JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME measurement in education
Check for updates
DOI: 10.1111/jedm.12383 *Journal of Educational Measurement*

A Dual-Purpose Model for Binary Data: Estimating Ability and Misconceptions

Wenchao Ma
The University of Alabama
Miguel A. Sorrel
Universidad Autonoma de Madrid
Xiaoming Zhai
University of Georgia
Yuan Ge
College Board

Most existing diagnostic models are developed to detect whether students have mastered a set of skills of interest, but few have focused on identifying what scientific misconceptions students possess. This article developed a general dual-purpose model for simultaneously estimating students' overall ability and the presence and absence of misconceptions. The expectation-maximization algorithm was developed to estimate the model parameters. A simulation study was conducted to evaluate to what extent the parameters can be accurately recovered under varied conditions. A set of real data in science education was also analyzed to examine the viability of the proposed model in practice.

Misconceptions, also known as alternative conceptions, refer to the naïve ideas or intuitive knowledge that learners develop when explaining or predicting phenomena (Confrey, 1990). Specifically, misconceptions are commonly discovered in mathematics and science education research, where students were found to hold preexisting inconsistent ideas and then generate incorrect understandings, concepts, or skills when explaining phenomena (Vosniadou, 2020). In addition, misconceptions hinder the acquisition of new expertise and the conceptualization of advanced knowledge (Thompson & Logue, 2006). Hence, the goal of science or mathematics learning for all children can be partially achieved by investigating the conceptual basis of misconceptions and facilitating students' learning progression (Li & Li, 2008; Zhai & Li, 2021).

Previous researchers have documented a variety of psychometrics approaches for identifying misconceptions (Bradshaw & Templin, 2014; DiBello et al., 2015; Kuo et al., 2018; Kuo et al., 2016; Ozaki et al., 2020) to assist teachers and other educators in implementing practical curriculum constructions as well as classroom instructions. For example, the rule space method (Tatsuoka, 2009) represents early work in psychometrics for dealing with misconceptions (e.g., Gao et al., 2020). Along this line, several diagnostic models have been recently developed and they can be grouped into two categories: models for multiple-choice (MC) items with some distractors associated with misconceptions and models for dichotomous data. Examples of cognitive diagnosis models (CDMs) in the former category include the scaling individuals and classifying misconceptions (SICM) model (Bradshaw & Templin, 2014), the

generalized diagnostic classification model (GDCM; DiBello et al., 2015), and the MC-M-DINO models (Ozaki et al., 2020). The MC-M-DINO models only consider misconceptions when defining item response functions, but the SICM and GDCM also take students' overall ability or misconceptions into account. As they currently stand, all three models involve complex formulations, and their parameters can only be estimated using the Markov chain Monte-Carlo (MCMC) method, which tends to be time-consuming. Examples of CDMs in the second category include the BUG deterministic inputs, noisy "or" gate (Bug-DINO; Kuo et al., 2016) model, and the model for simultaneously identifying skills and misconceptions (SISM; Kuo et al., 2018). Both models were developed to analyze dichotomous data, which may be obtained from MC items without coded distractors. Because of the simplicity of their formulations, their parameters can be estimated using the expectation-maximization (EM) algorithm, which is usually more efficient than the MCMC method.

Despite their usefulness, existing models for misconceptions are not without limitations. First, some models are designed for MC items with coded options but developing items of this type tends to be time-consuming and error prone. In addition, MC items may not always be the best item format, and researchers and practitioners may prefer questions of other types, such as fill-in-the-blank or constructed response items. In these cases, no distractors are explicitly identified. Second, other models, such as the Bug-DINO and the models by Ozaki et al. (2020), assume that misconceptions are the only person-related factor affecting item responses and ignore the impact of students' abilities or skills. This assumption seems to be very restrictive in most practical settings.

To overcome the limitations of existing approaches, we develop a novel psychometric model—a general dual-purpose model (GDPM) for binary data (i.e., correct vs. incorrect)—to estimate students' overall abilities and misconceptions at the same time. Compared with existing methods, the proposed GDPM has three salient features. First, we consider the interaction of persons' overall ability and misconceptions when defining the item response function, and the resulting GDPM is very general, and subsumes some existing models like Bug-DINO and SICM for binary data.[1] Second, the GDPM estimates both students' ability and misconceptions simultaneously, enabling the ranking of students based on their overall ability and providing diagnostic information for targeted remediation. Third, the GDPM offers flexibility by relaxing the orthogonal assumption between ability and misconceptions, a constraint in some existing models like SICM, and thus can accommodate a broader range of relationships among latent variables. Last, we develop the expectation-maximization algorithm for estimating the parameters of the GDPM, which is much faster than the MCMC algorithm adopted by some of the existing models, like the SICM and GDCM.

The Model Formulation

Let J be the number of items, K the number of misconceptions, and N the number of students in a random sample taking the test. Also, let $\mathbf{X}_i$ be the response vector of student i to J items, with X_{ij} being the jth element. Student i's ability was denoted by

θ_i, which was assumed to be a discrete latent variable with $H+1$ ordinal levels (i.e., $\theta_i \in \{0,\ldots,H\}$) as in Woodruff and Hanson (1996). The profile of misconceptions, or bugs, associated with student i is denoted by $\boldsymbol{\alpha}_i = (\alpha_{i1},\ldots,\alpha_{iK})$, where $\alpha_{ik} = 1$ if student i has the kth misconception and 0 otherwise. The probability that student i responds to item j correctly, denoted by $P(X_{ij} = 1|\theta_i,\boldsymbol{\alpha}_i)$, is assumed to be a function of the student's ability, misconception profile, and some key features of the item characterized by item parameters. To simplify the notation, let us assume item j is related to the first K_j misconceptions. The general dual-purpose model (GDPM) for estimating students' abilities and misconceptions is defined as

$$\log\left[\frac{P\left(X_{ij}=1|\theta_i,\boldsymbol{\alpha}_i\right)}{P\left(X_{ij}=0|\theta_i,\boldsymbol{\alpha}_i\right)}\right] = \delta_{j0} + \delta_{j1}\theta_i g\left(\boldsymbol{\alpha}_i\right) + h\left(\boldsymbol{\alpha}_i\right), \tag{1}$$

where the term $g(\boldsymbol{\alpha}_i)$ is a function of misconception profiles for controlling whether the overall ability affects item success probability for students under each misconception profile. The term $h(\boldsymbol{\alpha}_i)$ defines how misconceptions affect an item's success probabilities and can take various forms.

This study focuses on two special cases of the GDPM, each with a unique assumption about the relationship between overall ability, misconception profiles, and item performance.

We first consider a disjunctive rule for misconceptions such that possessing any misconception involved in item j leads to a low chance of success, regardless of the overall ability. The resulting model is referred to as the disjunctive-misconception GDPM, or DM-GDPM, which, for item j, is defined with

$$h\left(\boldsymbol{\alpha}_i\right) = \phi_j\left[1 - \prod_{k=1}^{K_j}\left(1-\alpha_{ik}\right)\right] \tag{2}$$

and

$$g\left(\boldsymbol{\alpha}_i\right) = \prod_{k=1}^{K_j}\left(1-\alpha_{ik}\right). \tag{3}$$

As a result, the item response function (IRF) of the DM-GDPM can be written by

$$\log\left[\frac{P\left(X_{ij}=1|\theta_i,\boldsymbol{\alpha}_i\right)}{P\left(X_{ij}=0|\theta_i,\boldsymbol{\alpha}_i\right)}\right] = \begin{cases}\delta_{j0}+\delta_{j1}\theta_i & \text{if } \boldsymbol{\alpha}_i = \mathbf{0},\\ \delta_{j0}+\phi_j & \text{otherwise.}\end{cases} \tag{4}$$

The DM-GDPM has three parameters for each item: parameter δ_{j0} is the log odds of success for students who possess none of the measured misconceptions and $\theta = 0$, and δ_{j1} is the increase in log odds of success for every one score increase in θ for those students who possess none of the measured misconceptions. For those who possess any of the measured misconceptions, $\delta_{j0} + \phi_j$ is the log odds of success to item j.

Although the assumption of DM-GDPM that the possession of any misconception degrades item performance to the lowest level is plausible, it is equally, if not more, likely that the possession of each misconception has a unique and distinct impact on

item performance. Therefore, we also define an additive-misconception GDPM, or AM-GDPM, with $g(\boldsymbol{\alpha}_i) = 1$ and

$$h(\boldsymbol{\alpha}_i) = \sum_{k=1}^{K_j} \phi_{jk}\alpha_{ik}. \tag{5}$$

The IRF of the AM-GDPM can then be written by

$$\log\left[\frac{P\left(X_{ij}=1|\theta_i,\boldsymbol{\alpha}_i\right)}{P\left(X_{ij}=0|\theta_i,\boldsymbol{\alpha}_i\right)}\right] = \delta_{j0} + \delta_{j1}\theta_i + \sum_{k=1}^{K_j}\phi_{jk}\alpha_{ik}. \tag{6}$$

The AM-GDPM assumes each misconception affects the item performance independently and separately. The AM-GDPM has $K_j + 2$ parameters, and ϕ_{jk}, which is assumed to be negative, is the impact of misconception k on item j.

For illustration, Figure 1 shows the IRFs of AM- and DM-GDPMs for an item involving two misconceptions with hypothetical parameters (for AM-GDPM, $\delta_0 = -3$, $\delta_1 = 1$, $\phi_1 = -1$, and $\phi_2 = -5$; for DM-GDPM, $\delta_0 = -3$, $\delta_1 = 1$, and $\phi_1 = -1$). It can be observed that the DM-GDPM divides all students into two groups: (1) those who exhibit at least one misconception measured by the item are in a homogenous group with the same probability of success regardless of their ability levels and (2) those who do not exhibit any measured misconceptions are in a heterogeneous group with the probability of success only depending on students' ability levels. In contrast, the AM-GDPM divides students into four groups, each with a unique misconception profile. It can also be observed that students without any misconceptions outperform students with some misconceptions in general, and students with two misconceptions have the lowest success probabilities. The impacts of the two misconceptions differ, as evidenced by the separate trace lines for 01 and 10 misconception profiles.

Model Estimation

The GDPM involves two types of parameters: structural and incidental parameters, where the number of incidental parameters increases in conjunction with sample size, but the number of structural parameters does not. The structural parameters of the GDPM, denoted by $\boldsymbol{\gamma}$, can be estimated using the marginal maximum likelihood estimation via the expectation-maximization algorithm (Bock & Aitkin, 1981). In particular, the EM algorithm consists of two steps: the expectation (E) step and the maximization (M) step. In the E-step, we calculate the so-called Q-function (Dempster et al., 1977), which is the expected log-likelihood of the complete data conditional on the observed data and current parameter estimates. With independent individuals, the Q-function takes the following form:

$$\begin{aligned} Q(\boldsymbol{\gamma}|\boldsymbol{\gamma}') = &\sum_{j=1}^{J}\sum_{\theta=0}^{H}\sum_{c=1}^{2^K}\left[r_{j\theta c}\log P(X_j=1|\theta,\boldsymbol{\alpha}_c) + (n_{\theta c} - r_{j\theta c})\log\left[1 - P(X_j=1|\theta,\boldsymbol{\alpha}_c)\right]\right] \\ &+ \sum_{\theta=0}^{H}\sum_{c=1}^{2^K} n_{\theta c}\log(\pi_{\theta c}), \end{aligned} \tag{7}$$

where $n_{\theta c}$ is the expected number of students with ability θ and misconception profile $\boldsymbol{\alpha}_c$ and $r_{j\theta c}$ is the expected number of students with ability θ and misconception profile $\boldsymbol{\alpha}_c$ who answer item j correctly. In addition, K is the number

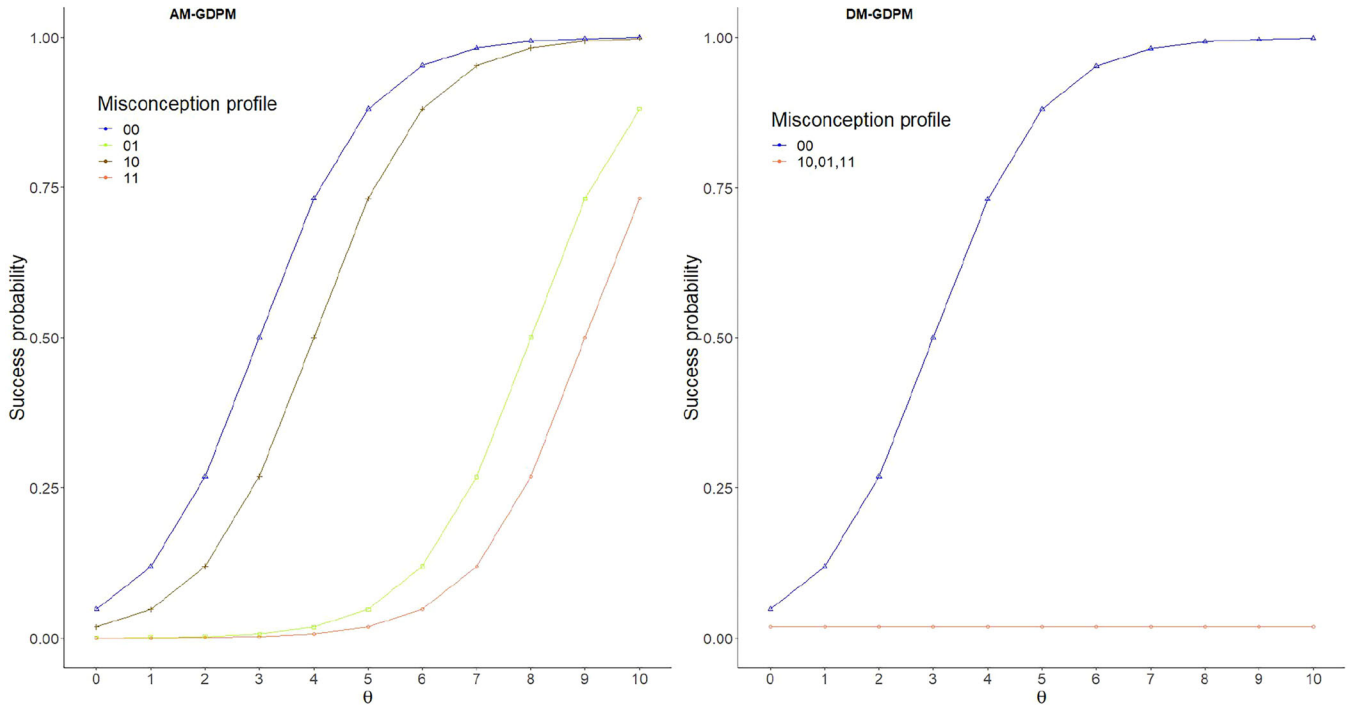

Figure 1. Item response functions of different GDPMs using hypothetical parameters. [Color figure can be viewed at wileyonlinelibrary.com]

of misconceptions, and $\pi_{\theta c}$ is the proportion of students with ability θ, and misconception profile $\boldsymbol{\alpha}_c$ in the population. In the M-step, the Q-function is maximized with respect to parameter $\boldsymbol{\gamma}$, which includes item parameters and $A \times (2^K - 1)$ parameters for the joint distribution of ability and misconceptions. The E- and M-steps are repeated until some convergence criteria are met. Although many approaches can be used to model the relationship between students' abilities and misconceptions, such as the higher-order method (e.g., de la Torre & Douglas, 2004; Ma, 2022), this paper adopts the multinomial model that estimates the proportion of each ability and misconception profile because of its flexibility and generality. The covariances between ability and misconceptions are not modeled as parameters but can still be obtained from the estimates of students' profiles. Students' ability and misconception profiles are estimated afterward by treating the estimated structural parameters as true, using the marginal modal assignment (Ma, 2022), which is equivalent to the expected a posteriori (EAP) method for binary latent variables when .5 is used as the cutoff. Specifically, for person i, we have $\hat{\alpha}_{ik} = \underset{\alpha_k \in \{0,1\}}{\text{argmax}} P(\alpha_k | \boldsymbol{X}_i)$ and $\widehat{\theta}_i = \underset{\theta \in \{0,1,\ldots,H\}}{\text{argmax}} P(\theta | \mathbf{X}_i)$, where

$$\begin{aligned} P(\alpha_k|\mathbf{X}_i) &= \sum_{\theta=0}^{H} \sum_{\alpha_1=0}^{1} \cdots \sum_{\alpha_{k-1}=0}^{1} \sum_{\alpha_{k+1}=0}^{1} \cdots \sum_{\alpha_K=0}^{1} P(\theta, \alpha_1, \ldots, \alpha_k, \ldots, \alpha_K|\mathbf{X}_i) \\ &= \sum_{\theta=0}^{H} \sum_{\alpha_1=0}^{1} \cdots \sum_{\alpha_{k-1}=0}^{1} \sum_{\alpha_{k+1}=0}^{1} \cdots \sum_{\alpha_K=0}^{1} \frac{P(\mathbf{X}_i|\theta,\alpha_1,\ldots,\alpha_k,\ldots,\alpha_K)}{P(\mathbf{X}_i)} p(\theta, \alpha_1, \ldots, \alpha_k, \ldots, \alpha_K) \end{aligned} \tag{8}$$

and $P(\theta|\mathbf{X}_i) = \sum_{c=1}^{2^K} P(\theta, \boldsymbol{\alpha}_c|\mathbf{X}_i) = \sum_{c=1}^{2^K} \frac{P(\mathbf{X}_i|\theta,\boldsymbol{\alpha}_c)}{P(\mathbf{X}_i)} p(\theta, \boldsymbol{\alpha}_c)$.

Simulation Study

A simulation study was conducted to evaluate the performance of the EM algorithm in estimating model parameters of the two special cases proposed for the general dual-purpose model (AM- and DM-GDPMs) under varied conditions.

Method

Factors. Three factors were manipulated: sample size, attribute correlation, and generating model. Levels of the factors were chosen to reflect realistic scenarios based on the existing literature and the empirical applications.

***Sample size* (N).** This study considers four levels of sample sizes: $N = 500$, 1,000, 2,000, and 4,000. These levels were chosen according to Sessoms and Henson (2018), where the mean and median of sample sizes of 36 CDM applications were 1,788 and 1,255, respectively, and 30% of them involved samples of 2,000 or more.

***Attribute correlation* (R).** Previous studies showed that attributes could have varied levels of associations. For example, Ma et al. (2020) showed that the correlations between attributes measured in a diagnostic test ranged from .07 to .95. The strengths of associations between misconceptions, however, remain unclear. The results from the real data analysis in the next section showed that the absolute values of

Table 1
Q-Matrix for Simulation Study

Item	M1	M2	M3	M4	Item	M1	M2	M3	M4
7	1	0	0	0	19	1	1	0	0
8	0	1	0	0	20	1	0	1	0
9	0	0	1	0	21	1	0	0	1
10	0	0	0	1	22	0	1	1	0
11	1	0	0	0	23	0	1	0	1
12	0	1	0	0	24	0	0	1	1
13	0	0	1	0	25	1	1	0	0
14	0	0	0	1	26	1	0	1	0
15	1	0	0	0	27	1	0	0	1
16	0	1	0	0	28	0	1	1	0
17	0	0	1	0	29	0	1	0	1
18	0	0	0	1	30	0	0	1	1

Note: Items 1 to 6 were omitted because they do not measure any misconceptions. M1–M4: four misconceptions.

correlations between misconceptions ranged from .1 to .7. As a result, we considered three levels of correlations in this study: $R = .3, .5, .8$, representing low, moderate, and high correlations, respectively. Note that we assume misconceptions are positively correlated but have negative correlations with the overall ability. This is a plausible assumption and is also in line with the findings from the real data analysis in the next section.

Generating models (M). In this study, we considered two generating models—the AM—and DM-GDPMs for data generation. The same models were used to fit the data.

In addition to the manipulated factors above, we fixed the following factors to make the simulations manageable. In particular, the number of misconceptions was fixed at $K = 4$, the same as the real data analysis in the next section. Test length was fixed at 30, which has been considered in previous simulation studies (Ma & Jiang, 2021; Nájera et al., 2020) and is also in line with the diagnostic test reported by Ma et al. (2020) The Q-matrix is given in Table 1. This Q-matrix contains six items measuring none of the misconceptions, 12 measuring a single misconception, and 12 measuring two misconceptions. Each misconception is measured by nine items.

Data generation. The item parameters were informed by the results from the real data analysis in the next section. In particular, the success probability for students with the lowest level of proficiency without possessing misconceptions is denoted by $P(X_j = 1|\theta = 0, \boldsymbol{\alpha} = \mathbf{0})$ and drawn from a truncated $N(.2, .1)$ with values between [.05, .35]. This is a plausible setting, given that all items are MC questions. Also, similar to the real data analysis in the next section, where the probability of success for those who had the highest level of ability yet without any misconceptions had a mean of .96 and a standard deviation of .04, we drew $P(X_j = 1|\theta = 0, \boldsymbol{\alpha} = \mathbf{0})$ from

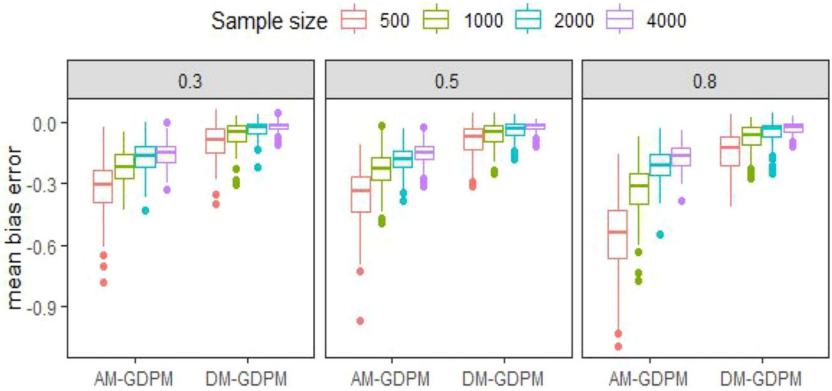

Figure 2. Mean bias errors of item parameter estimates according to sample size and attribute correlations. [Color figure can be viewed at wileyonlinelibrary.com]

truncated $N(.95, .25^2)$ with values between [.9, 1]. In addition, we set $P(X_j = 1|\theta, \boldsymbol{\alpha} = \mathbf{1}) \sim U[0, P(X_j = 1|\theta = 0, \boldsymbol{\alpha} = \mathbf{0})]$. Based on the above settings, all item parameters were calculated directly.

Examinees' person parameters were drawn from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ has diagonal elements of unity. The absolute values of off-diagonal elements of $\boldsymbol{\Sigma}$ are equal to R. The off-diagonal elements are negative when they represent the covariances between the overall ability and misconceptions but positive when they represent the covariances between two misconceptions. The discrete ability and binary misconceptions were obtained by categorizing the multivariates from the distribution. We generated Rep = 200 data sets under each condition to reduce Monte-Carlo sampling errors. The AM- and DM-GDPM were fitted to the simulated data. The estimation code was written in R software (R Core Team, 2021), with various functions from the *GDINA* R package (Ma & de la Torre, 2020b), and can be requested from the corresponding author. The estimation is terminated if the maximum absolute difference in structural parameters between two successive iterations is less than .001 or the number of iterations reaches 2,000.

Dependent variables. To assess the item parameter recovery, we calculated bias and root mean square error (RMSE) for item parameter estimates. To evaluate person parameter recovery, we explored the correlation between the generating ability θ and the estimated ability $\hat{\theta}$ ($r_{\theta\hat{\theta}}$) and the proportions of correctly classified misconception vectors (PCV), defined as $\text{PCV} = \frac{1}{N \times \text{Rep}} \sum_{r=1}^{\text{Rep}} \sum_{n=1}^{N} I(\boldsymbol{\alpha}_i = \widehat{\boldsymbol{\alpha}}_i)$, where $\hat{\boldsymbol{\alpha}}_i$ and $\boldsymbol{\alpha}_i$ are the estimated and true misconception profiles for examinee i, respectively.

Results

Figure 2 shows the mean biases of item parameter estimates under varied conditions. It can be observed that as sample size increased, mean biases for both AM- and DM-GDPM decreased. In particular, the mean biases were –.41, –.26, –.19, and –.16 for samples of size 500, 1,000, 2,000, and 4,000, respectively, using AM-GDPM; and –.11, –.07, −04, and –.03 for samples of size 500, 1,000, 2,000, and 4,000,

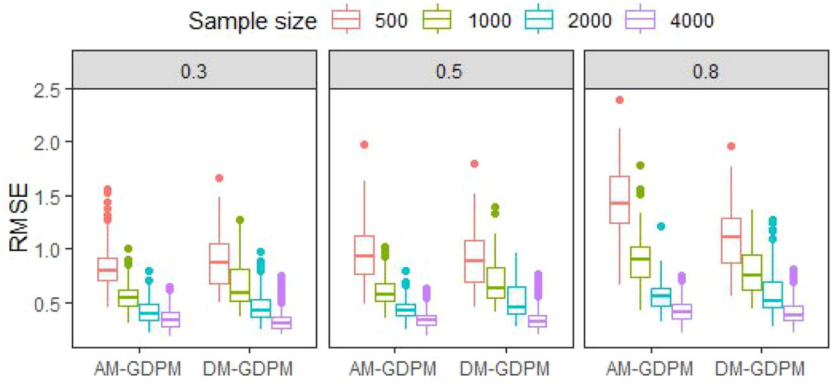

Figure 3. Root mean square errors of item parameter estimates according to model, sample size, and attribute correlations. [Color figure can be viewed at wileyonlinelibrary.com]

respectively, using DM-GDPM. In addition, as attribute correlation increased, both models produced larger biases. In particular, the mean biases were –.22, –.23, and –.32 for attribute correlation of .3, .5, and .8, respectively, using AM-GDPM; and –.06, –.06, and –.08 for attribute correlation of .3, .5, and .8, respectively, using DM-GDPM. Overall, the DM-GDPM had lower mean biases than AM-GDPM (i.e., mean bias = –.26 for AM-GDPM and –.06 for DM-GDPM).

Figure 3 shows the RMSEs of item parameter estimates under varied conditions. Several findings can be observed. First, as sample size increased, the RMSEs for both AM- and DM-GDPM decreased. In particular, the average RMSEs were 1.08, .68, .47, and .36 for samples of size 500, 1,000, 2,000, and 4,000, respectively, using AM-GDPM; and .97, .71, .52, and .36 for samples of size 500, 1,000, 2,000, and 4,000, respectively, using DM-GDPM. Second, as attribute correlation increased, the RMSEs increased for both models. In particular, the average RMSEs were .53, .58, and .83 for attribute correlation of .3, .5, and .8, respectively, using AM-GDPM; and .59, .61, and .72 for attribute correlation of .3, .5, and .8, respectively, using DM-GDPM. Last, the AM-GDPM and DM-GDPM had similar RMSEs in general (i.e., average RMSE = .65 for AM-GDPM and .64 for DM-GDPM).

Figure 4 gives the PCVs of both models under varied simulated conditions. It can be observed that the classification accuracy increased as the sample size increased. In particular, the mean PCVs were .83, .85, .86, and .86 for samples of size 500, 1,000, 2,000, and 4,000, respectively, using AM-GDPM; .80, .82, .83, and .84 for samples of size 500, 1,000, 2,000, and 4,000, respectively, using DM-GDPM. In addition, as attribute correlation increased, classifications became more accurate. Specifically, mean PCVs were .81, .84, and .90 for attribute correlation of low, medium, and high, respectively, using AM-GDPM; and .79, .81, and .87 for attribute correlation of low, medium, and high, respectively, using DM-GDPM. Overall, AM-GDPM and DM-GDPM had similar classification accuracy, with AM-GDPM having slightly higher PCVs (i.e., mean PCV = .85 and .82 for AM- and DM-GDPM, respectively).

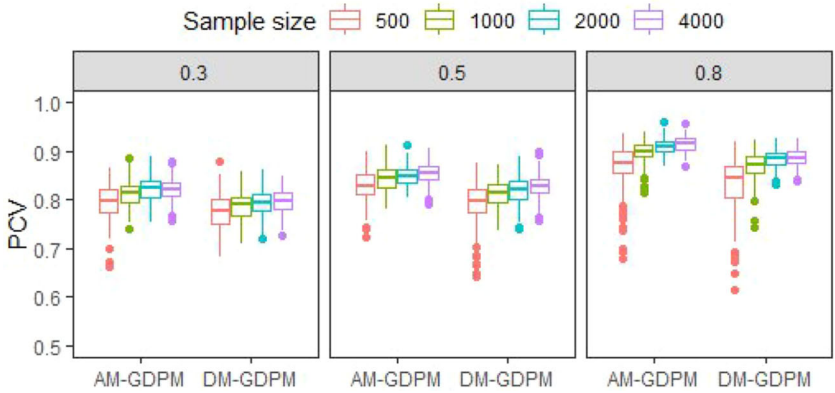

Figure 4. Proportions of correctly classified misconception vectors according to model, sample size, and attribute correlations. [Color figure can be viewed at wileyonlinelibrary.com]

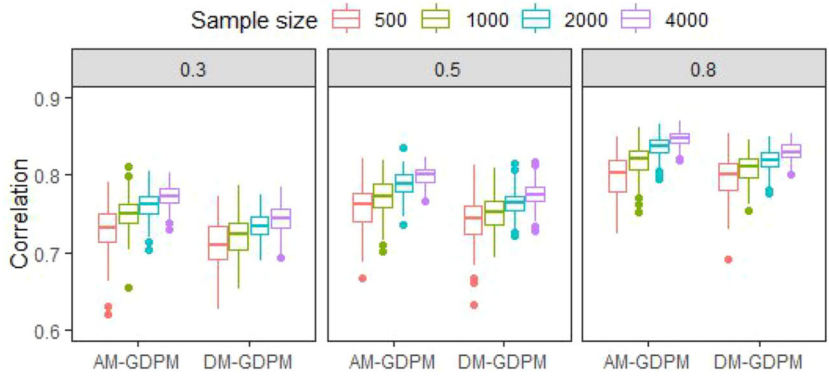

Figure 5. Pearson's correlation for overall ability recovery according to model, sample size, and attribute correlations. [Color figure can be viewed at wileyonlinelibrary.com]

Figure 5 displays the averaged correlations between the estimated and true overall ability. Several findings can be observed. First, the correlation became stronger as the sample size and the strength of attribute correlation increased. In particular, mean correlations were .76, .78, .79, and .81 for samples of size 500, 1,000, 2,000, and 4,000, respectively, using AM-GDPM; and .75, .76, .77, and .78 for samples of size 500, 1,000, 2,000, and 4,000, respectively, using DM-GDPM. In addition, mean correlations were .75, .78, and .83 for attribute correlation of low, medium, and high, respectively, using AM-GDPM; .73, .76, and .81 for attribute correlation of low, medium, and high, respectively, using DM-GDPM. Overall, AM-GDPM and DM-GDPM had similar correlations between the estimated and true overall ability, with AM-GDPM having slightly stronger correlations (i.e., mean correlation = .79 and .77 for AM- and DM-GDPM, respectively).

The results of the simulation study showed that the parameters of both AM-GDPM and DM-GDPM can be recovered reasonably well, especially when sample size was

Table 2
Q-Matrix for Real Data

Item	α_1	α_2	α_3	α_4
1	1	0	0	0
2	1	0	0	0
5	1	0	0	0
7	1	1	0	1
10	0	0	0	1
12	0	1	0	1
13	1	0	0	0
14	0	0	1	0
16	0	0	1	0
17	0	0	1	0
20	0	1	0	0

Note: Items were omitted if they do not measure any misconceptions.

large. Stronger attribute correlation produced less accurately estimated item parameters, but the associations among attributes provide additional information for estimating person attribute profiles and thus lead to more accurate person classifications. Under all simulated conditions, the EM algorithm converged with the set criteria. The average computational time used for parameter estimation was 18.7 seconds with the maximum being 59.9 seconds.

Real Data Analysis

Data

The data comprise item responses of 614 11th-grade Chinese students to 20 MC items in a thermal concept evaluation test in science (Yeo & Zadnik, 2001). Among the 614 students, 329 were male and 285 were female. At the time of testing, students had completed the learning of thermal concepts. Domain experts identified misconceptions involved in each item, within which four misconceptions were considered in this study because they were measured more than three times: (α_1) Objects of different temperatures that are in contact with each other or in contact with air at a different temperature, do not necessarily move toward the same temperature; (α_2) Temperature is a property of a particular material or object; (α_3) Perceptions of hot and cold are unrelated to energy transfer; (α_4) Material has the ability to attract, hold, intensify, or absorb heat and cold. The Q-matrix is given in Table 2, where items that do not measure any misconceptions were omitted.

Analysis

Both AM- and DM-GDPMs were fitted to the data. The absolute model-data fit was evaluated using $r_{jj'}$, which is the absolute difference between the observed and model-implied Fisher-transformed correlations between items j and j', and $l_{jj'}$, the absolute difference between the observed and model-implied log odds ratio between

Table 3
Absolute Fit Statistics

	$r_{jj'}$			$l_{jj'}$		
Model	Max $r_{jj'}$	z	Adj p Value	Max $l_{jj'}$	z	Adj p Value
AM-GDPM	.14	3.34	.16	.87	3.15	.31
DM-GDPM	.13	3.17	.29	.82	2.96	.58

Table 4
Relative Fit Statistics

	Number of Parameters	AIC	BIC	SABIC	CAIC
AM-GDPM	389	**12,286**	**13,909**	**12,675**	**14,298**
DM-GDPM	386	12,354	13,965	12,740	14,351

Note: Smaller values were presented in boldface.

items j and j'. Based on the approximate standard errors of those statistics, z statistics can be obtained for assessing whether the residuals differ significantly from zero. Note that one can calculate $r_{jj'}$ and $l_{jj'}$ for each pair of items, and thus there are $J(J-1)/2$ $r_{jj'}$ and $l_{jj'}$ statistics to examine, where $J = 20$. Like Chen et al. (2013), we focused on the maximum z-scores of these statistics to assess whether the worst-fitting item pair could fit the data well or not. Because this analysis implicitly involved multiple tests, the Holm-Bonferroni procedure was used to control family-wise Type I errors. The relative fit of AM- and DM-GDPMs to the data were compared using various information criteria.

Results

During the model fit analysis of the thermal concept evaluation questions using the GPDMs, we observed that both AM-GDPM and DM-GDPM fit data adequately. Specifically, Table 3 includes the absolute fit statistics from both models, including the absolute difference between the observed and model-implied Fisher-transformed correlations ($r_{jj'}$), and the absolute difference between the observed and model-implied log odds ratio between the item pair of j and j' ($l_{jj'}$). The maximum values of $r_{jj'}$ and $l_{jj'}$ quantified the goodness-of-fit of the worst-fitting item pairs. It can be observed that both models fit the data adequately, so we chose the model that fits the data better according to relative fit statistics. Based on AIC, BIC, SABIC, and CAIC indices, as shown in Table 4, the AM-GDPM was preferred. The following analyses were conducted under the AM-GDPM.

Figure 6 shows the item characteristic curves (ICCs) of the AM-GDPM for four items as an example. The plots were constructed based on the estimated item parameters. The x-axis shows students' ability on a 0 to 20 scale, and the y-axis shows the probability of students correctly answering a specific item. Item 3 does not involve any of the misconceptions, and thus, there is a single curve. Item 10 involves

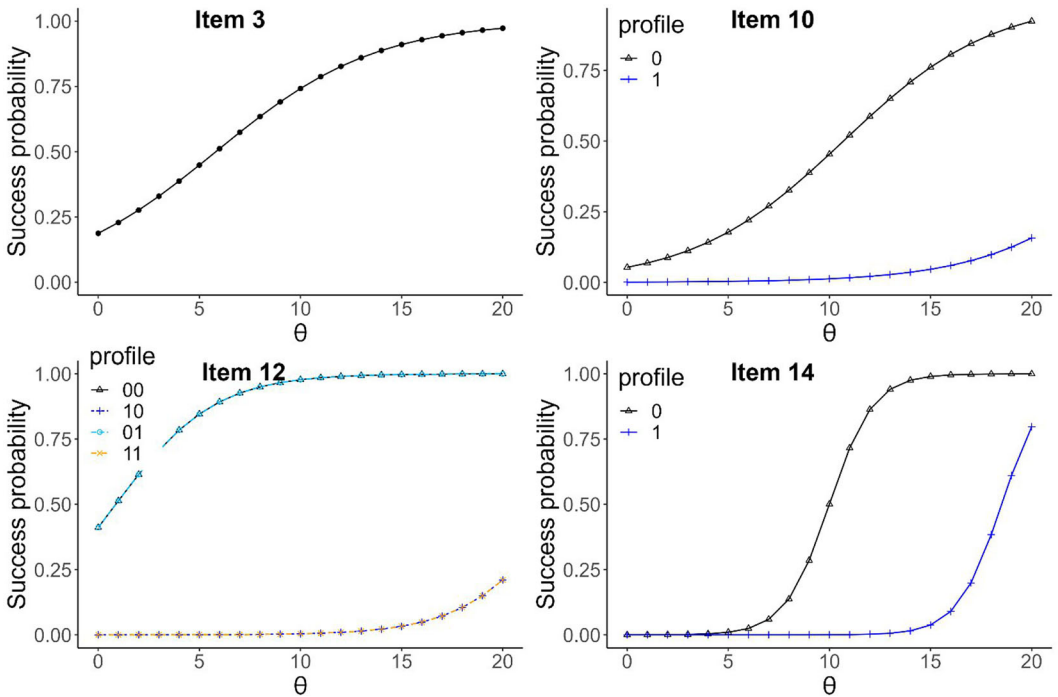

Figure 6. Item characteristic curves based on AM-GDPM. [Color figure can be viewed at wileyonlinelibrary.com]

Table 5
Correlations between Estimated Person Parameters and Total Scores

	α_1	α_2	α_3	α_4	Correlation with Sum Score
θ	–.31	–.45	–.25	–.30	.88
α_1	1	.19	.28	.20	–.38
α_2		1	.36	.28	–.54
α_3			1	.26	–.40
α_4				1	–.41
Prevalence	34.6%	23.0%	22.7%	12.0%	

one misconception, so there are two curves plotted: one for students who possess the measured misconception and one for other students. It can be observed that students having the measured misconception had relatively low probabilities of success regardless of their overall ability. In contrast, Item 14 also measures a single misconception, but students with the misconception can have quite a high chance of answering it correctly when their overall abilities are high.

The ICCs can be used to assess the utility of the items. First, the separability between curves reflects the diagnostic power of misconceptions. For example, the two curves of Item 14 were quite separable, suggesting that students with the misconception performed quite differently from those without the misconception. However, for Item 12, the curves for misconception profiles 00 and 01, and 10 and 11 overlap, suggesting a poor diagnostic power for α_4 and a potential Q-matrix misspecification. Second, the slope of the curve reflects the discrimination of the item for a certain group of students. Take Item 10 as an example; it has relatively high discrimination power for students who do not possess the misconception (i.e., when overall ability increases from low to high, probability of success increases substantially too), but relatively low discrimination power for those who has the misconception (i.e., when overall ability increases from low to high, probability of success does not change substantially).

To evaluate the reliability of person parameter estimation, we use a Monte-Carlo approach. In particular, a sample with one million students with ability and misconception profiles drawn from the posterior distribution of the above fitted AM-GDPM was obtained. Their responses were simulated using AM-GDPM based on the estimated item parameters and refit with fixed item parameters. The classification accuracy for a misconception can be defined as the proportion students who are correctly classified for that misconception. Results showed that the classification accuracy for four misconceptions was .78, .99, .92, and .91, suggesting a reasonably high reliability of misconception classifications, especially for misconceptions 2, 3, and 4. Pearson's correlation between estimated and true overall abilities was .82, suggesting a good reliability for ability estimation.

Details about the correlations between estimated ability, misconceptions, and total scores of the test are given in Table 5. It can be observed that students' estimated

ability regarding thermal concepts was negatively correlated with the possession of misconceptions (from $-.25$ with α_3 to $-.45$ with α_2). In addition, the students' estimated ability was positively correlated with their total scores (.88), and the correlation among the misconception parameters ranged from .19 (between α_1 and α_2) to .36 (between α_2 and α_3), indicating that misconceptions tend to cooccur. Also, misconception 1 is the most prevalent, being found in 34.6% of students, while misconception 4 is the least prevalent, observed in 12.0% of students. The percentages of students who were found to have misconceptions 2 and 3 were estimated at 23.0% and 22.7%, respectively.

Summary and Discussion

Diagnosing misconceptions allows teachers to gain deeper insights into students' understanding, whereas scaling students on a continuum scale allows teachers to know students' overall proficiency. To simultaneously estimate students' overall ability and misconceptions based on dichotomously scored items, this article proposed the general dual-purpose model (GDPM) for binary data. The EM algorithm was developed to estimate its item parameters. Two special cases of the GDPM with substantive interpretations were carefully studied via both simulated and real data—one assumes that misconceptions work in a disjunctive way, whereas the other assumes misconceptions function additively and independently. The simulation study showed that the item parameters of the proposed models could be accurately estimated, especially when the sample size was large. The estimation accuracy of students' overall ability and misconception profiles was primarily affected by sample size and attribute correlations. An interesting observation is that although item parameters tend to be less accurately estimated as attribute correlation increases (i.e., larger biases and RMSEs), the overall ability and misconception profiles tend to be estimated more accurately. This may be caused by the fact that information can be borrowed from other dimensions when they are highly correlated.

The real data analysis shows that the additive version of the GDPM fits data better than the disjunctive version of the GDPM, though both could fit data adequately in an absolute sense probably because many items do not involve any misconceptions, and thus two models are equivalent for them. The analysis of real data also shows that the misconceptions were negatively associated with overall ability, suggesting that students of lower proficiency were prone to some misconceptions. Also, the positive correlation between misconceptions implies that misconceptions tend to cooccur.

This paper defines the overall ability as a categorical variable for several reasons. First, a continuous ability is often assumed to be normally distributed, but this assumption is not needed for categorical variables. Second, as the number of levels increases for the categorical variable, the categorical variable can provide detailed information that is similar to the continuous variable. Third, the categorical variable is easier to understand for practitioners than the continuous variable, which usually has a mean of 0 and standard deviation of 1 for identifiability purpose. Last but not least, the joint distribution of ability and misconception can be parameterized as a multinomial distribution when the ability is categorical, and hence, the estimation is simplified. In this paper, the number of categories for the overall

ability, for both simulation study and real data analysis, was set at the number of items plus one to mimic the total score. It turns out that as long as the number of categories is large enough, it has little impact on model-data fit and person classifications.

This study shows that the GDPM is a promising tool for scaling students and identifying their misconceptions, but additional research is needed to fulfill its potential. First, it is equally important, if not more, to have a well-designed test and carefully developed Q-matrix to support the inferences from the GDPM. Issues related to test development for cognitive diagnosis (de la Torre & Minchen, 2014; Leighton & Gierl, 2007) and Q-matrix completeness (Chiu et al., 2009; Köhn & Chiu, 2017) have been extensively discussed in the literature. The importance of the Q-matrix cannot be overstated for CDM, and the validity of the person classifications from the GDPM also relies on the correctly specified Q-matrix. Although many methods have been developed for estimating or refining Q-matrix (e.g., Ma & de la Torre, 2020a; Tu et al., 2023), they may not be used for the GDPM. Future research should investigate how to detect and correct the misspecifications in the Q-matrix when using the GDPM. Second, the number of misconceptions is often large, and thus, how to measure them stably remains challenging. The test may need to be long enough to measure each misconception a sufficient number of times and at the same time, the estimation algorithm also needs to be able to handle a large number of latent variables. The naïve EM algorithm may need to be modified for this purpose, and recently many more advanced algorithms have been proposed. Also, in the simulation study, a multivariate normal distribution was assumed to govern the relation between latent variables, but this assumption can be relaxed, and data can be simulated from a multinomial distribution directly. Future research may explore how to achieve this while maintaining control for correlations among latent variables, as in Ma et al. (2023). In addition, the proposed model is for cross-sectional data. To better support learning, it is important to extend the model to handle longitudinal data or to embed the model in an adaptive testing framework. Last, the proposed model was only applied to dichotomously scored items, but in practice, polytomously scored items are also widely used. Future research may explore how to extend the model to accommodate polytomous data.

Acknowledgments

The first author acknowledges the financial support provided by the National Science Foundation (Grant No. SES-2150601). The second author acknowledges the financial support provided by the Community of Madrid through the Pluriannual Agreement with the Universidad de Universidad Autónoma de Madrid in its Programa de Estímulo a la Investigación de Jóvenes Doctores (Reference SI3/PJI/2021-00258).

Note

[1]Note that the SICM can handle polytomous data, but the GDPM cannot; however, when dealing with binary data, SICM can be viewed as a special case of the GDPM.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/bf02293801

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*(3), 403–425. https://doi.org/10.1007/s11336-013-9350-4

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*(2), 123–140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

Chiu, C. Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*(4), 633–665. https://doi.org/10.1007/s11336-009-9125-0

Confrey, J. (1990). Chapter 1: A review of the research on student conceptions in mathematics, science, and programming. *Review of Research in Education*, *16*(1), 3–56. https://doi.org/10.3102/0091732x016001003

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. https://doi.org/10.1007/bf02295640

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*(2), 89–97. https://doi.org/10.1016/j.pse.2014.11.001

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), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement*, *39*(1), 62–79. https://doi.org/10.1177/0146621614561315

Gao, Y., Zhai, X., Andersson, B., Zeng, P., & Xin, T. (2020). Developing a learning progression of buoyancy to model conceptual change: A latent class and rule space model analysis. *Research in Science Education*, *50*, 1369–1388. https://doi.org/10.1007/s11165-018-9736-5

Köhn, H.F., & Chiu, C.Y. (2017). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika*, *82*(1), 112–132. https://doi.org/10.1007/s11336-016-9536-7

Kuo, B.C., Chen, C.H., & de la Torre, J. (2018). A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Applied Psychological Measurement*, *42*(3), 179–191. https://doi.org/10.1177/0146621617722791

Kuo, B.C., Chen, C.H., Yang, C.W., & Mok, M. M. C. (2016). Cognitive diagnostic models for tests with multiple-choice and constructed-response items. *Educational Psychology*, *36*(6), 1115–1133. https://doi.org/10.1080/01443410.2016.1166176

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Li, X., & Li, Y. (2008). Research on students' misconceptions to improve teaching and learning in school mathematics and science. *School Science and Mathematics*, *108*(1), 4–7. https://doi.org/10.1111/j.1949-8594.2008.tb17934.x

Ma, W. (2022). A higher-order cognitive diagnosis model with ordinal attributes for dichotomous response data. *Multivariate Behavioral Research*, *57*(2-3), 408–421. https://doi.org/10.1080/00273171.2020.1860731

Ma, W., Chen, J., & Jiang, Z. (2023). Attribute continuity in cognitive diagnosis models: Impact on parameter estimation and its detection. *Behaviormetrika*, *50*(1), 217–240. https://doi.org/10.1007/s41237-022-00174-y

Ma, W., & de la Torre, J. (2020a). An empirical Q-matrix validation method for the sequential generalized DINA model. *The British Journal of Mathematical and Statistical Psychology*, *73*(1), 142–163. https://doi.org/10.1111/bmsp.12156

Ma, W., & de la Torre, J. (2020b). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*(14), 1–26. https://doi.org/10.18637/jss.v093.i14

Ma, W., & Jiang, Z. (2021). Estimating cognitive diagnosis models in small samples: Bayes modal estimation and monotonic constraints. *Applied Psychological Measurement*, *45*(2), 95–111. https://doi.org/10.1177/0146621620977681

Ma, W., Minchen, N., & de la Torre, J. (2020). Choosing between CDM and unidimensional IRT: The proportional reasoning test case. *Measurement: Interdisciplinary Research & Perspective*, *18*(2), 87–96. https://doi.org/10.1080/15366367.2019.1697122

Nájera, P., Sorrel, M. A., de la Torre, J., & Abad, F. J. (2020). Improving robustness in Q-matrix validation using an iterative and dynamic procedure. *Applied Psychological Measurement*, *44*(6), 431–446. https://doi.org/10.1177/0146621620909904

Ozaki, K., Sugawara, S., & Arai, N. (2020). Cognitive diagnosis models for estimation of misconceptions analyzing multiple-choice data. *Behaviormetrika*, *47*(1), 19–41. https://doi.org/10.1007/s41237-019-00100-9

R Core Team. (2021). *R: A language and environment for statistical computing [Computer software]*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*(1), 1–17. https://doi.org/10.1080/15366367.2018.1435104

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. Routledge.

Thompson, F., & Logue, S. (2006). An exploration of common student misconceptions in science. *International Education Journal*, *7*(4), 553–559.

Tu, D., Chiu, J., Ma, W., Wang, D., Cai, Y., & Ouyang, X. (2023). A multiple logistic regression-based (MLR-B) Q-matrix validation method for cognitive diagnosis models: A confirmatory approach. *Behavior Research Methods*, *55*(4), 2080–2092. https://doi.org/10.3758/s13428-022-01880-x

Vosniadou, S. (2020). Students' misconceptions and science education. In S. Vosniadou (Ed.), *Oxford research encyclopedia of education*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.965

Woodruff, D. J., & Hanson, B. A. (1996). Estimation of item response models using the EM algorithm for finite mixtures. *ACT Research Report Series (96-6)*.

Yeo, S., & Zadnik, M. (2001). Introductory thermal concept evaluation: Assessing students' understanding. *The Physics Teacher*, *39*(8), 496–504. https://doi.org/10.1119/1.1424603

Zhai, X., & Li, M. (2021). Validating a partial-credit scoring approach for multiple-choice science items: An application of fundamental ideas in science. *International Journal of Science Education*, *43*(10), 1640–1666. https://doi.org/10.1080/09500693.2021.1923856

Authors

WENCHAO MA is Associate Professor in the Department of Educational Studies in Psychology, Research Methodology and Counseling at The University of Alabama; 520 Colonial Dr., Tuscaloosa, AL 35487, wenchao.ma@ua.edu. His primary research interests include educational measurement, cognitive diagnosis modeling and item response theory.

MIGUEL A. SORREL is an Assistant Professor at Universidad Autónoma de Madrid, Faculty of Psychology, Ciudad Universitaria de Cantoblanco, Madrid, 28049, Spain; miguel.sorrel@uam.es. His research focuses on cognitive diagnosis modeling, item response theory, and computerized adaptive testing.

XIAOMING ZHAI is an Associate Professor of Science Education and Director of the AI4STEM Education Center at the University of Georgia. 125M Aderhold Hall, 110 Carlton St., Athens, GA 30602; xiaoming.zhai@uga.edu. His primary research interests include applying AI in science teaching, learning, and assessment.

YUAN GE is an Associate Psychometrician at The College Board, 800 Township Line Road, Yardley, PA 19067; yge@collegeboard.org. Her primary research interests include psychometrics and assessment.