Estimating Cognitive Diagnosis Models in Small Samples: Bayes Modal Estimation and Monotonic Constraints

Wenchao Ma

Zhehan Jiang

Citation of published version:
Ma, W., & Jiang, Z. (2020). Estimating Cognitive Diagnosis Models in Small Samples: Bayes Modal Estimation and Monotonic Constraints. In Applied Psychological Measurement (Vol. 45, Issue 2, pp. 95–111). SAGE Publications. https://doi.org/10.1177/0146621620977681

Article

Applied Psychological Measurement
2021, Vol. 45(2) 95–111

DOI: 10.1177/0146621620977681
journals.sagepub.com/home/apm
SAGE

Estimating Cognitive Diagnosis Models in Small Samples: Bayes Modal Estimation and Monotonic Constraints

Wenchao Ma[1], and Zhehan Jiang[2]

Abstract

Despite the increasing popularity, cognitive diagnosis models have been criticized for limited utility for small samples. In this study, the authors proposed to use Bayes modal (BM) estimation and monotonic constraints to stabilize item parameter estimation and facilitate person classification in small samples based on the generalized deterministic input noisy "and" gate (G-DINA) model. Both simulation study and real data analysis were used to assess the utility of the BM estimation and monotonic constraints. Results showed that in small samples, (a) the G-DINA model with BM estimation is more likely to converge successfully, (b) when prior distributions are specified reasonably, and monotonicity is not violated, the BM estimation with monotonicity tends to produce more stable item parameter estimates and more accurate person classification, and (c) the G-DINA model using the BM estimation with monotonicity is less likely to overfit the data and shows higher predictive power.

Keywords

cognitive diagnosis, diagnostic classification, EM algorithm, Bayes modal, monotonic constraints, G-DINA

Introduction

Cognitive diagnosis models (CDMs) have been developed recently and, unlike classical test theory and unidimensional item response models, CDMs aim to pinpoint whether students have already mastered a set of skills or attributes. This has the potential to provide finer grained information about students' strengths and weaknesses, and thus could be used to facilitate classroom instruction and learning. In addition to the increasing applications in educational assessments (e.g., Ma et al., 2020; Wu, 2019), CDMs have also been used in psychology for psychological disorder diagnosis (e.g., de la Torre et al., 2018) and personnel selection (Sorrel et al., 2016).

[1]The University of Alabama, Tuscaloosa, USA
[2]Peking University, Beijing, China

Corresponding Authors:
Wenchao Ma, The University of Alabama, Box 870231, Tuscaloosa, AL 35487, USA.
Email: wenchao.ma@ua.edu
Zhehan Jiang, Institute of medical Education, Peking University, 5 Yiheyuan Rd, Beijing 100871, Haidian District, China.
Email: jiangzhehan@gmail.com

Despite the rising popularity, CDMs have been criticized for limited utility for small samples (e.g., Henson, 2009; Sessoms & Henson, 2018). This could be attributed to various reasons, but one is the unsatisfactory performance of existing algorithms for model parameter estimation in small samples. As noted by Jiang and Ma (2018), parameters of parameteric CDMs are usually estimated using either the Markov chain Monte Carlo (MCMC) method or the expectation-maximization (EM; Dempster et al., 1977) algorithm, both of which, however, have some limitations. On one hand, the MCMC method (e.g., Culpepper & Hudson, 2018; Jiang & Carter, 2019; Zhan et al., 2019) tends to be very time-consuming, which makes it less applicable in real settings. On the other hand, although the EM algorithm is relatively fast for moderate number of attributes and theoretically guaranteed to converge to local maxima, it has been observed to fail to converge sometimes in practice. For example, Templin and Bradshaw (2014) conducted a simulation study on log-linear cognitive diagnostic model (LCDM) using Mplus, which implements the EM algorithm, and found that of 500 replications in each condition, only 330 to 447 converged successfully. It can be expected that the nonconvergence issue would be more severe when sample size is small. Also, the EM algorithm has been observed to produce boundary or near-boundary solutions (e.g., Ma & Guo, 2019). Chiu et al. (2018) noted that the EM algorithm may fail to ''produce reasonable parameter estimates when samples are small'' as in their real data analysis many estimates of success probabilities were either 0 or 1. Furthermore, the boundary solutions pose a challenge to inferences. For example, obtaining standard errors of boundary solutions could be challenging or even impossible, as noted by Ma and Guo (2019) and Philipp et al. (2018), and in turn, affects various hypothesis testing procedures for model comparison, differential item functioning detection and Q-matrix validation using the Wald and score tests (e.g., Ma & de la Torre, 2019; Sorrel et al., 2017).

Researchers have attempted to improve the performance of CDMs in small samples from several aspects. First, some researchers (Culpepper, 2015; Culpepper & Hudson, 2018; Zhang & Culpepper, 2017) have derived the Gibbs sampling methods for several CDMs, which tends to be much faster than a commonly employed Metropolis–Hastings (MH) algorithm (da Silva et al., 2018). Jiang and Carter (2019) investigated using a Hamilitonian procedure for estimating the LCDM, which is potentially more efficient than the MH and Gibbs samplers. Despite these advanced developments, the MCMC algorithms are still rather slow compared with the EM algorithm. The second direction that researchers have explored is nonparametric approaches (Chiu et al., 2009, 2018). Chiu et al. (2018) proposed a general nonparametric classification (GNPC) method, which is suitable for data conforming to a variety of parameteric forms. The GNPC approach has been shown to produce higher classification accuracy than the parametric CDMs estimated using the EM algorithm for small samples (Chiu et al., 2018), but it also has several limitations. For example, it only focuses on person classification and cannot be used to examine item properties, and psychometric tools for assessing the classification accuracy based on the GNPC method are lacking.

Unlike the aforementioned two directions, this study intends to address the issue of CDM applications in small samples by improving the performance of the EM algorithm. The challenges of the EM algorithm in small samples, as discussed above, do not only pertain to parametric CDMs, but also occur in many other modern psychometric models. For example, extremely large or implausible estimates for item response theory (IRT) models (Mislevy, 1986) and boundary solutions for latent class models (Garre & Vermunt, 2006) are often observed when the EM algorithm is employed in small samples. It has been well recognized that when sample size is small or data do not contain sufficient information, the use of prior distribution may become important for inferences (e.g., Gelman, 2002). The prior information about model parameters could be incorporated into the EM algorithm straightforwardly, which is usually referred to as Bayes modal (BM) estimation or posterior mode estimation (McLachlan &

Krishnan, 2008). The BM estimation can be viewed as a computationally efficient variant of the MCMC algorithm. The BM estimation has been used for calibrating complex IRT models such as the three-parameter logistic model (Birnbaum, 1968), and Garre and Vermunt (2006) showed that the BM estimation could provide more accurate parameter estimation than the EM algorithm when some parameters were close to the boundary in latent class models. In CDMs, DeCarlo (2011) considered making use of the BM estimation to avoid the boundary problems for the estimation of joint attribute distribution parameters when analyzing C. Tatsuoka's (2002) fraction subtraction data using the deterministic input noisy ''and'' gate (DINA) model.

The current study, however, focuses on the use of BM algorithm for the estimation of item parameters of the generalized deterministic input noisy ''and'' gate (G-DINA) model (de la Torre, 2011). The G-DINA model is considered in this study because it is one of the most general CDMs with many applications. The item parameter estimates of the G-DINA model have played a central role in assessing the quality of items, detecting potential misspecifications in the Q-matrix (de la Torre & Chiu, 2016), and comparing the G-DINA model with many reduced models it subsumes (de la Torre & Lee, 2013; Ma et al., 2016), and thus, the estimation precision is critical.

De la Torre (2011) derived the closed-form solutions for estimating item parameters of the G-DINA model, which, however, cannot accommodate monotonic constraints. Monotonicity is a fundamental assumption for many psychometric approaches such as the Mokken scale (Sijtsma & van der Ark, 2017) and most IRT models (De Ayala, 2013). In CDMs, the monotonic constraints are said to be satisfied for an item if mastering an additional required attribute would not yield a lower probability of success. The monotonicity is based on an ''intuitively plausible assumption'' (Rupp et al., 2010, p. 121), and imposing such constraints would be theoretically reasonable in many CDM applications. However, when analyzing real data using the G-DINA model, researchers sometimes chose not to impose such constraints because of the additional efforts required in model calibration (e.g., Sorrel et al., 2016). By doing so, the G-DINA model may produce parameter estimates that are hard to interpret (Chiu et al., 2018; Sorrel et al., 2016).

In sum, this study aims to investigate whether, or to what extent, imposing priors and monotonic constraints on item parameters could improve person classification accuracy for CDMs in small samples. The remainder of this article is laid out as follows. The next section provides an overview of the G-DINA model. In section ''BM Estimation and Monotonicity Constraints,'' the BM estimation and the monotonicity for the G-DINA model were introduced. Section ''Simulation Study'' describes in detail a simulation study for evaluating the utilities of priors and monotonic constraints in small samples. Then, a set of data was analyzed to further illustrate their performance in practice. The authors conclude in section ''Summary and Discussion'' with a brief summary as well as a discussion of directions for future studies.

An Overview of the G-DINA Model

Suppose a test with J items measures K binary attributes. The association between items and attributes is specified in a $J \times K$ Q-matrix (K. K. Tatsuoka, 1983), with element $q_{jk} = 1$ indicating item j measures attribute k and $q_{jk} = 0$ indicating item j does not measure attribute k. For notational simplicity, the first K_j^* attributes are assumed to be required for item j, and $\boldsymbol{\alpha}_{lj}^* = (\alpha_{lj1}, \ldots, \alpha_{ljK_j^*})^\top$ is the reduced attribute vector consisting of the columns of the required attributes, where $l = 1, \ldots, 2^{K_j^*}$. Let Y_{ij} be a response variable of individual i to item j, following a Bernoulli distribution. The success probability of individuals with attribute profile $\boldsymbol{\alpha}_{lj}^*$ on item j is denoted as $P(\boldsymbol{\alpha}_{lj}^*) = P(Y_{ij} = 1 | \boldsymbol{\alpha}_{lj}^*)$, and based on the identity link G-DINA model:

$$P\left(\boldsymbol{\alpha}_{lj}^{*}\right)=\delta_{j0}+\sum_{k=1}^{K_j^{*}}\delta_{jk}\alpha_{ljk}+\sum_{k'=k+1}^{K_j^{*}}\sum_{k=1}^{K_j^{*}-1}\delta_{jkk'}\alpha_{ljk}\alpha_{ljk'}+\cdots+\delta_{j12\cdots K_j^{*}}\prod_{k=1}^{K_j^{*}}\alpha_{ljk}, \tag{1}$$

where δ_{j0} is the intercept for item j, δ_{jk} is the main effect due to attribute k, $\delta_{jkk'}$ is the interaction effect due to attributes k and k', and $\delta_{j12\ldots K_j^{*}}$ is the interaction effect due to all required attributes. De la Torre *(2011)* showed that Equation 1 can also be written as the following:

$$\boldsymbol{P}_j=\boldsymbol{M}_j\boldsymbol{\delta}_j, \tag{2}$$

where $\boldsymbol{\delta}_j=(\delta_{j0},\ldots,\delta_{j12\ldots K_j^{*}})^{\top}$, $\boldsymbol{P}_j=(P(\boldsymbol{\alpha}_{1j}^{*}),\ldots,P(\boldsymbol{\alpha}_{2^{K_j^{*}}j}^{*}))^{\top}$, and $\boldsymbol{M}_j$ is a $2^{K_j^{*}}\times 2^{K_j^{*}}$ design matrix. Because $\boldsymbol{M}_j$ is an invertable square matrix, either $\boldsymbol{P}_j$ or $\boldsymbol{\delta}_j$ can be treated as item parameters. In this study, $\boldsymbol{P}_j$ was focused, which has elements ranging between 0 and 1 with straightforward interpretations.

BM Estimation and Monotonicity Constraints

Based on the EM algorithm, de la Torre (2011) showed that the maximum likelihood estimate of $P(\boldsymbol{\alpha}_{lj}^{*})$ can be expressed as:

$$\hat{P}\left(\boldsymbol{\alpha}_{lj}^{*}\right)=\frac{r_{jl}}{n_l}, \tag{3}$$

where n_l and r_{jl} represent the expected number of individuals in latent group l and the expected number of individuals in latent group l who answer item j correctly, respectively. Because $P(\boldsymbol{\alpha}_{lj}^{*})$ represents the probability of individuals in latent group l answering item j correctly, the natural conjugate prior is the Beta distribution, Beta(β_1,β_2). As shown in the Online Appendix, the BM estimate of $P(\alpha_{lj}^{*})$ can be expressed as:

$$\hat{P}\left(\boldsymbol{\alpha}_{lj}^{*}\right)=\frac{r_{jl}+(\beta_1-1)}{n_l+(\beta_1+\beta_2-2)}. \tag{4}$$

Note that the closed-form solution above was derived when the monotonicity is not assumed. The monotonicity of success probabilities of different latent groups can be represented using a Hasse diagram of a power set (Koshy, 2004), with an example displayed in Figure 1 for an item measuring three attributes. In the Hasse diagram, rounded rectangles represent latent groups (labeled by the corresponding reduced attribute profiles) and lines are used to connect latent groups that monotonic constraints should be imposed to. Specifically, when two latent groups are connected, the one in a higher position should not have a lower item success probability. Thus, the number of lines in a Hasse diagram represents the number of monotonic constraints for an item. From Figure 1, it is straightforward to identify the following constraints for monotonicity:

$$\begin{aligned}
&P(000)\leq P(100),P(010),P(001)\\
&P(100)\leq P(110),P(101)\\
&P(010)\leq P(110),P(011)\\
&P(001)\leq P(101),P(011)\\
&P(111)\geq P(110),P(101),P(011).
\end{aligned} \tag{5}$$

Note that some inequality constraints are not given in Equation 5 because they are implied. For example, it is obvious that $P(000)\leq P(110)$.

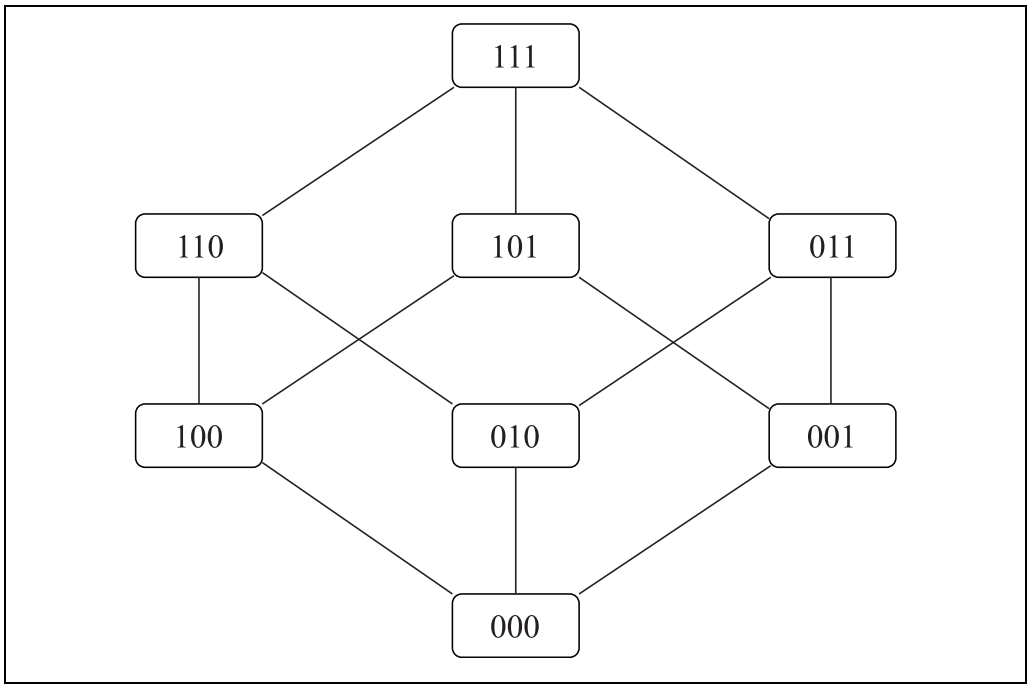

Figure 1. A Hasse diagram for monotonic constraints for an item measuring three attributes.

To accommodate monotonic constraints, Hong et al. (2016) proposed to adjust the parameter estimates of the G-DINA model based on the EM algorithm. However, their adjustments are implemented in a post hoc manner and thus is deemed statistically suboptimal. In this study, the constraints were accommodated in the process of model calibration using some optimization routines in M-step of the EM algorithm. For optimization purpose, the constraints in a Hasse diagram can be represented using a constraint matrix $\boldsymbol{C}$, with the number of rows equal to the number of constraints and the number of columns equal to the number of latent groups. The monotonic constraints in Figure 1 can be written by:

$$\boldsymbol{CP} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \times \begin{bmatrix} P(000) \\ P(100) \\ P(010) \\ P(001) \\ P(110) \\ P(101) \\ P(011) \\ P(111) \end{bmatrix} \geq \mathbf{0}. \tag{6}$$

Many optimization routines can be employed to accommodate the monotonic constraints, and in this study, the sequential quadratic programming (Kraft, 1988) implemented in R package nloptr was used (Johnson, 2019).

Simulation Study

Design

The simulation study intends to investigate the performance of the G-DINA model using different estimation algorithms, including the EM algorithm, EM algorithm with monotonic constraints, BM algorithm, and BM algorithm with monotonic constaints. The GNPC method serves as a benchmark because of its promising performance in small samples (Chiu et al., 2018). To compare the performance of different estimation algorithms, three factors were manipulated:

Sample size (N). Although most studies in CDMs used large samples of size more than 1,000 (Sessoms & Henson, 2018), this study focuses on samples of small to moderate sizes because CDMs are usually believed to be most useful in these settings. In the literature, Sessoms and Henson (2018) referred samples of between 50 and 150 as ''relatively small'' (p. 9). For studies with a focus on small samples, Chiu et al. (2018) considered samples of 30 to 500 and Chang et al. (2018) considered samples of 30, 50, and 100. This study considers samples of $30, 50, 100, 200,$ and 500, where, for discussion purpose, $N \leq 100$ represents small sample sizes and $N = 200$ or 500 represents moderate sizes.

Test length (J). As noted by Nájera et al. (2019), previous simulation studies in CDMs usually involved tests with 11 to 30 items. For CDM applications, Jurich and Bradshaw (2013) reported a diagnostic test with 17 items and Ma et al. (2020) reported a diagnostic test with 30 items. Therefore, this study considers two levels of test lengths: $J = 15$ and 30.

Attribute correlation (R). It has been observed that attributes tend to be highly correlated in retrofitting applications (Sessoms & Henson, 2018), but a recent study found that attributes in a diagnostic assessment had correlations ranging between .07 and .95 (Ma et al., 2020). To consider varied levels of attribute associations, and to be consistent with previous simulation studies, three levels of attribute correlation were considered, namely, $R = .3$, .5, and .8, representing weak, moderate, and strong attribute associations, respectively.

Regarding item parameters, like J. Chen (2017) and Liu et al. (2009), the authors simulated $P(\mathbf{0})$ and $1 - P(\mathbf{1})$ from $U(0, 0.3)$. For items measuring two or more attributes, the success probabilities for individuals who master some but not all required attributes were generated from uniform distribution $U(P(\mathbf{0}), P(\mathbf{1}))$ subject to the monotonic constraints. The number of attributes was fixed at 5, which is in line with existing studies (e.g., J. Chen & de la Torre, 2013; Ma et al., 2016). Attribute patterns were generated by dichotomizing a set of multivariate latent abilities from the multivariate normal threshold model (Chiu et al., 2009). Specifically, for individual i, $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK})^\top$ is randomly generated from $MVN(\mathbf{0}, \Sigma)$, where the diagonal elements of Σ were 1s and the off-diagonal elements were equal to R. With $\boldsymbol{\theta}_i$, $\alpha_{ik} = 1$ if $\theta_{ik} \geq \Phi^{-1}(\frac{k}{K+1})$ and 0 otherwise. The Q-matrix used in J. Chen et al. (2013) was used in this study, which ensures all attributes were measured the same number of times. The Q-matrix can be found in the Online Appendix. Based on the G-DINA model, 200 data sets were generated under each condition.

For each replication, the G-DINA model was fit to the data using different algorithms. To simplify the presentation, the EM algorithm with monotonic constraints and BM algorithm with monotonic constaints were abbreviated to the EM + M and BM + M algorithms, respectively. For the BM and BM + M algorithms, Beta (1.5, 2.5) was used as the prior distribution of $P(\mathbf{0})$, Beta (2.5, 1.5) the prior distribution of $P(\mathbf{1})$, and Beta (2, 2) the prior distribution of $P(\boldsymbol{\alpha}_{lj}^* \notin \{\mathbf{0}, \mathbf{1}\})$. The code for estimating the G-DINA model using different estimation algorithms was written in R programming language (R Core Team, 2017) and can be downloaded

from https://osf.io/z9kxe/. These algorithms are also available in the latest version of the GDINA R package (Ma & de la Torre, 2020).

Criteria

First, person parameter recovery was evaluated using the proportion of correctly classified attribute vectors (PCV) defined as $\frac{1}{T\times N}\sum_{t=1}^{T}\sum_{i=1}^{N} I^{(t)}_{[\boldsymbol{\alpha}_i=\hat{\boldsymbol{\alpha}}_i]}$, where $I^{(t)}_{[\boldsymbol{\alpha}_i=\hat{\boldsymbol{\alpha}}_i]}$ is an indicator variable with an outcome of 1 if the estimated attribute vector matches the true for the tth replication and 0 otherwise. Second, to evaluate the recovery of item parameters of the G-DINA model, the bias and absolute bias were calculated. Let θ^{t} and $\hat{\theta}^{\mathrm{t}}$ represent an item success probability parameter[1] and its estimate in tth replication under a certain condition, respectively. The bias and absolute bias were defined as $\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}(\hat{\theta}^{t}-\theta^{t})$ and $\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}|\hat{\theta}^{t}-\theta^{t}|$, respectively. The mean bias and mean absolute bias were reported by averaging the biases and absolute biases across all parameters and all replications. To better understand the results, four-way (Sample size $\times$ Test length $\times$ Attribute correlation $\times$ Method) analyses of variance (ANOVAs) were performed for different criteria. To examine the sizes of different effects, the generalized η^2, denoted by η^2_G, was calculated as the effect size measure, which was recommended to be used in mixed ANOVA (Bakeman, 2005; Olejnik & Algina, 2003). An effect is said to be nontrivial or essentially meaningful when $\eta^2_G \geq .01$, according to the guideline in Cohen (1988, p. 287).

Results

Before evaluating the performance of different estimation algorithms, it was examined whether the G-DINA model converged under all conditions. Table 1 gives the proportion of replications that the G-DINA model failed to converge[2] under varied conditions. Note that the G-DINA model using BM and BM + M algorithms converged under all conditions, so they are not presented in the table. Several findings can be observed from Table 1. First, the G-DINA model with the EM algorithm had higher nonconvergence rates than that with the EM + M algorithm consistently. For example, when $N = J = 30$ and $R = .8$, the G-DINA model with the EM algorithm failed to converge under 23% of replications, whereas the G-DINA model with the EM + M algorithm failed to converge in 7% of replications. Second, the nonconvergence rates tended to increase as sample size decreased, attribute correlation increased, or test lengthened. A close scrutiny of the estimation process revealed that the failure to converge was primarily caused by the fact that n_l in Equation 3 was estimated to be 0. If the G-DINA model with any estimation algorithm failed to converge for a replication, that replication was not included for further analysis to ensure a fair comparison of all estimation algorithms in terms of parameter recovery.

Attribute classification. With PCV as the dependent variable, mixed ANOVA showed that in addition to four main effects, two two-way interactions had nontrivial effects (i.e., Sample size $\times$ Method with $\eta^2_G = .08$ and Attribute correlation $\times$ Method with $\eta^2_G = .02$). Several findings can be observed from the interaction plots in Figure 2. Regarding different algorithms for the G-DINA model, the BM + M algorithm consistently performed the best and the EM algorithm the worst in terms of PCV, especially when sample size was small. For example, when $N = 30$, the BM + M algorithm produced an averaged PCV of .591, and the EM algorithm produced an averaged PCV of .514. In contrast, when $N = 500$, the BM + M algorithm produced an averaged PCV of .751 and the EM algorithm an averaged PCV of .731. Second, the BM algorithm outperformed the EM + M algorithm when $N \leq 200$, but produced similar averaged PCVs when $N = 500$.

Table 1. The Proportion of Replications Where the G-DINA Model Failed to Converge.

		G-DINA with EM algorithm			G-DINA with EM + M algorithm		
J	*N*	*R* = .3	*R* = .5	*R* = .8	*R* = .3	*R* = .5	*R* = .8
15	30	0.015	0.030	0.045		0.010	0.005
	50	0.005	0.010	0.015			0.015
	100	0.005	0.005	0.010		0.005	
	200		0.005			0.005	
	500						
30	30	0.140	0.170	0.230	0.025	0.045	0.070
	50	0.110	0.175	0.215	0.030	0.020	0.045
	100	0.030	0.080	0.200	0.010	0.020	0.005
	200	0.015	0.050	0.050			0.010
	500			0.040			

Note. Zero elements were removed for readability. *J* = test length; *N* = sample size; *R* = attribute correlation; G-DINA = generalized deterministic input noisy and gate; EM = expectation-maximization.

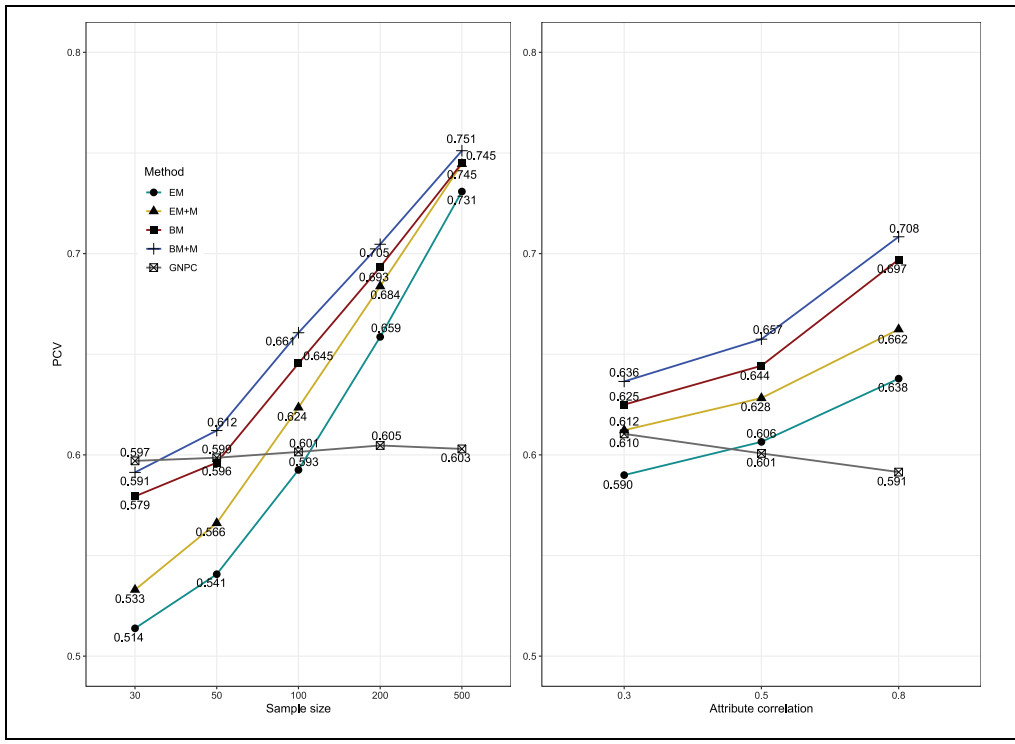

Figure 2. Two-way interactions of Sample size × Method and Attribute correlation × Method with PCV as the dependent variable.
Note. PCV = proportion of correctly classified attribute vectors.

Figure 2 reveals that the classification accuracy of the GNPC method was not affected by sample size (e.g., the averaged PCV = .597 when $N = 30$ and .603 when $N = 500$), but the classification accuracy of the G-DINA model improved as sample size increased (e.g., the averaged

PCVs ranged between .514 and .591 when $N = 30$ and between .731 and .751 when $N = 500$). Also, as attribute correlations became stronger, the G-DINA model provided higher classification accuracy in terms of PCV, but the GNPC yielded slightly lower classification accuracy.

Concerning the comparison between the G-DINA model and GNPC method, it can be observed from Figure 2 that the GNPC method had higher classification accuracy than the G-DINA model when sample size was small but the G-DINA model gradually outperformed the GNPC method in terms of PCV as sample size increased. Specifically, when $N = 30$, the GNPC outperformed the G-DINA model, regardless of the estimation method used. When $N = 50$, the G-DINA model with the BM + M algorithm outperformed the GNPC method in terms of averaged PCV. When $N = 100$, the G-DINA model with all algorithms except the EM algorithm produced higher averaged PCVs than the GNPC method. When $N \geq 200$, the G-DINA model consistently outperformed the GNPC approach. Finally, although it is not displayed in Figure 2, test length had a large main effect (i.e., $\eta_G^2 = .67$) and the classification accuracy increased as test lengthened.

Mean bias of item parameter estimates of the G-DINA model. Mixed ANOVA showed that all factors had nontrivial but small main effects for the bias of item parameters (η_G^2 ranged between .01 and .04) and that none of the interaction effects were practically meaningful. Specifically, the EM and EM + M algorithms produced slightly less biased item parameter estimates than the BM and BM + M algorithms (i.e., mean biases for EM and EM + M were $-.02$ and for BM and BM + M were $-.03$). This is not unexpected because the BM methods are known to produce biased estimates by pulling estimates toward the mean of the prior distribution. In addition, the mean biases of item parameter estimates decreased as sample size increased, test lengthened, or attribute correlations became weaker.

Mean absolute bias of item parameters of the G-DINA model. Mixed ANOVA showed that the main effects of all factors as well as several two-way interactions were nontrivial for the mean absolute bias. The two-way interaction effects involving the Method factor were displayed in Figure 3, and several findings can be observed. First, the EM algorithm consistently produced highest mean absolute biases of item parameter estimates, whereas the BM + M algorithm, the smallest mean absolute biases. The BM algorithm performed better than the EM + M algorithm when $N \leq 100$ in terms of the mean absolute bias, but both algorithms performed similarly when $N \geq 200$. Second, all algorithms produced smaller mean absolute biases as sample size increased, test length increased, or attribute association strengthened.

Real Data Illustration

For illustrative purposes, a set of data collected from a learning experiment at the University of Tuebingen in Germany in 2010 was analyzed in this study. The test consists of 12 items in elementary probability theory, which measured four attributes. Responses of 504 participants from the first part of the experiment were used. The same data have been previously analyzed by Ma and de la Torre (2020) and Philipp et al. (2018). The data and Q-matrix, as well as other relevant information, are available from the R package pks (Heller & Wickelmaier, 2013).

The G-DINA model was first fit to the data using the EM, EM + M, BM, and BM + M algorithms. The estimated item success probabilities and delta parameters of all items can be found in Online Appendix. Figure 4 gives the estimated success probabilities of Item 12 based on different estimation algorithms as an example. It can be observed that some of the estimates based on the EM and BM algorithms violated the assumption of monotonicity. For example, $P(101)$ was estimated to be less than $P(000)$ based on the EM algorithm and less than $P(100)$ based on the BM algorithm. The estimates conformed to the assumption of monotonicity when

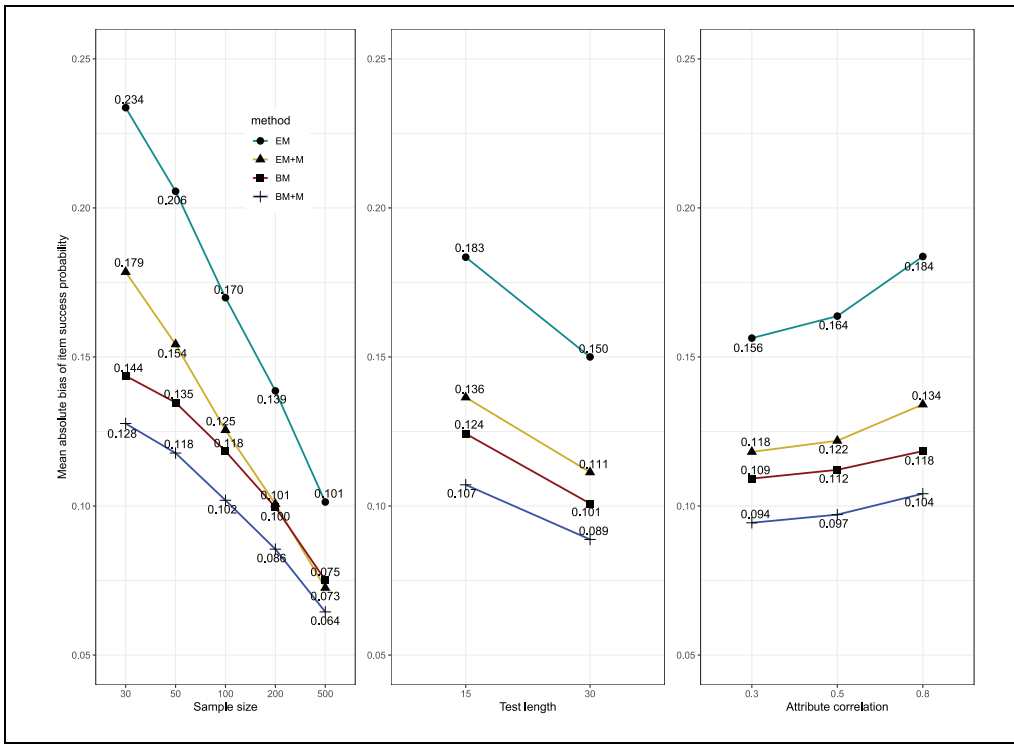

Figure 3. Two-way interactions with mean absolute bias as the dependent variable.

monotonic constraints were imposed. In addition, the EM algorithm produced boundary solutions for some item parameters such as $P(100)$, $P(010)$, $P(001)$, and $P(101)$, but the BM and BM + M algorithms avoided boundary solutions. The EM + M algorithm did not produce boundary solutions in this study, but may not be able to avoid boundary estimates in other data sets based on the authors' experience. It can also be observed that the estimated success probabilities based on the BM and BM + M algorithms for latent groups mastering one or two attributes tended to be higher than those based on the EM and EM + M algorithms due to the priors imposed.

Boundary or near-boundary solutions may indicate model overfitting, which occurs when the parameters were estimated such that the model fits the data too closely. To assess the accuracy or error of the G-DINA model's predictions based on different estimation algorithms, the k-fold cross-validation, a well-known resampling approach in supervised learning, was used in this study. The k-fold cross-validation first randomly splits the sample into k equally sized folds, and each time, $k-1$ folds are used as the training sample for model calibration, and the remaining fold is used for model validation (i.e., test sample). This process is repeated k times, each time using a different fold as the test sample. In particular, for the mth replication, let Y_m^{training} denote the training sample and $\boldsymbol{Y}_m^{\text{test}}$ the test sample. Also, let $\hat{\boldsymbol{\theta}}_m^{(s)}$ represent the item and mixing proportion parameters estimated using estimation algorithm s, which could be the EM, EM + M, BM, or BM + M algorithm, based on $\boldsymbol{Y}_m^{\text{training}}$. Furthermore, let $\ell[\hat{\boldsymbol{\theta}}_m^{(s)}|Y_m^{\text{training}}]$ and $\ell[\hat{\boldsymbol{\theta}}_m^{(s)}|Y_m^{\text{test}}]$ denote the marginalized log likelihood of observing the training sample and the test sample, respectively, given the set of parameter estimates $\hat{\boldsymbol{\theta}}_m^{(s)}$ from the training sample. Due to the fact that the sample is divided into k folds, one can define $\bar{\ell}_s^{\text{training}} = \frac{1}{k}\sum_{m=1}^{k} \ell[\hat{\boldsymbol{\theta}}_m^{(s)}|\boldsymbol{Y}_m^{\text{training}}]$

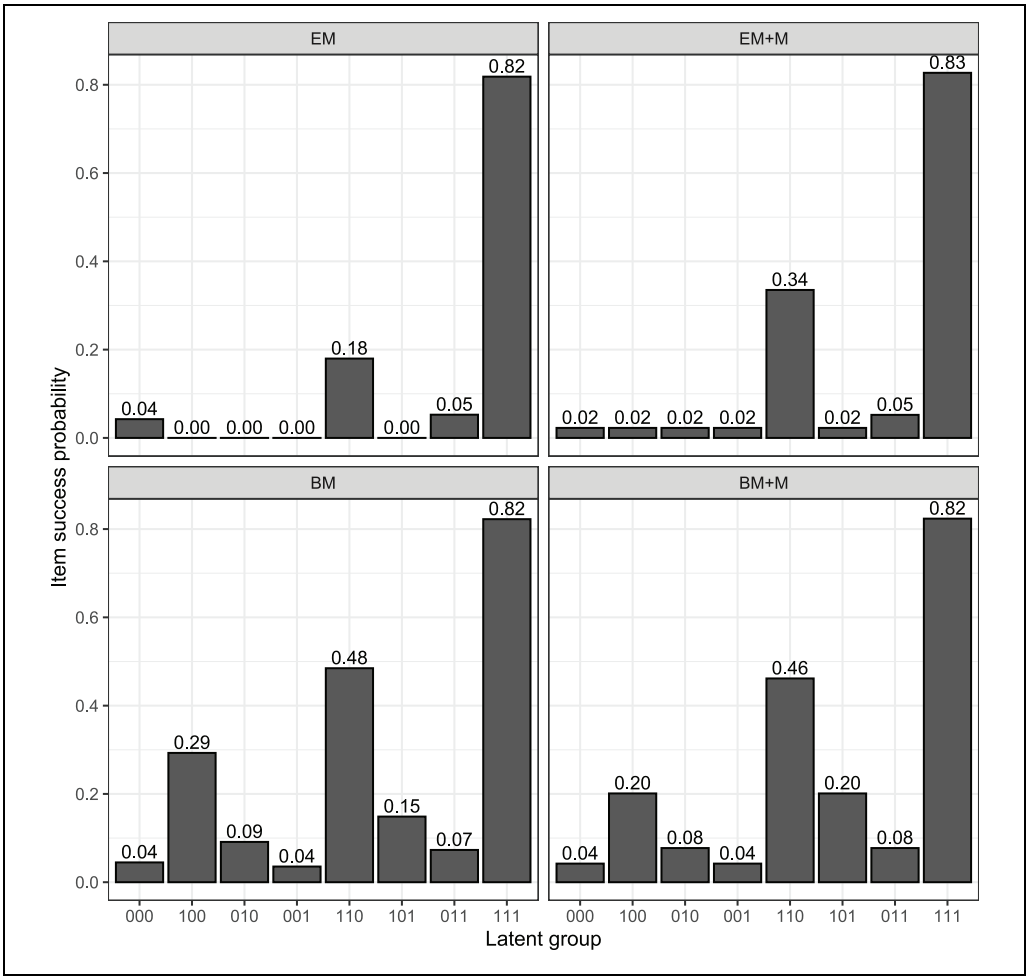

Figure 4. Estimated success probabilities for Item 12 based on different algorithms.

and $\bar{\ell}_s^{\text{test}} = \frac{1}{k}\sum_{m=1}^{k} \ell[\hat{\boldsymbol{\theta}}_m^{(s)} | \boldsymbol{Y}_m^{\text{test}}]$. The latter is also known as the log predictive score (Gelman et al., 2013; Smyth, 2000). A higher value of $\bar{\ell}_s^{\text{training}}$ indicates that the estimation algorithm s produced a set of parameters that can better describe the training data, and a higher value of $\bar{\ell}_s^{\text{test}}$ suggests that the estimation algorithm produced a set of parameters that have higher predictive power for the test data.

Determining the value of k for a k-fold cross-validation is not trivial, and researchers usually suggest 5 or 10 folds (Kuhn & Johnson, 2013, p. 70). In this study, 2-, 5-, and 10-fold cross-validation was considered, each of which was repeated 50 times. The averaged $\bar{\ell}_s^{\text{training}}$ and $\bar{\ell}_s^{\text{test}}$ across all 50 repetitions were reported in Table 2. When comparing parameters estimated based on four different estimation algorithms, it can be observed that averaged $\bar{\ell}_{\text{EM}}^{\text{training}}$>averaged $\bar{\ell}_{\text{EM+M}}^{\text{training}}$>averaged $\bar{\ell}_{\text{BM}}^{\text{training}}$>averaged $\bar{\ell}_{\text{BM+M}}^{\text{training}}$ and that averaged $\bar{\ell}_{\text{EM}}^{\text{test}}$<averaged $\bar{\ell}_{\text{EM+M}}^{\text{test}}$<averaged $\bar{\ell}_{\text{BM}}^{\text{test}}$<averaged $\bar{\ell}_{\text{BM+M}}^{\text{test}}$. These findings suggest that the parameter estimates of the EM algorithm can best capture the characteristics of the training data, but failed to perform on the test data as well as other algorithms. In other words, the EM algorithm not only captured the pattern

Table 2. Average $\bar{\ell}_s^{\text{training}}$ and $\bar{\ell}_s^{\text{test}}$ Based on *k*-Fold Cross-Validations.

	Average $\bar{\ell}_s^{\text{test}}$				Average $\bar{\ell}_s^{\text{training}}$			
k	EM	EM + M	BM	BM + M	EM	EM + M	BM	BM + M
2	−1,321.9	−1,274.9	−1,259.1	**−1,257.8**	**−1,193.7**	−1,197.9	−1,203.4	−1,204.6
5	−507.3	−501.5	−498.5	**−498.1**	**−1,932.7**	−1,937.3	−1,941.0	−1,941.9
10	−251.8	−250.3	−249.0	**−248.7**	**−2,179.2**	−2,183.5	−2,186.8	−2,187.5

Note. Largest values are in bold. EM = expectation-maximization; BM = Bayes modal.

of the training data but also took the random fluctuations in the data into account, and thus overfit the data. In contrast, other algorithms, most notably the BM + M approach, reduced the risk of overfitting by imposing additional constraints. As a result, they produced better predictive performance on the test data.

To further examine the impact of sample size on person classifications, 500 samples were randomly selected from the original data under each of the four sample size conditions: $n = 30, 50, 100$, and 200. Individuals in each sample were drawn without replacement. For each of those selected samples, individuals were classified using the G-DINA model and the GNPC method. Based on a classification method s, which could be the EM, EM + M, BM, BM + M, or the GNPC approach here, the estimated attribute profile of individual i in a selected sample is denoted by $\tilde{\boldsymbol{\alpha}}_i^s$ and the estimated attribute profile of this individual based on the full sample of size 504 is denoted by $\hat{\boldsymbol{\alpha}}_i^s$. The authors defined a measure of classification agreement by $\frac{1}{n}\sum_{i=1}^{n} I[\tilde{\boldsymbol{\alpha}}_i^s = \hat{\boldsymbol{\alpha}}_i^s]$, with $I[\cdot]$ being 1 when $\tilde{\boldsymbol{\alpha}}_i^s$ is the same as $\hat{\boldsymbol{\alpha}}_i^s$ and 0 otherwise. Note that here, we implicitly treat the classifications based on the full sample as benchmarks. A higher value of classification agreement does not necessarily mean the corresponding classification method is better in terms of the accuracy of classifications, but does imply that the method provides more stable and consistent person classifications across varied sample sizes.

The classification agreements for different estimation methods were given in Table 3. Although 500 samples were drawn under each sample size condition, some were removed because the G-DINA model using a certain algorithm failed to converge. The number of valid samples under each condition was given in Table 3 too. It can be observed that as n increased, the classification agreements increased for the G-DINA model, regardless of the estimation method used. Nevertheless, compared with the BM and BM + M algorithms, the EM and EM + M algorithms were more sensitive to sample sizes. For example, when n decreased from 200 to 30, the classification agreement decreased from .90 to .75 for the EM algorithm, from .91 to .76 for the EM + M algorithm, from .94 to .86 for the BM algorithm, and from .93 to .85 for the BM + M algorithm. This is not unexpected, given that the prior information of item parameters could play an important role when the data have limited information (i.e., sample size is small). A perplexing finding, however, is that the classification agreement improved as sample size increased for the GNPC method, though the simulation study showed that the classification accuracy of the GNPC method is not affected by sample size. When comparing the GNPC method with the G-DINA model, it can also be observed that the GNPC method produced higher classification agreement than the EM algorithm did for small samples but similar classification agreement for samples of moderate size. Nevertheless, the BM and BM + M algorithms consistently yielded higher classification agreement than the GNPC method, regardless of the sample size.

Table 3. Person Classification Agreement Among Different Methods.

n	Number of valid replications	EM	EM + M	BM	BM + M	GNPC
30	385	0.75	0.76	0.86	0.85	0.83
50	434	0.80	0.80	0.87	0.87	0.84
100	479	0.85	0.85	0.91	0.91	0.86
200	491	0.90	0.91	0.94	0.93	0.89

Note. EM = expectation-maximization; BM = Bayes modal; GNPC = general nonparametric classification.

Summary and Discussion

In response to the criticism of limited utility for small samples of CDMs, this study systematically examined the challenges the EM algorithm faces. Specifically, the results of the simulation study, which echoed the findings of previous studies, show that the G-DINA model based on the EM algorithm may fail to converge in small samples. The real data analysis also shows that the G-DINA model based on the EM algorithm tend to overfit the data in small samples and thus performed poorly in new samples.

This study examined whether incorporating priors and monotonic constraints into the EM algorithm could facilitate parameter estimation. By incorporating prior information into model calibration, the BM and BM + M algorithms are analogous with the MCMC algorithm. Compared with the MCMC algorithm, the BM and BM + M algorithms only provide point estimation of parameters of interest, but the computational time is typically much shorter. For the real data calibration in the previous section, the G-DINA model with the EM, EM + M, BM, and BM + M algorithms all converged in 2 s. It is observed that when items conform to monotonicity, imposing monotonic constraints and priors could improve the accuracy of person classification, especially when sample size is small, and imposing monotonic constraints and priors also prevents overfitting, at least to some extent, and thus demonstrates better predictive power in new samples.

Despite the potential advantages, monotonic constraints and priors need to be imposed with caution in practice. For one thing, the findings of the simulation study were observed when data were assumed to conform to the monotonicity, which appear reasonable but may not always be the case. de la Torre and Sorrel (2017) proposed an effect size measure to quantify the size of the monotonicity violation and also examined whether the likelihood ratio test could be used to determine the violation of monotonicity empirically. In addition to their approaches, model-data fit at both test and item level may be used to assess whether the model with monotonicity can fit data adequately. For another, the hyperparameters of the prior distributions used in this study may not be appropriate under some circumstances. A body of research has showed that priors could have a substantial impact on the estimation of model parameters (e.g., van Erp et al., 2018), and carelessly selected priors could produce misleading or even erroneous results, especially when sample size is small. In the CDM context, for example, if the item response models can be reasonably assumed to be conjunctive or disjunctive in nature, the priors should be adjusted to reflect this belief accordingly.

In addition, this study focuses on the identity link G-DINA model, but it can be expected that the nonconvergence and overfitting issues of the EM algorithm also pertain to other widely used CDMs. Therefore, examining the impact of imposing priors and monotonic constraints to other CDMs in small samples is needed. In addition, the Q-matrix is assumed to be known and correctly specified in this study, but it is unclear about the impact of misspecifications in the Q-matrix on the performance of the G-DINA model with the EM + M, BM, and BM + M

algorithms. It is worth mentioning that a variety of approaches have been developed to estimate the Q-matrix either with or without experts' input (e.g., Y. Chen, Culpepper, Chen, & Douglas, 2018; Y. Chen et al., 2015, 2020; Culpepper, 2019), which could be used when Q-matrix is not available or is potentially misspecified. Finally, this study only considers improving parameter estimation of CDMs based on cross-sectional data. It is possible that in small-scale educational programs, diagnostic assessments are administered multiple times, producing longitudinal data. An array of CDMs have been developed for analyzing longitudinal data (e.g., Y. Chen & Culpepper, 2020; Y. Chen, Culpepper, Wang, & Douglas, 2018; Kaya & Leite, 2017; Wang, Yang, Culpepper, & Douglas, 2018; Wang, Zhang, Douglas, & Culpepper, 2018; Zhan, 2020), and future research may explore if monotonic constraints and priors could be used in conjunction with these models.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Zhehan Jiang https://orcid.org/0000-0002-1376-9439

Supplemental Material

Supplementary material is available for this article online.

Notes

1. In this study, the bias and absolute bias of both item success probability and δ parameters in Equation 1 were examined, but due to the limit of space, the results on δ parameters were only presented in Online Appendix. The bias and absolute bias of δ parameters were averaged across all replications and parameter types (i.e., intercept, main effect, and interaction denoted by $\delta_{\text{intercept}}$,$\delta_{\text{maineffect}}$, and $\delta_{\text{interaction}}$, respectively).
2. A calibration is said to fail to converge when the expectation-maximization (EM) algorithm terminated with an error or when the number of EM iterations was 2,000, and the maximum absolute difference in item parameters between the last two iterations was greater than .0001.

References

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Chang, Y.-P., Chiu, C.-Y., & Tsai, R.-C. (2018). Nonparametric CAT for CD in educational settings with small samples. *Applied Psychological Measurement*, *43*(7), 543–561.

Chen, J. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement*, *41*, 277–293.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, *37*, 419–437.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*, 123–140.

Chen, Y., Culpepper, S., & Liang, F. (2020). A sparse latent class model for cognitive diagnosis. *Psychometrika*, *85*(1), 121–153. https://doi.org/10.1007/s11336-019-09693-2

Chen, Y., & Culpepper, S. A. (2020). A multivariate probit model for learning trajectories: A fine-grained evaluation of an educational intervention. *Applied Psychological Measurement*, *44*(7–8), 5151–530. https://doi.org/10.1177/0146621620920928

Chen, Y., Culpepper, S. A., Chen, Y., & Douglas, J. (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika*, *83*, 89–108. https://doi.org/10.1007/s11336-017-9579-4

Chen, Y., Culpepper, S. A., Wang, S., & Douglas, J. (2018). A hidden Markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied Psychological Measurement*, *42*(1), 5–23.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850–866.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633–665. https://doi.org/10.1007/s11336-009-9125-0

Chiu, C.-Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika*, *83*(2), 355–375.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press.

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics*, *40*(*5*), 454–476. https://doi.org/10.3102/1076998615595403

Culpepper, S. A. (2019). Estimating the cognitive diagnosis Q matrix with expert knowledge: Application to the fraction-subtraction dataset. *Psychometrika*, *84*(2), 333–357.

Culpepper, S. A., & Hudson, A. (2018). An improved strategy for Bayesian estimation of the reduced reparameterized unified model. *Applied Psychological Measurement*, *42*(2), 99–115.

da Silva, M. A., de Oliveira, E. S., von Davier, A. A., & Bazán, J. L. (2018). Estimating the DINA model parameters using the No-U-Turn sampler. *Biometrical Journal*, *60*(2), 352–368.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Press.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*(1), 8–26. https://doi.org/10.1177/0146621610377081

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273. https://doi.org/10.1007/s11336-015-9467-8

de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*, 355–373. https://doi.org/10.1111/jedm.12022

de la Torre, J., & Sorrel, M. A. (2017, July 18–21). *Attribute classification accuracy improvement: Monotonicity constraints on the G-DINA model* [Conference session]. Annual Meeting of Psychometric Society, Zurich, Switzerland.

de la Torre, J., van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, *51*(4), 281–296. https://doi.org/10.1080/07481756.2017.1327286

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*, 1–38.

Garre, F. G., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika*, *33*, 43–59. https://doi.org/10.2333/bhmk.33.43

Gelman, A. (2002). Prior distribution. *Encyclopedia of Environmetrics*, *3*(*4*), 1634–1637.

Gelman, A., Hwang, J., & Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Heller, J., & Wickelmaier, F. (2013). Minimum discrepancy estimation in probabilistic knowledge structures. *Electronic Notes in Discrete Mathematics*, *42*, 49–56.

Henson, R. A. (2009). Diagnostic classification models: Thoughts and future directions. *Measurement: Interdisciplinary Research and Perspectives*, *7*, 34–36.

Hong, C.-Y., Chang, Y.-W., & Tsai, R.-C. (2016). Estimation of generalized DINA model with order restrictions. *Journal of Classification*, *33*, 460–484.

Jiang, Z., & Carter, R. (2019). Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnosis model via Stan. *Behavior Research Methods*, *51*, 651–662.

Jiang, Z., & Ma, W. (2018). Integrating differential evolution optimization to cognitive diagnostic model estimation. *Frontiers in Psychology*, *9*, Article 2142.

Johnson, S. G. (2019). *The NLopt nonlinear-optimization package* (version 1.2.2) [Computer software]. https://CRAN.R-project.org/package=nloptr

Jurich, D. P., & Bradshaw, L. P. (2013). An illustration of diagnostic classification modeling in student learning outcomes assessment. *International Journal of Testing*, *14*(1), 49–72. https://doi.org/10.1080/15305058.2013.835728

Kaya, Y., & Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: An evaluation of model performance. *Educational and Psychological Measurement*, *77*(3), 369–388.

Koshy, T. (2004). *Discrete mathematics with applications*. Elsevier.

Kraft, D. (1988). *A software package for sequential quadratic programming* (Technical report DFVLR-FB 88–28). Institut fuer Dynamik der Flugsysteme.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement*, *33*, 579–598.

Ma, W., & de la Torre, J. (2019). An empirical Q-matrix validation method for the sequential G-DINA model. *British Journal of Mathematical and Statistical Psychology*, *73*(*1*), 142–163. https://doi.org/10.1111/bmsp.12156

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*(14). https://doi.org/10.18637/jss.v093.i14

Ma, W., & Guo, W. (2019). Cognitive diagnosis models for multiple strategies. *The British Journal of Mathematical and Statistical Psychology*, *72*(2), 370–392.

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*, 200–217. 10.1177/0146621615621717

Ma, W., Minchen, N., & de la Torre, J. (2020). Choosing between cdm and unidimensional IRT: The proportional reasoning test case. *Measurement: Interdisciplinary Research and Perspectives*, *18*, 87–96.

McLachlan, G., & Krishnan, T. (2008). *The EM algorithm and extensions* (2nd ed.). John Wiley & Sons.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(*2*), 177–195.

Nájera, P., Sorrel, M. A., & Abad, F. J. (2019). Reconsidering cutoff points in the general method of empirical Q-matrix validation. *Educational and Psychological Measurement*, *79*(4), 727–753. https://doi.org/10.1177/0013164418822700

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434–447.

Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics*, *43*, 88–115.

R Core Team. (2017). *R: A language and environment for statistical computing* [Computer software manual]. https://www.R-project.org/

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

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*, 1–17.

Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 137–158.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, *10*(*1*), 63–72. https://doi.org/10.1023/a:1008940618127

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*, *41*(8), 614–631. https://doi.org/10.1177/0146621617707510

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, *19*, 506–532. https://doi.org/10.1177/1094428116630065

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.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

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

van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods*, *23*, 363–388.

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden markov model with covariates. *Journal of Educational and Behavioral Statistics*, *43*(1), 57–87.

Wang, S., Zhang, S., Douglas, J., & Culpepper, S. (2018). Using response times to assess learning progress: A joint model for responses and response times, *16*, 45–58.

Wu, H.-M. (2019). Online individualised tutor for improving mathematics learning: A cognitive diagnostic model approach. *Educational Psychology*, *39*(10), 1218–1232.

Zhan, P. (2020). A Markov estimation strategy for longitudinal learning diagnosis: Providing timely diagnostic feedback. *Educational and Psychological Measurement*, *80*(6), 1145–1167. https://doi.org/10.1177/0013164420912318

Zhan, P., Jiao, H., Man, K., & Wang, L. (2019). Using JAGS for bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics*, *44*(*4*), 473–503.

Zhang, S., & Culpepper, S. A. (2017, April). *Bayesian estimation of a general class of identified restricted latent class models* [Conference session]. Annual Meeting of the National Council on Measurement in Education, San Antonio, TX, United States.