures for the diagnostic models.

Keywords**:** posterior predictive model checking, discrepancy measures, cognitive diagnostic model, DINA, general diagnostic model

1. Introduction

Cognitive diagnostic models (CDMs; DiBello, Roussos, & Stout, 2007) are a type of latent class models assuming that each item in an assessment measures a small number of discrete cognitive skills or attributes. Most CDMs treat the attributes as categorical latent variables; typically, they are binary variables indicating whether examinees have mastered or failed to master skills or attributes. Thus, the purpose of the models is to diagnose and categorize each examinee with a fine-grained attribute profile or pattern of skills they possess. Over the last decade, CDMs have received increasing attention in educational measurement and statistics because of their extensive diagnostic potential.

There is a wide range of models that fall within the framework of CDMs. The models are often characterized by the rules for how the attributes combine, and include both conjunctive and compensatory models (Rupp, Templin, & Henson, 2010). The 'conjunctive' models assume that a lack of a required attribute cannot be compensated by other attributes; for example, the DINA (Deterministic-Input, Noisy-And-gate) model discussed by Junker and Sijtsma (2001), the NIDA (Noisy-Input, Deterministic-And-gate) model by Junker and Sijtsma (2001), and the NC-RUM (Non-Compensatory Reparameterized Unified Model) model by Hartz (2002). In contrast, 'compensatory' models assume that a lack of an attribute can be compensated by another attribute required for the item; for example, the DINO (Deterministic-Input, Noisy-Or-gate) model by Templin and Henson (2006), the NIDO (Noisy-Input, Deterministic-Or-gate) model by Templin (2006), or the compensatory GDM (General Diagnostic Model) by von Davier (2005).

The DINA model (Junker & Sijtsma, 2001; Haertel, 1989; Macready & Dayton, 1977) is one of the most researched conjunctive (or non-compensatory) CDMs for researchers in educational measurement. The model assumes that a correct response on an item depends on

possessing *all* the attributes required to solve that item. To reduce the complexity of estimation in a fully unstructured attribute space, various attribute models have been suggested; examples include the independence model (Maris, 1999) and the higher-order model (HO-DINA; de la Torre & Douglas, 2004).

On the other hand, the GDM (von Davier, 2005; von Davier & Yamamoto, 2004a, 2004b) is a type of compensatory diagnostic model. It is based on extensions of latent class and item response theory models and incorporates a Q-matrix (specifying the attributes required for each item) in the model. This model allows the assumption that a subset of all the required attributes can also contribute to a correct response for the item.

Despite the current popularity of the aforementioned diagnostic models, research on assessing model-data fit is considerably limited. This is problematic because checking model fit is crucial to evaluating the strengths and weaknesses of the proposed models. This provides the motivation for this study where we focus on model diagnostic measures for the DINA model and the GDM, using a Bayesian approach. Specifically, we use posterior predictive model checks (PPMC; Rubin, 1984; Guttman, 1967; Gelman, Meng, & Stern, 1996; Meng, 1994) as the tool for the assessment of model fit.

The PPMC methodology has been recognized as a promising technique for evaluating psychometric models such as unidimensional item response theory (IRT) models (Sinharay, Johnson, & Stern, 2006; Hojitink, 2001; Janssen, Tuerlinckx, Meulders, & De Boeck, 2000), multidimensional IRT models (Levy, Mislevy, & Sinharay, 2009; Levy, 2011), and Bayesian networks (Sinharay, 2006). Sinharay (2006) suggested that the proposed diagnostics for Bayesian networks can also be useful for other CDMs.

In this study, we aim to investigate the performance of different PPMC methods in detecting model-data misfits from various diagnostic models using a real data example and a simulation study. An important issue with the application of PPMC is the choice of discrepancy measures. This study examines the performance of three discrepancy measures for assessing different aspects of model fit; the discrepancy measures we use are the observed total-score distribution, association of item pairs, and correlation of attribute pairs. The diagnostic models include DINA models with three attribute models including (1) independence model, (2) higher-order model, (3) unstructured model, and GDM models with the assumption of (1) constant slopes and (2) varying slopes. We also compared them with the standard two-parameter logistic (2PL) IRT model.

The outline of this paper is as follows. In Section 2.1, we discuss the DINA model formulation and several choices for its attribute components. In Sections 2.2 and 2.3, we describe formulations for GDMs and 2PL IRT models, respectively. Section 3.1 provides a brief discussion of the PPMC method, discrepancy measures, and the posterior predictive p-value. Section 3.2 introduces the discrepancy measures to be used for assessing different aspects of model fit. Section 4 describes implementation of the Bayesian inference using OpenBUGS. In Sections 5 and 6, we apply the suggested model fit measures to a real data example and simulation studies. Section 7 presents a summary and conclusions.

2. The item response models examined in this study

The goal of this paper is to examine the ability of three posterior predictive discrepancy measures to detect model misfit. Below we summarize the models we examine, (1) the DINA

model with (a) the unstructured attribute distribution, (b) the higher-order attribute model, and (c) the independent attribute model; (2) the general diagnostic model; and (3) the 2PL IRT model.

2.1 The DINA

2.1.1 The DINA Measurement Model

Let Y_j be the random response of an examinee to item j, where $= 1, \dots, J$. Then, $Y_j = 1$ if the examinee answers the item correctly or $Y_j = 0$, otherwise. Further, let the observed response pattern be denoted by $\boldsymbol{y} = (y_1, \dots, y_J)^T$. We also denote an examinee's latent attribute by α_k , $k = 1, \dots, K$ where $\alpha_k = 1$ indicates if an examinee masters attribute k or $\alpha_k = 0$, otherwise. Thus the attribute pattern of the examinee can be denoted by $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^T$. In order to implement the CDMs, we need to describe whether the attribute k is necessary for answering item j. The Q-matrix (Tatsuoka, 1985) is a binary matrix, where, $q_{jk} = 1$ indicates attribute k is required to solve item j correctly, and $q_{jk} = 0$, otherwise. Typically, the Q-matrix is determined by subject matter experts before implementing CDMs.

The DINA model (Junker & Sijtsma, 2001; Haertel, 1989; Macready & Dayton, 1977) is typically characterized with a latent response, $\eta_j = \prod_{k=1}^{K} \alpha_k^{q_{jk}}$. That is, $\boldsymbol{\eta} = \left(\eta_1, \dots, \eta_J\right)^T$ represents the deterministic prediction of task performance from the attribute pattern of the examinee. The latent response definition implies that one should master all required attributes in order to respond correctly. The latent response pattern $\boldsymbol{\eta}$ is linked to observed responses in a probabilistic relationship with two "noisy" parameters; slipping (s_j) and guessing (g_j) parameters. The guessing rate is the probability that an examinee responds correctly to item j even though he/she does not possess all the required attributes, i.e., $g_j = P(Y_j = 1 | \eta_j = 0)$. The slipping rate is the probability that an examinee fails to respond to item j correctly even

though he/she possesses all the required attributes, i.e., $s_j = P(Y_j = 0|\eta_j = 1)$. The item response function (IRF) is written as $P(Y_j = 1|\boldsymbol{\alpha}) = (1 - s_j)^{\eta_j} g_j{}^{1-\eta_j}$. The local independence assumption gives that $P(\boldsymbol{Y} \mid \boldsymbol{\alpha}) = \prod_{j=1}^{J} P(Y_j = y_j|\boldsymbol{\alpha})$. The Y_js are independent for all js, given the attribute pattern $\boldsymbol{\alpha}$.

For Bayesian inference, we now define prior distributions on the item parameters, guessing (g_j) and anti-slipping ($a_j = 1 - s_j$). We assume the guessing parameter follows a beta distribution while the anti-slipping parameter follows a uniform distribution i.e. g_j ~ Beta (1 , 2) and $a_j|g_j$ ~ U (g_j, 1). The particular choices of prior distributions were adopted because it is reasonable to assume that the anti-slipping rate is greater than the guessing rate (monotonicity assumption). Also, this form of the prior distribution leads to a uniform joint distribution of the parameters as shown below. Similar prior distributions of item parameters in the DINA model are found in Tseng (2010). Given these priors, the joint prior density function for the two parameters can be derived as:

$$P(a_j, g_j) = P(a_j \mid g_j)P(g_j) \quad (1)$$

$$= \frac{1}{1-g_j}\frac{\Gamma(3)}{\Gamma(1)\Gamma(2)}(1 - g_j)$$

$$= \begin{cases} 2, & \text{where } 0 \le g_j \le a_j \le 1 \\ 0, & \text{otherwise} \end{cases}$$

2.1.2 DINA Attribute Models

Unstructured Model. Recall that $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^T$ denotes a vector of attribute mastery indicators, where $\alpha_k \in \{0,1\}$, $k = 1, \ldots, K$. This model assumes that the probability of each of the different realizations of α_k is unrestricted, allowing for all possible dependencies for pairwise relationships between α_k and $\alpha_{k'}$, where $k \neq k'$. Because α_k is binary, the total number of

possible attribute patterns of $\boldsymbol{\alpha}$ is 2^K. The probability that the attribute vector $\boldsymbol{\alpha}$ takes on the l-th pattern is characterized by a single parameter, π_l, $l = 1, \ldots, 2^K$. Thus the model can be described by the following probability mass function,

$$P(\boldsymbol{\alpha} = \boldsymbol{\alpha_l}) = \pi_l, \tag{2}$$

where $\boldsymbol{\alpha_l}$ is the l-th pattern, $l = 1, \ldots, 2^K$.

We place a prior distribution on the structural parameters of the saturated-DINA model as follows. Assuming a categorical distribution for $P(\boldsymbol{\alpha})$, we are interested in estimating the hyperparameters, $\pi_1, \ldots, \pi_{2^K}$, and need to choose a hyperprior distribution. The Dirichlet distribution is used because it is a conjugate distribution of a categorical distribution. We assume $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{2^K})$ has a Dirichlet ($\boldsymbol{m}$) distribution with 2^K categories and parameters $\boldsymbol{m} = (m_1, m_2, \ldots, m_{2^K})$, where $m_l = 0.5$ for $l = 1, \ldots, 2^K$. Then the density function is $Dir_m(\pi) = Dir(\pi|\boldsymbol{m}) = \frac{\Gamma(m_0)}{\prod_{l=1}^{2^K}\Gamma(m_k)}\prod_{l=1}^{2^K}\pi_l^{m_k-1}$, where $m_l > 0$, $m_0 = \sum_l m_l$, and $\sum_l \pi_l = 1$.

There is a shortcoming in the unstructured attribute model. The unstructured attribute distribution is computationally intensive due to the total number of latent classes increasing dramatically as the number of attributes increases. For example, when there are K = 3 attributes in the DINA model, 7 ($= 2^3 - 1$) latent class parameters are required to be estimated. However, if K = 9, then 511($= 2^9 - 1$) latent class parameters need to be estimated. For this reason, more informative and parsimonious ways to summarize associations among the attributes are desired.

Higher-order Model. The purpose of the higher-order DINA (HO-DINA) model is to model the joint distribution of attribute patterns using higher-order latent traits. The model is motivated by settings where the notion of higher-order latent traits represents measures of general knowledge defined more broadly than the individual attributes in the cognitive diagnostic model (de la Torre

& Douglas, 2004). The HO-DINA model assumes conditional independence—specifically, that the attributes are conditionally independent under the general knowledge, θ i.e. $P(\boldsymbol{\alpha}|\theta) = \prod_{k=1}^{K} P(\alpha_k|\theta)$. For a more parsimonious model, in this study, we assume that the higher order model for the attributes is a one-parameter logistic (1 PL) model:

$$P(\alpha_k = 1|\theta) = \frac{\exp(\lambda_1(\theta - \lambda_{0k}))}{[1+\exp(\lambda_1(\theta - \lambda_{0k}))]}, \quad (3)$$

where λ_{0k} indicates the attribute difficulty parameter and λ_1 indicates the attribute discrimination parameter, which is assumed to be constant across items. By denoting $\pi_k = P(\alpha_k = 1|\theta)$, this model is characterized by $P(\boldsymbol{\alpha}|\theta) = \prod_{k=1}^{K} {\pi_k}^{\alpha_k} (1 - \pi_k)^{1-\alpha_k}$. Finally, the marginal probability of having the attribute has the form, $P(\boldsymbol{\alpha}) = \int P(\boldsymbol{\alpha}|\theta)P(\theta)\, d(\theta)$, where $P(\boldsymbol{\alpha}|\theta)$ is given by the higher order model.

The prior distributions for λ_1 and λ_{0k} are as follow: $\lambda_{0k} \sim N(0,1)$ and $\lambda_1 \sim N(0,1)I[\lambda_1 \in (0,\infty)]$. As in a typical IRT model framework, the general knowledge state θ follows a standard normal distribution, $\theta \sim N(0,1)$.

Independence Model. The simplest model for α_k is to assume that the α_ks are independently distributed (Maris, 1999). Because α_k is binary, it is characterized by K parameters π_k which are defined by a probability of mastering attribute k, i.e., $P(\alpha_k = 1)$. Then, the probability mass function is $P(\boldsymbol{\alpha}) = \prod_{k=1}^{K} {\pi_k}^{\alpha_k} (1 - \pi_k)^{1-\alpha_k}$. The prior for each π_k is a uniform distribution, $\pi_k \sim \text{Unif}(0,1), k = 1, \ldots, K$.

2.2. General Diagnostic Model

The general class of GDMs (von Davier, 2005; von Davier & Yamamoto, 2004a 2004b) was developed with the goal of maintaining similarities to previous approaches using ideas from IRT, log-linear models, and latent class analysis. A central idea of the GDM is that the Q-matrix

generates a matrix of relations between items and attributes required to solve those items. One of the advantages of the GDM lies in its applicability of attribute models for polytomous item responses and for attributes with more than two proficiency levels.

Let us assume that both the response outcome and attribute proficiency are dichotomous. If we assume a logistic link function, the model can be formulated as

$$P\left(Y_j = 1 \middle| \beta_j, \gamma_j, \alpha\right) = \frac{\exp\left[\beta_j + \gamma_j^T h(\boldsymbol{q}_j, \boldsymbol{\alpha})\right]}{1 + \exp\left[\beta_j + \gamma_j^T h(\boldsymbol{q}_j, \boldsymbol{\alpha})\right]}\ , \tag{4}$$

with K attributes (discrete latent traits), $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ is the attribute proficiency and $\boldsymbol{q}_j$ is the set of attributes influencing item j as given by the *j-th* row of the Q-matrix. As shown in the equation, the probability of a correct response can be thought to incorporate two parts: the overall difficulty i.e. β_j and a linear combination of interactions of attributes required and attributes present: $h\left(\boldsymbol{q}_j, \boldsymbol{\alpha}\right) = \left(h_1\left(\boldsymbol{q}_j, \boldsymbol{\alpha}\right), \ldots., h_m\left(\boldsymbol{q}_j, \boldsymbol{\alpha}\right)\right)$. In this study, we use $h\left(\boldsymbol{q}_j, \boldsymbol{\alpha}\right) = \left(q_{j1}\alpha_1, \ldots, q_{jK}\alpha_K\right)$. Then, given a nonzero Q-matrix entry, the slope parameter, γ_{jk} determines how much the particular skill components in $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ contribute to the response probabilities for item j. In another formulation of the model we use $h\left(\boldsymbol{q}_j, \boldsymbol{\alpha}\right) = \sum_{k=1,\ldots K} q_{jk}\alpha_k$. In this case the probability of correct response is only influenced by the number of required attributes present. This model can also be thought of as a constant-slope simplification of the previous GDM (i.e. all the γ_{jk}'s are equal for $k = 1, \ldots, K$).

2.3. The Two Parameter Logistic Model

In the two parameter logistic model (2PL), the item response model is defined by the following:

$$P\left(Y_j = 1 \middle| \ \xi,\ \nu_{0j}, \nu_{1j}\right) = \frac{\exp\left(\nu_{0j}\left(\xi - \nu_{1j}\right)\right)}{1 + \exp\left(\nu_{0j}(\xi - \nu_{1j})\right)}. \tag{5}$$

The prior distributions for ν_{1j} and ν_{0j} are as follow: $\nu_{1j} \sim N(0,100)$ and $\nu_{0j} \sim N(0,100)I[\nu_{0j} \in (0,\infty)]$, $j = 1, \ldots, J$. As in the typical IRT model framework, the general knowledge state ξ follows a standard normal distribution, $\xi \sim N(0,1)$.

3. Posterior Predictive Model Checks, Test Quantities, and P-values

3.1 Overview of the Method

The basic idea of posterior predictive model checks (PPMC; Rubin, 1984; Guttman, 1967; Gelman, Meng, & Stern, 1996; Meng, 1994) is that data generated from a model should be similar to the observed data under the condition that the model describes the underlying structure of the data properly. In the posterior predictive model checking method, data is generated from the posterior predictive distribution and the fit of the model to the observed data is compared to the fit of the model to the generated data.

Let us denote the prior distribution of the parameters by $p(\boldsymbol{\psi})$ and the likelihood function of the model with $p(\boldsymbol{Y}|\boldsymbol{\psi})$. Then the posterior distribution of the $\boldsymbol{\psi}$, $p(\boldsymbol{\psi}|\boldsymbol{Y})$, is proportional to the product of $p(\boldsymbol{\psi})$ and $p(\boldsymbol{Y}|\boldsymbol{\psi})$. The posterior predictive distribution, $p(\boldsymbol{Y}^{rep}|\boldsymbol{Y})$, is viewed as a combination of the likelihood function for replicated data and the posterior distribution of the $\boldsymbol{\psi}$; that is, $p(\boldsymbol{Y}^{rep}|\boldsymbol{Y}) = \int p(\boldsymbol{Y}^{rep}|\boldsymbol{Y},\boldsymbol{\psi})\, p(\boldsymbol{\psi}|\boldsymbol{Y})\, d\boldsymbol{\psi} = \int P(\boldsymbol{Y}^{rep}|\boldsymbol{\psi})\, P(\boldsymbol{\psi}|\boldsymbol{Y})\, d\boldsymbol{\psi}$ as $\boldsymbol{Y}^{rep}$ and $\boldsymbol{Y}$ conditionally are independent given $\boldsymbol{\psi}$.

The posterior predictive distribution of the replicated data can be considered as an empirical distribution of the fitted model, and hence the amount of discrepancy between the observed data and the fitted model can be summarized by calculating empirical p-values from the replicated data sets. One of the advantages of the PPMC method is that it is flexible enough to use different test quantities (so called discrepancy measures; Gelman, Carlin, Stern, & Dunson,

2013) according to the researchers' need. The empirical p-value is equal to the probability that the test quantities using replicated data sets are greater than the test quantity using an observed data set; that is $P[D(\boldsymbol{Y}^{rep}, \boldsymbol{\psi}) \geq D(\boldsymbol{Y}, \boldsymbol{\psi})|\boldsymbol{Y}]$, where $D(.)$ is a test quantity. P-values that are extremely small (.05 or smaller) or high (.95) might indicate that the model does not capture some particular aspect of the observed data; even moderate p-values (less than .2 or greater than .8) can be indicative of problems with the fit of the model.

3.2. Discrepancy measures for diagnostic models

Observed Score Distribution The first discrepancy measure we use in this study is the observed total-score distribution (e.g., Hambleton & Han, 2004; Sinharay et al., 2006; Levy, 2006). Let n_j denote the number of examinees answering exactly j items correct, $j = 0, \ldots, J$. Then, the proportion of examinees who obtained observed sum-score j is

$$PC_j = \frac{n_j}{n} \ . \tag{6}$$

Association between Item Pairs As a measure of bivariate association of item pair, the odds ratio was used. Let $n_{kk'}$ denote the number of examinees scoring k on item j and k' on item j'. Then the sample odds ratio is denoted by

$$OR_{jj'} = \frac{(n_{11})(n_{00})}{(n_{10})(n_{01})}, \ j \neq j' \text{ and } k, k' = 0,1. \tag{7}$$

The measure has appeared to detect violations of the local independence assumption in unidimensional IRT (e.g., Chen & Thissen, 1997) and was also found to be useful to detect model misfit under various conditions within the context of IRT and Bayesian network (Sinharay et al., 2006).

Associations among the attribute pairs In order to measure bivariate association between attribute pairs, we employed per attribute sum-scores (Henson, Templin, & Douglas, 2007; Chiu, Douglas, & Li, 2009). Let us define a sum-score for an attribute as

$$W_{ik} = \sum_{j=1}^{J} Y_{ij} q_{jk} \,, \tag{8}$$

for examinee i. Then W_{ik} ranges from 0 to the total number of items that require attribute *k* i.e. $W_{ik} \in [\, 0 \,,\, J_k \,]$, where J_k denotes the total number of items for attribute *k*. The measure was originally used to cluster subjects into correct underlying latent classes (Chiu, Douglas, & Li, 2009). In this study, we calculated Pearson product-moment correlation coefficients between the attribute sum-scores W_k and $W_{k'}$ to quantify the strength between pairs of attributes.

4. Implementation

The MCMC algorithm is implemented with R 3.1.1 (R Core Team, 2013), R2OpenBUGS (Sturtz, Ligges, & Gelman, 2005) and OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009). For each analysis, four chains were run, and each ran for 25,000 iterations. We used a thinning parameter of 10 and used the first half as burn-in. The resulting 1,250 iterations from each chain were pooled and randomly mixed, thus totaling 5,000 iterations after burn-in to be used as samples from the posterior distribution of the parameters. The use of multiple chains and thinning serves to reduce the dependencies among the iterations and ensures adequate convergence to and mixing from the posterior distribution.

For each sample from the posterior distribution of the parameters we generated a replicated data set, giving us a total of 5000 replicated datasets for each model. While generating the replicated datasets we started by simulating the parameters of the highest possible level of the attribute distribution, e.g., $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{2^K})$ for unstructured DINA models and GDMs,

$\theta \sim N(0,1)$, λ_{0k} and λ_1 for the higher-order DINA model , $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$ for the independence DINA model and $\theta \sim N(0,1)$ for the 2PL model. With the data generation by first sampling the parameters at this level, we decrease the association between the observed and replicated data sets. After the attribute distributions were sampled, we simulated an attribute pattern for each individual in the data set, and concluded by sampling the individual item responses.

5. Real Data Example: Fraction Subtraction Data

We demonstrate the performances of diagnostic models by applying the PPMC method to the fraction subtraction data, which was originally used in K. Tatsuoka (1990) and also appeared in C. Tatsuoka (2002), de la Torre and Douglas (2004), and de la Torre (2009). The data is comprised of 15 fraction subtraction test items from 536 middle school students. Table 1 shows the corresponding Q-matrix (de la Torre, 2009) which specifies the five required attributes to solve each of the 15 questions. This section demonstrates the various aspects of model-data misfit with regard to (1) observed total score distribution, (2) item-pair associations, and (3) attribute-pair association. The three discrepancy measures mentioned in the previous section are used to evaluate the appropriateness of the diagnostic models.

Results

Figure 1 compares the observed total score distribution with the corresponding posterior predictive distributions by fitting the six models to the fraction subtraction data. As shown in the figure, the proportion of examinees for each observed total score (dots in black) and the posterior predictive means are presented (dots in grey) with corresponding 90% posterior predictive intervals (dotted lines in black). If the posterior predictive intervals do not contain the proportion

of examinees associated with the observed total scores, it suggests lack of fit of the models to predict the score distribution. Overall, the result suggests that overall the six models perform similarly. It was found that choosing unstructured or higher-order structures of the DINA model hardly improve the accuracy in predicting the total-score distribution. It appears that the GDMs performed slightly better than DINAs in predicting the percentage of total examinees for each total score.

Figure 2 shows the posterior predictive p-values corresponding to the odds ratio for each item pair (with 15 items). The p-values are represented visually with heat maps, where color gradations represent the level of p-values. For maximal p-values between 1.00 and 0.95 the color is red, and for minimal p-values between 0.00 and 0.05 the color is blue. The p-values between 0.8 and 0.95 are colored in orange and those between 0.05 and 0.2 are colored in light blue. Finally, the lightest color corresponds to moderate p-values, between 0.20 and 0.80. Note that the plot is symmetric with respect to the diagonal (top-left to right-bottom). Because we analyzed 15 questions, there are a total of 105 pairs of associations to be considered.

We can see from the figure that the 2PL model failed to fit item-pair associations for 33 percent of the items (35 out of 105 pairs). In particular, most of the misfit involves item 2 and item 3.

It appears that all the three DINA models also failed to predict many of the item-by-item associations. In particular, the DINA model with an independence structure failed to predict item-pair odds ratios too often when item 1, item 5, item 8, and item 10 are involved. The four items led to extremely low p-values ($p \leq 0.05$). The higher-order and unstructured DINA models showed better performances than the independence structure but often failed to predict the item-pair associations when item 1 and item 10 are involved. Overall, the DINA models resulted in

extreme p-values ($p \leq 0.05$) for item-pairs in nearly 50 percent (52 out of 105 pairs) for the independence model, 38 percent (40 out of 105 pairs) for the higher-order model, and 35 percent (37 out of 105 pairs) for the unstructured model.

The GDM with the constant slopes resulted in 21 percent (22 out of 105 pairs) of the item pairs producing extreme p-values ($p \leq 0.05$ or $p \geq 0.95$). Specifically, nine of the extreme p-values involve item 1. But the GDM with varying slopes resulted in only 9 percent (9 out of 105 pairs) of the extreme p-values. It was found that allowing different slopes for each attribute significantly reduced misfit involving item 1.

Finally, attribute-pair correlations were evaluated with the Pearson product-moment correlation coefficients between 'per-skill sum score' (Henson et al., 2007; Chiu et al., 2009). The posterior predictive p-values are shown in Table 2. Results suggest that the three DINA models with independence, higher-order, unstructured models failed to predict associations between attribute masteries for nine out of 10 pairs and the 2PL model for 7 out of the 10 pairs. Compared to the conjunctive models, compensatory GDMs predicted the attribute-pair correlations more accurately. The GDM with constant slopes led to four extreme p-values i.e. (α_1, α_2), (α_1, α_3), (α_2, α_3), and (α_2, α_4) out of 10 pairs. Finally, the GDM with the varying slopes for attribute led to only one extreme p-value i.e. (α_1, α_4).

Given these results, it appears that the compensatory GDMs we examined provide the best fit to the fraction subtraction data. Given that the fraction subtraction data has been a popular example data set for demonstrating the various versions of the DINA model, it is important to examine if the types of misfit we discovered are in fact atypical for data actually generated from the DINA model. The simulation study we discuss below is meant to examine

the expected behavior of the PPMC measures we used here, when we know that the data actually has come from the DINA model.

6. Simulation Study

Design

We generate our simulation data sets using the DINA as the measurement model with the unstructured attribute space. In order to mimic real data situations, we employed the Q-matrix designed for Examination for the Certificate of Proficiency in English (ECPE; Buck & Tatsuoka, 1998; Henson et al., 2007). The Q-matrix was originally constructed with three attributes (morphosyntactic, cohesive, and lexical attributes) to solve a total of 28 questions. The matrix was modified to create three different Q-matrices by varying test lengths to be 10, 20, and 30 (see Table 3). True guessing and slipping parameters for all items were fixed at 0.2.

For the purpose of simulating from the attribute space, the simulated data was generated by specifying the tetrachoric correlation between attribute pairs using the R package 'CDM' (Robitzsch, Kiefer, George, & Uenlue, 2014). Specifically, we considered the true attribute space where attribute masteries are correlated weakly (ρ=0.2 for each pair), moderately (ρ=0.5 for each pair), and strongly (ρ=0.8 for each pair). Attribute mastery probabilities for the three attributes were set at 0.5. The sample size for each simulation condition was n=1,000.

In summary, we examine simulated data generated under nine conditions defined by two factors: (A) three levels of test length (10 items, 20 items, and 30 items) and (B) three levels of attribute correlation (weak, moderate, and high relationships). Each simulation data set was analyzed with a total of six models, i.e. DINA models with (1) unstructured, (2) higher-order, and (3) independent attribute space; two GDM models with (4) constant slopes and (5) varying

slopes; and (6) the 2PL model. In the simulation study, results in relation to the total-score distribution was not considered as the real data example suggests that the measure provides limited information as compared to the other two measures.

Results of item-pair associations

Posterior predictive p-values were calculated for the item-pair odds ratios using the 5,000 replicated data sets. Because test lengths varied (10, 20, and 30 items), we consider p-values corresponding to (10×9)/2 = 45, (20×19)/2 = 190, and (30×29)/2 = 435 pairs of odds ratios, respectively. Table 4 shows the percentage of the p-values either greater than .95 or less than .05. The numbers within parentheses refer to the p-values either greater than .80 or less than .20.

Results suggested that the 2PL model performs better in predicting the item-pair odds ratios as the strength between attribute masteries get stronger and as the test length decreases. When the strengths are moderate to high, over 35 percent of p-values are more extreme than (.05, .95) and 65% of them are more extreme than (.20, .80). The DINA with independent attributes showed a great deal of model misfit when strengths between attributes are moderate to strong. Specifically, the model failed to predict more than 50 percent of the item-pair associations. In contrast, the higher-order and unstructured attribute models for the DINA performed well across all conditions. The model fits were improved with the strength between attribute masteries.

The GDM model fits were the best across most conditions. The GDM with constant slopes worked almost as good as varying slopes except when the simulated data had 20 items with a weak to moderate association among the attributes. It appears that the model fits from the unstructured DINA model and the GDM with varying slopes were the best but tended to show mild over-fittings when the magnitude of the associations between attribute pairs is strong.

Results of attribute-pair associations

In order to examine the model-data misfit of the diagnostic models in predicting bivariate relationships between pairs of attribute masteries, p-values were calculated for the attribute-pair correlation coefficients using the posterior predictive distribution based on the 5,000 replicated data sets (Table 5). Because the simulation condition was created using three attributes, three pairwise associations between attribute masteries (attribute 1 vs. attribute 2, attribute2 vs. attribute 3, and attribute 1 vs. attribute 3) were examined.

As expected, the 2PL model failed to predict almost all attribute-pair associations; almost all p-values are greater than .95 (nearly 1), which implies that it consistently over-predicted the associations between pairs of attribute masteries.

In contrast, independence structures in the DINA model consistently under-predicted the associations. When relations between attribute masteries are moderate to strong, the p-values are nearly zero. The model also failed to predict them when the true attributes are weakly related. The higher-order model performed sufficiently well enough to predict all possible associations; p-values range from .08 to .81. Also, the p-values from the unstructured DINA model range from .32 to .61 across all simulation conditions.

The GDM with constant slopes revealed a few misfits in the relations between attribute 1 and attribute 3. The model tended to under-predict the relations when the underlying condition was weak or moderate and test lengths were 20 or 30 but over-predict the relations when the underlying condition was strong. Except for those conditions, the p-values range from .31 to .61. Similarly, the GDM with varying slopes under-predicted the relation between attribute 1 and attribute 3 when the true attribute-pairs were weakly correlated and the test length was 20 but

over-predicted those associations when the true attribute-pairs were strongly correlated and the test length was 20. For the remaining conditions, the p-values range from .21 to .79.

7. Summary and Discussion

Since its introduction a decade ago, there have been quite a number of published articles related to CDMs, with the models receiving increasing attention from educational researchers. Despite the recent popularity, however, there has been a lack of research focusing on evaluating model-data misfit. The PPMC method has been considered to be an attractive method to detect various perspectives of model-data misfit in psychometric models (e.g., Sinharay et al., 2006; Levy, 2006, 2011; Sinharay, 2005). However, the technique has not been fully applied to a variety of diagnostic models, an issue the current study addressed. In this article, we explored the extent to which each model is capable of predicting some systematic aspects of data structure. We considered diagnostic models based on conjunctive and compensatory rules. Also, we examined the impact of inadequately defined attribute structures on the DINA model.

Three discrepancy measures were used, total-score distributions, odds ratios between item pairs, and correlations between attribute pairs. The observed total-score distribution was first examined in the real data example; it provided a simple graphical way to detect if shape of response function can be replicated by the DINA models in the real data example. The odds ratios as an item-fit measure has been widely considered as a powerful discrepancy measure to detect associations among item pairs (e.g., IRT, MIRT, and BN). Therefore, the measure was adapted to real data analysis and simulation of current study. In the real data example, the posterior predictive p-values associated with this measure were efficiently presented by using a visualization technique. Finally, we introduced correlation coefficients between per attribute

sum-scores (Henson et al., 2007; Chiu et al., 2009) as a measure of checking the associations between pairs of attributes. This allows for the detection of correlations among attributes to see if the model can reproduce the direction and strength among attribute pairs.

The real data analysis example revealed that the DINA model showed considerable lack of fit to the fraction subtraction data. In particular, assuming independence among attributes is an overly restrictive approach towards modeling the attribute space. The models showed serious misfits in predicting true associations in item pairs. It was noticeable that the 2PL IRT model performed better than the DINA models, which implied that the conjunctive rule in the DINA models is not appropriate for the fraction subtraction data. The compensatory GDMs with constant slopes and varying slopes showed minimum amount of misfit. It implied that the conjunctive rule is overly restrictive in this case.

While the simulation study was somewhat limited in scope by the fact that it did not replicate data for each condition, and examined only a single sample size condition it did produce some interesting results. The simulation study suggests that the independence structure of the DINA model has serious misfits when the true correlation structure among attributes is moderate to strong. The higher-order model worked almost as well as the unstructured model. Perhaps it is because we considered only three attributes in the simulation conditions. Ideally, the goal is to find a structural model which is parsimonious yet closer to the true attribute space. The simulation results suggested that the two GDMs performed well in most conditions although the true generating model was the unstructured DINA model. Particularly when test length is 10, the GDMs worked even better than the unstructured DINA model with regard to item-pair associations. One possible explanation for the GDMs being favored is related to the fact that the

DINA model is a constrained GDMs; the DINA model is mathematically equivalent to a particular type of the GDMs with reduced number of parameters (see von Davier, 2014).

After finding particular model misfits, it is important to take the best possible action to address them. As previously indicated, modifying structural parts of the DINA model is not the only way to address model misfit. Sinharay (2007) suggests that if a particular model would be the final model for a given data set, particular items which involve serious misfits with the model should be removed and then the model should be fit with the remaining items, which would result in an adequate model fit using PPMC. Depending upon which discrepancy measure is used and how sensitive it is, there could always be occasions to find at least some defects in the finalized model. Therefore, it is important to emphasize that practitioners should be aware of its defects and possible consequences when choosing models to make inferences (Gelman et al., 1996).

For future research, the PPMC method needs to be applied in a more diverse way. For example, a more efficient discrepancy measure needs to be developed to address the lack of accuracy in determining item-attribute associations as specified in the Q-matrix. Or different methods for quantifying the magnitude of discrepancies, other than calculating the posterior predictive p-value, can be developed. For example, Wu, Yuen, and Leung (2014) proposed relative entropy posterior predictive model checking method (RE-PPMC) which utilizes the information of the whole distribution to measure the difference between the realized and predictive distribution using the relative entropy. Also, we are interested in potential lack of fit due to the dichotomized scale of the attributes; dividing individual cognitive attributes into two categories (mastery or non-mastery) is perhaps too restrictive for inferences to be made, so there could be considerable loss of information. Finally, more thorough simulation study is desirable

with multiple simulation data sets, varying sample sizes, and different data-generating models (e.g. GDM).

References

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

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using Item Response Theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289.

Chiu, C., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633-655.

Davier, M., von, & Yamamoto, K. (2004a, October). A class of models for cognitive diagnosis. Paper presented at the 4th Spearman Conference, Philadelphia, PA.

Davier, M., von, & Yamamoto, K. (2004b, December). A class of models for cognitive diagnosis and some notes on estimation. Paper presented at the ETS Tucker Workshop Seminar, Princeton, NJ.

von Davier, M. (2005). A General Diagnostic Model Applied to Language Testing Data. Research Report RR-05-16. ETS: Princeton, NJ.

von Davier, M (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*, 49-71.

DiBello, L., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C.R Rao & S. Sinharay (Eds.) *Handbook of Statistics, 26*, (pp. 979-1030). Amsterdam: Elsevier

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333-353.

de la Torre, J. (2009). DINA model and parameter estimation: a didactic. *Journal of Educational*

and Behavioral Statistics, 34 , 115-130.

Gelman, A., Meng, X. L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica, 6,* 733-807.

Gelman A., Carlin, J. B., Stern, H.S., & Dunson, D. B. (2013). *Bayesian data analysis*. 3rd edition. New York: Chapman & Hall.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society B, 29*, 83-100.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 333-352.

Hambleton, R. K., & Han, N. (2004, April). *Assessing the fit of IRT models: Some approaches and graphical displays*. Paper resented at the annual meeting of the National Council on Measurement in Education. San Diego.

Hartz, S. (2002). A Bayesian framework for the Unified Model for assessing cognitive abilities: blending theory with practice. Unpublished doctoral thesis, University of Illinois at Urbana-Champain.

Henson, R., Templin, J., & Douglas, J. (2007). Using efficient model based sum-scores for conducting skill diagnoses. *Journal of Educational Measurement, 44*, 361-376.

Hoijtink,H.(2001).Conditional independence and differential item functioning in the two-parameter logistic model. In A. Boomsma, M. A. J.van Duijn, & T. A.B. Snijders(Eds.),Essaysinitemresponsetheory(pp.109-130).NewYork,NY: Springer-Verlag.

Jannsen, R., Tuerlinckx, F., Meulders, M.,& De Boeck, P.(2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285-306

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33*, 519-537.

Levy, R. (2006). *Posterior predictive model checking for multidimensionality in item response theory and Bayesian networks*. Unpublished doctoral dissertation, University of Maryland.

Levy, R. (2011). Posterior predictive model checking for conjunctive multidimensionality in item response theory. *Journal of educational and behavioral statistics, 36,* 672-694.

Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009) The BUGS project: Evolution, critique and future directions (with discussion), *Statistics in Medicine* 28**:** 3049--3082.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 33,* 379-416.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics*, 22, 1142-1160.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2013). *CDM: Cognitive*

Diagnosis Modeling. R package version 2.4-9. http://CRAN.R-project.org/package=CDM.

Rupp, A. A., & Templin, J. (2008). The effect of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and psychological measurement, 68*, 78-96.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of statistics, 12,* 1151-1172.

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

Sinharay, S. (2006). Model diagnostics for Bayesian network. *Journal of educational and behavioral statistics, 31*, 1-33.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied psychological measurement, 30*, 298-321.

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

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software*, *12*(3), 1–16.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12*, 55-73.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge Acquisition* (pp. 453–488). Mahwah, NJ: Erlbaum.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification

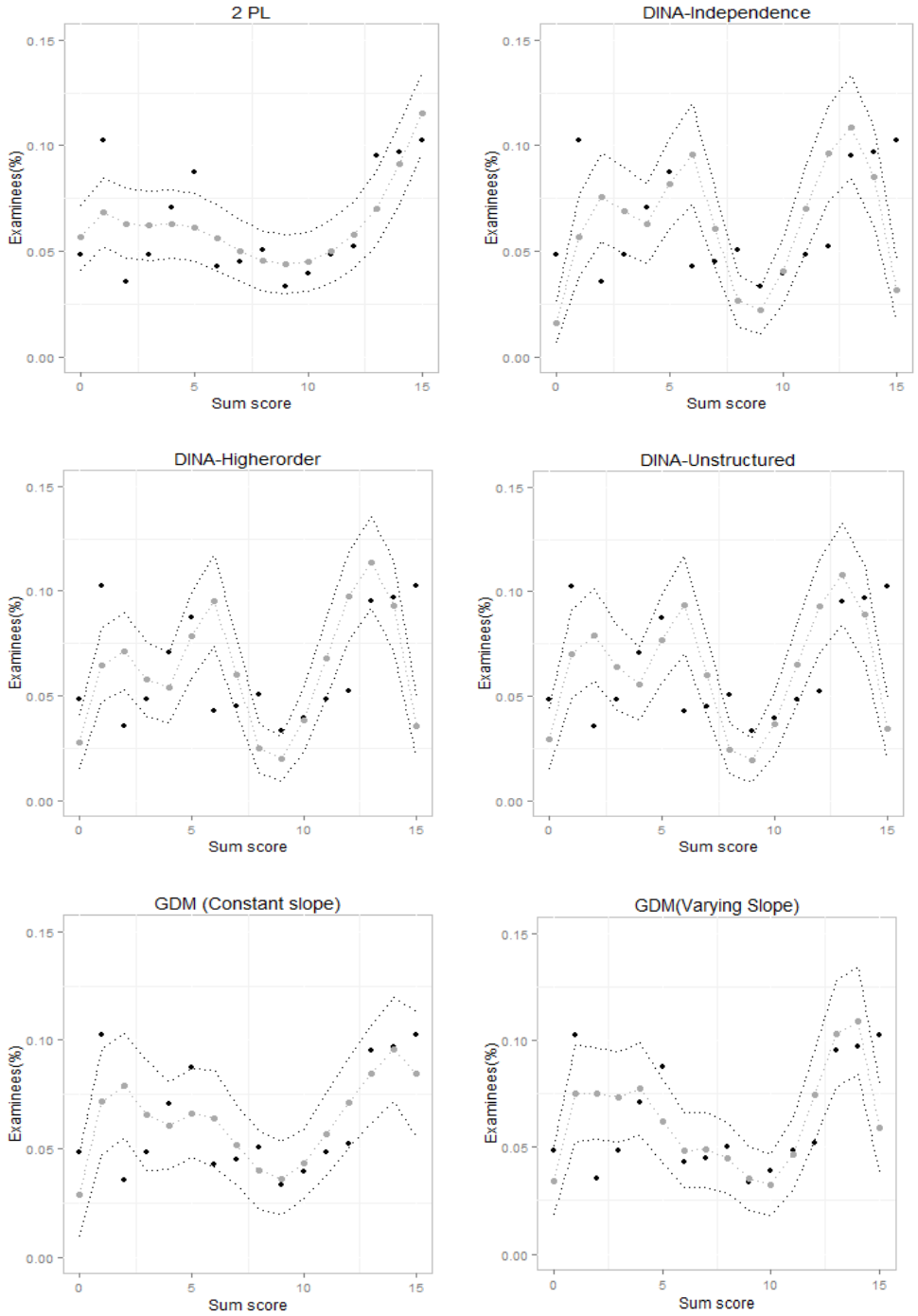
models. *Journal of the royal statistical society: series C (Applied statistics), 51*, 337-350.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11* , 287-305.

Tseng, H. (2010). *A linear compensatory counterpart to and generalization of the DINA model*. Unpublished doctoral dissertation, Columbia University.

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

Figure 1. Summaries of the posterior predictive distributions of the observed scores for the Fraction Subtraction data



Note. Dots (black) = Observed proportion of examinees; Dots (grey) = Posterior mean for the proportion of examinees; Dotted line (black) = 90% posterior predictive interval.

Figure 2. Posterior predictive *p*-values for item odds ratios: Fraction Subtraction data

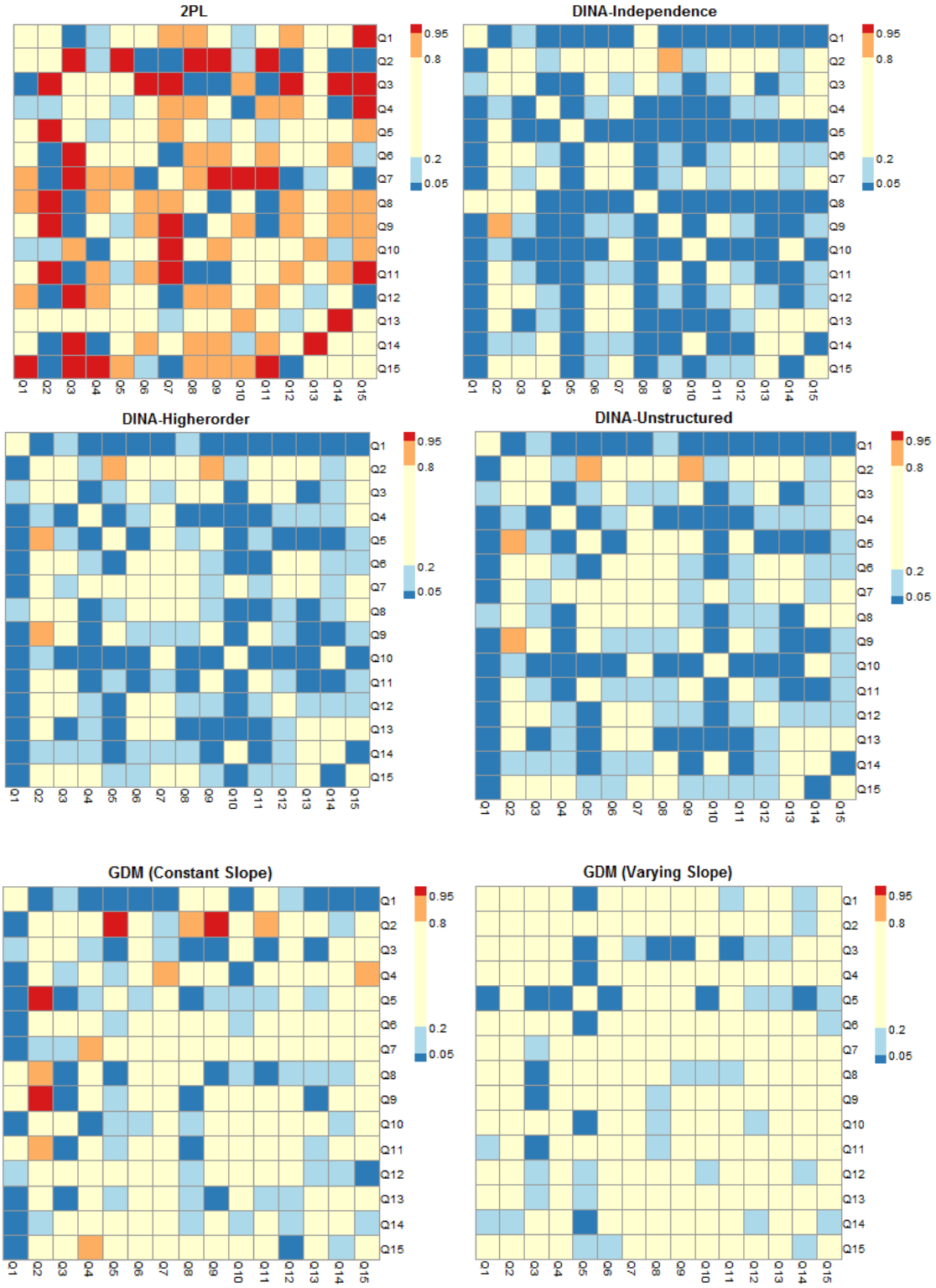


Table 1. Q-matrix designed for Fraction Subtraction data (15 items)

Item No.	Item	α_1	α_2	α_3	α_4	α_5
1	3/4 - 3/8	1	0	0	0	0
2	3 1/2 - 2 3/2	1	1	1	1	0
3	6/7 - 4/7	1	0	0	0	0
4	3 - 2 1/5	1	1	1	1	1
5	3 7/8 - 2	0	0	1	0	0
6	4 4/12 - 2 7/12	1	1	1	1	0
7	4 1/3 - 2 4/3	1	1	1	1	0
8	1 1/8 - 1/8	1	1	0	0	0
9	3 4/5 - 3 2/5	1	0	1	0	0
10	2 - 1/3	1	0	1	1	1
11	4 5/7 - 1 4/7	1	0	1	0	0
12	7 3/5 - 4/5	1	0	1	1	0
13	4 1/10 - 2 8/10	1	1	1	1	0
14	4 - 1 4/3	1	1	1	1	1
15	4 1/3 - 1 5/3	1	1	1	1	0

Table 2. Posterior predictive p-values from the Pearson product-moment correlation between attribute-scores for the Fraction Subtraction data (five attributes)

		DINA (Conjunctive)			GDM (Compensatory)	
Pairs	2 PL	Independence	Higherorder	Unstructured	Constant Slope	Varying Slope
(α_1, α_2)	0.991	<.001	<.001	<.001	0.020	0.495
(α_1, α_3)	0.849	<.001	<.001	<.001	<.001	0.149
(α_1, α_4)	1.000	<.001	<.001	<.001	0.560	0.951
(α_1, α_5)	1.000	<.001	<.001	0.002	0.546	0.454
(α_2, α_3)	0.733	0.002	0.010	0.024	0.035	0.063
(α_2, α_4)	0.186	<.001	0.004	0.002	0.031	0.181
(α_2, α_5)	1.000	0.016	0.030	0.035	0.492	0.308
(α_3, α_4)	1.000	0.041	<.001	0.001	0.887	0.105
(α_3, α_5)	1.000	0.004	0.004	0.004	0.500	0.127
(α_4, α_5)	1.000	0.045	0.060	0.066	0.665	0.222
# Extreme p-values	7	10	9	9	4	1

Table 3. Q-matrix designed for simulated data

Item	α_1	α_2	α_3	10 items	20 items	30 items
1	1	1	0	x	x	x
2	0	1	0	x	x	x
3	1	0	1	x	x	x
4	0	0	1	x	x	x
5	0	0	1	x	x	x
6	0	0	1	x	x	x
7	1	0	1	x	x	x
8	0	1	0	x	x	x
9	0	0	1	x	x	x
10	1	0	0	x	x	x
11	1	0	1	.	x	x
12	1	0	1	.	x	x
13	1	0	0	.	x	x
14	1	0	0	.	x	x
15	0	0	1	.	x	x
16	1	0	1	.	x	x
17	0	1	1	.	x	x
18	0	0	1	.	x	x
19	0	0	1	.	x	x
20	1	0	1	.	x	x
21	1	0	1	.	.	x
22	0	0	1	.	.	x
23	0	1	0	.	.	x
24	0	1	0	.	.	x
25	1	0	0	.	.	x
26	0	0	1	.	.	x
27	1	0	0	.	.	x
28	0	0	1	.	.	x
29	1	1	0	.	.	x
30	0	1	0	.	.	x

Table 4. % of extreme *p*-values for item-pair odds ratios (Simulation)

			DINA (Conjunctive)			GDM (Compensatory)	
Strength	# Items	2 PL	Independence	Higherorder	Unstructured	Constant Slopes	Varying Slopes
Weak	10	.38 (.67)	.20 (.36)	.09 (.33)	.07 (.20)	.04 (.24)	.04 (.17)
	20	.46 (.72)	.13 (.35)	.07 (.32)	.04 (.26)	.09 (.29)	.01 (.21)
	30	.44 (.69)	.14 (.39)	.04 (.32)	.03 (.24)	.06 (.35)	.06 (.32)
Moderate	10	.36 (.69)	.53 (.67)	.04 (.29)	.07 (.24)	.02 (.24)	.02 (.24)
	20	.36 (.59)	.38 (.62)	.01 (.16)	.02 (.15)	.14 (.43)	.03 (.21)
	30	.41 (.66)	.45 (.63)	.04 (.23)	.03 (.20)	.03 (.28)	.03 (.29)
Strong	10	.18 (.47)	.64 (.73)	.02 (.24)	.02 (.27)	.00 (.20)	.00 (.24)
	20	.18 (.39)	.54 (.73)	.01 (.15)	.01 (.17)	.02 (.24)	.04 (.23)
	30	.23 (.54)	.62 (.76)	.01 (.23)	.01 (.23)	.02 (.30)	.05 (.31)

Table 5. Posterior predictive p-values from the Pearson product-moment correlation between attribute-scores

				DINA (Conjunctive)			GDM (Compensatory)	
Strength	# Items	Pair	2 PL	Independence	Higherorder	Unstructured	Constant Slopes	Varying Slopes
Weak	10	12	0.22	0.08	0.73	0.46	0.33	0.32
		13	1.00	0.00	0.11	0.40	0.38	0.44
		23	0.98	0.03	0.78	0.43	0.33	0.35
	20	12	0.99	0.06	0.70	0.51	0.33	0.22
		13	1.00	0.00	0.10	0.25	0.13	0.25
		23	1.00	0.10	0.76	0.50	0.46	0.23
	30	12	1.00	0.02	0.74	0.47	0.27	0.26
		13	1.00	0.03	0.08	0.39	0.03	0.03
		23	1.00	0.00	0.86	0.50	0.34	0.33
Moderate	10	12	0.83	0.00	0.59	0.55	0.39	0.30
		13	1.00	0.00	0.21	0.36	0.52	0.75
		23	1.00	0.00	0.42	0.45	0.39	0.45
	20	12	1.00	0.00	0.54	0.35	0.25	0.21
		13	1.00	0.00	0.12	0.45	0.01	0.75
		23	1.00	0.00	0.73	0.49	0.28	0.31
	30	12	1.00	0.00	0.54	0.35	0.32	0.33
		13	1.00	0.00	0.12	0.45	0.18	0.19
		23	1.00	0.00	0.73	0.49	0.36	0.41
Strong	10	12	0.92	0.00	0.28	0.48	0.45	0.42
		13	1.00	0.00	0.36	0.45	0.58	0.61
		23	0.99	0.00	0.12	0.41	0.36	0.47
	20	12	1.00	0.00	0.42	0.57	0.72	0.50
		13	1.00	0.00	0.40	0.45	0.92	0.99
		23	1.00	0.00	0.39	0.61	0.55	0.65
	30	12	1.00	0.00	0.26	0.32	0.42	0.35
		13	1.00	0.00	0.31	0.43	0.61	1.00
		23	1.00	0.00	0.44	0.55	0.47	0.79