Title

Diagnostic Classification Models for Ordinal Item Responses

Authors

Liu, Ren

Jiang, Zhehan

Publication Date

2018

DOI

10.3389/fpsyg.2018.02512

Peer reviewed

frontiers in Psychology

ORIGINAL RESEARCH
published: 11 December 2018
doi: 10.3389/fpsyg.2018.02512

Diagnostic Classification Models for Ordinal Item Responses

Ren Liu[1]* *and Zhehan Jiang*[2]*

[1] Psychological Sciences, University of California, Merced, Merced, CA, United States, [2] University Libraries, University of Alabama, Tuscaloosa, AL, United States

The purpose of this study is to develop and evaluate two diagnostic classification models (DCMs) for scoring ordinal item data. We first applied the proposed models to an operational dataset and compared their performance to an epitome of current polytomous DCMs in which the ordered data structure is ignored. Findings suggest that the much more parsimonious models that we proposed performed similarly to the current polytomous DCMs and offered useful item-level information in addition to option-level information. We then performed a small simulation study using the applied study condition and demonstrated that the proposed models can provide unbiased parameter estimates and correctly classify individuals. In practice, the proposed models can accommodate much smaller sample sizes than current polytomous DCMs and thus prove useful in many small-scale testing scenarios.

Keywords: diagnostic classification model, ordinal item responses, partial credit model, rating scales, Bayesian estimation, Markov Chain Monte Carlo (MCMC)

OPEN ACCESS

Edited by:
Dubravka Svetina,
Indiana University Bloomington,
United States

Reviewed by:
Yong Luo,
National Center for Assessment in Higher Education (Qiyas), Saudi Arabia
Roy Levy,
Arizona State University, United States

***Correspondence:**
Zhehan Jiang
zjiang17@ua.edu
Ren Liu
rliu45@ucmerced.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *12 August 2018*
Accepted: *26 November 2018*
Published: *11 December 2018*

Citation:
Liu R and Jiang Z (2018) Diagnostic Classification Models for Ordinal Item Responses. Front. Psychol. 9:2512. doi: 10.3389/fpsyg.2018.02512

Grouping people into different categories are often of interest in educational and psychological tests. For example, the Five Factor Personality Inventory-Children (McGhee et al., 2007) aims to identify which personalities a child possesses. In another case of career assessment, the Strong Interest Inventory (Prince, 1998; Staggs, 2004; Blackwell and Case, 2008) aims to categorize individuals into occupational themes for identifying their career interest areas. From a psychometric standpoint, those tests share at least three commonalities. First, they are usually multidimensional tests, meaning that multiple latent traits are assessed. Second, the purpose of such tests is to label individuals through assigning them with one of the pre-defined categories. Third, they usually allow for ordinal item responses such as strongly disagree, disagree, agree and strongly agree. For scoring tests with such features, diagnostic classification models (DCMs) have provided an attractive framework in psychometrics because they are designed to classify individuals into pre-defined latent categories (Rupp and Templin, 2008; Rupp et al., 2010). However, most current DCMs for polytomous items consider item response categories as nominal without using the ordered category information (Templin et al., 2008; e.g., de la Torre, 2010; Ma and de la Torre, 2016). As a result, those models are often large and require a sample size hardly attainable for parameter estimation. The purpose of this study is to create smaller ordinal DCMs that are designed to score individuals on an ordinal scale. In this article, we first review current polytomous DCMs. Then, we explain the theoretical development of the proposed models. Next, we fit the proposed models to an operation dataset and compare their performance with a current polytomous DCM in which the ordered structure is ignored. Afterwards, we performed a small simulation study using the applied study condition to evaluate the parameter recovery of the proposed models. Finally, we discuss the application and advantages of the models and offer future research recommendations.

REVIEW OF CURRENT POLYTOMOUS DCMS

Existing literature has considered DCMs from either the perspective of Bayesian networks or confirmatory latent class models. In the Bayesian networks literature, the Dibello-Samejima modeling framework advanced by Almond et al. (2001, 2009, 2015), and Levy and Mislevy (2004) is an example of scoring polytomous item data. In this article, we consider DCMs as confirmatory latent class models with two outstanding features. First, the latent traits, commonly referred to as attributes, are defined *a priori*. The possible possession status of all latent traits forms latent classes, commonly referred to as attribute profiles. In this article, we use $k = 1, \ldots, K$ to index latent traits and $\boldsymbol{\alpha}_c = \{\alpha_1, \ldots, \alpha_K\}$ to index attribute profiles for latent class c. Second, the measurement relationship between items and attributes is defined *a priori*. This information is contained in an item-by-attribute incidence matrix, commonly referred to as the Q-matrix (Tatsuoka, 1983), where an entry $q_{ik} = 1$ when item i measures attribute k, and $q_{ik} = 0$ otherwise.

To our knowledge, eight DCMs have been developed to score polytomous item data. Each model is constructed through applying a polytomous extension method to a dichotomous DCM. We listed such information in **Table 1**. Most polytomous DCMs are developed based on the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009) or its equivalent: the generalized deterministic input noisy "and" gate (GDINA; de la Torre, 2011) model. The NRDM, GDM, PC-DINA, and SG-DINA utilize the concept of the nominal response model (NRM; Bock, 1972) in item response theory where each response option in each item has its own intercept and slope; The P-LCDM, DINA-GD, and GPDM utilize the concept of the graded response model (GRM; Samejima, 1969) where the differences between cumulative probabilities of adjacent options are modeled; the RSDM utilize the concept of the rating scale model (RSM; Andrich, 1978) where items measuring the same set of attributes share response option parameters. To summarize, many current DCMs are built to accommodate nominal response data. For example, the NRDM defines the probability of individuals in latent class c selecting response option m on item i, such that

$$P(X_i = m|\boldsymbol{\alpha}_c) = \frac{\exp[\lambda_{0,i,m} + \boldsymbol{\lambda}_{i,m}^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)]}{\sum_{m=0}^{M-1} \exp[\lambda_{0,i,m} + \boldsymbol{\lambda}_{i,m}^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)]}, \quad (1)$$

where $\lambda_{0,i,m}$ is the intercept parameter associated with option m on item i, and $\boldsymbol{\lambda}_{i,m}^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)$ index all the main effects and higher-order interaction effects of the k attributes associated with option m on item i, which can be expressed as $\sum_{k=1}^{K} \lambda_{1,i,k,m}(\alpha_{c,k} q_{i,k}) + \sum_{k=1}^{K-1} \sum_{k'=K+1}^{K} \lambda_{2,i,k,k',m} \left(\alpha_{c,k} \alpha_{c,k'} q_{i,k} q_{i,k'}\right) + \ldots$.

Let us break down the summation symbol in Equation 1 for an instructional example. On item i with four response options ($M = 4$): 0,1,2, and 3, the probability of selecting response option 2 is expressed as

$$P\left(X_i = 2|\boldsymbol{\alpha}_c\right) = \frac{\exp[\lambda_{0,i,2} + \boldsymbol{\lambda}_{i,2}^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)]}{\begin{array}{c}\exp[\lambda_{0,i,0} + \boldsymbol{\lambda}_{i,0}^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)] + \\ \exp\left[\lambda_{0,i,1} + \boldsymbol{\lambda}_{i,1}^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)\right] + \\ \exp\left[\lambda_{0,i,2} + \boldsymbol{\lambda}_{i,2}^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)\right] + \\ \exp[\lambda_{0,i,3} + \boldsymbol{\lambda}_{i,3}^T \mathbf{h}\left(\boldsymbol{\alpha}_c, \mathbf{q}_i\right)]\end{array}}. \quad (2)$$

It should be clear in Equation 2 that each option in item i is associated with its own set of intercept, main effects and higher-order interaction parameters. As a result, the NRDM is able to accommodate polytomous response options that can be either ordered or not ordered.

MODEL DEVELOPMENT

To develop DCMs that utilize the ordered structure of response options in many polytomous items (e.g., 0 = never, 1 = seldom, 2 = sometimes, 3 = usually), we contemplated on how the parameters on the NRM can be constrained to create the Generalized Partial Credit Model (GPCM; Muraki, 1992) and Generalized Rating Scale Model (GRSM; Muraki, 1992) in item response theory. The probability of selecting option m on item i

TABLE 1 | Previous DCMs for scoring polytomous item data.

Model	Full Name	Dichotomous Core	Similar Extension Method in IRT Models
NRDM	The nominal response diagnostic model (Templin et al., 2008)	LCDM	The nominal response model (NRM; Bock, 1972)
GDM	The general diagnostic model (von Davier, 2008)	LCDM	NRM
PC-DINA	The partial-credit deterministic input noisy "and" gate model (de la Torre, 2010)	DINA	NRM
P-LCDM	The polytomous log-linear cognitive diagnosis model (Hansen, 2013)	LCDM	The graded response model (GRM; Samejima, 1969)
PC-DINA	The sequential generalized DINA model (Ma and de la Torre, 2016)	LCDM	NRM
DINA-GD	The DINA model for graded data (Tu et al., 2017)	DINA	GRM
GPDM	The general polytomous diagnosis model (Chen and de la Torre, 2018)	LCDM	GRM
RSDM	The rating scale diagnostic model (Liu and Jiang, submitted)	LCDM	The rating scale model (RSM; Andrich, 1978)

given a unidimensional latent trait θ for examinee e is defined as

$$P(X_i = m|\theta_e) = \frac{\exp\left(d_{im}\theta_e + b_{im}\right)}{\sum_{m=0}^{M-1}\exp\left(d_{im}\theta_e + b_{im}\right)}, \quad (3)$$

for the NRM,

$$P(X_i = m|\theta_e) = \frac{\exp\sum_{m=0}^{m}\left[d_i\left(\theta_e + b_{im}\right)\right]}{\sum_{s}^{M-1}\exp\sum_{m=0}^{s}\left[d_i\left(\theta_e + b_{im}\right)\right]}, \quad (4)$$

for the GPCM, and

$$P(X_i = m|\theta_e) = \frac{\exp\sum_{m=0}^{m}\left[d_i\left(\theta_e + b_i + t_m\right)\right]}{\sum_{s}^{M-1}\exp\sum_{m=0}^{s}\left[d_i\left(\theta_e + b_i + t_m\right)\right]}, \quad (5)$$

for the GRSM. The d_{im} and b_{im} in Equation 3 are the slope parameter and intercept parameter for option m in item i, respectively. In Equation 4, the slope parameter d_i loses the subscript m; instead, summation symbols are used such that the d_{im} in Equation 3 is represented by $m \times d_i \; \forall \; m > 0$ in Equation 4. To obtain Equation 5, an extra constraint is imposed on Equation 4 where the b_{im} is decomposed into a general item intercept for item i: b_i and a general response option intercept for option m: t_m that is applicable to all items.

Inspired by how the NRM can be constrained to arrive at the GPCM and GRSM, we propose two ordinal DCMs through applying constraints to the NRDM so that the proposed models are targeted for scoring ordered item data. We refer to these models as the Ordinal Response Diagnostic Model (ORDM) and the Modified Ordinal Response Diagnostic Model (MORDM). The ORDM is defined as

$$P(X_i = m|\boldsymbol{\alpha}_c) = \frac{\exp\sum_{m=0}^{m}\left[\lambda_{0,i,m} + \boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]}{\sum_{s}^{M-1}\exp\sum_{m=0}^{s}\left[\lambda_{0,i,m} + \boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]}, \quad (6)$$

where $\boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right) = \sum_{k=1}^{K}\lambda_{1,i,k}(\alpha_{c,k}q_{i,k}) + \sum_{k=1}^{K-1}\sum_{k'=K+1}^{K}\lambda_{2,i,k,k'}\left(\alpha_{c,k}\alpha_{c,k'}q_{i,k}q_{i,k'}\right) + \ldots$. For identifiability purposes, we impose three sets of constraints on the ORDM. First, in order to fix the scale, we adopt Thissen (1991)'s approach and fix all parameters associated with the first response option to 0, such that

$$\sum_{m=0}^{0}\left(\lambda_{0,i,m}\right) = 0 \; \forall \; i,$$
$$\sum_{m=0}^{0}\left(\lambda_{1,i,k}\right) = 0 \; \forall \; i, \; k,$$
$$\sum_{m=0}^{0}\left(\lambda_{2,i,k,k'}\right) = 0 \; \forall \; i, k, k',$$

and for all higher-order interactions. Second, we constrain parameters associated with main effects and higher-order interactions to be <0 so that the possession of more attributes increases the probability of selecting a higher response option:

$$\lambda_{1,i,k} > 0 \; \forall \; k,$$
$$\lambda_{2,i,k,k'} > 0 \; \forall \; k, k',$$

and for other higher-order interactions. Third, we constrain intercept parameters of a higher response option to be smaller than those of a lower response option so that the probability of selecting a higher response option is smaller for individuals without the measured attributes such that

$$\lambda_{0,i,m} \geq \lambda_{0,i,m+1} \forall \; i, m.$$

Comparing Equation 6 to Equation 1, the $\boldsymbol{\lambda}_{i,m}^T$ in Equation 1 loses the subscript m. The $\boldsymbol{\lambda_i}$ parameters in Equation 6 are summated for their associated response options.

Let us break down the summation symbol in Equation 6 for an instructional example. On item i with four response options: 0,1,2, and 3, the probability of selecting response option 2 is expressed as

$$P(X_i = 2|\boldsymbol{\alpha}_c) = \frac{\exp\left[0 + \left[\lambda_{0,i,1} + \boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right] + \left[\lambda_{0,i,2} + \boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]\right]}{\begin{array}{c}\exp(0) + \exp\left[0 + \left[\lambda_{0,i,1} + \boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]\right] + \exp\left[0 + \left[\lambda_{0,i,1} + \boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right] + \left[\lambda_{0,i,2} + \boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]\right] + \\ \exp\left[0 + \left[\lambda_{0,i,1} + \boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right] + \left[\lambda_{0,i,2} + \boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right] + \left[\lambda_{0,i,3} + \boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]\right]\end{array} \quad (7)$$

Equation 7 is similar to Equation 2 in two ways. First, both equations ask what the probability is that an individual in latent class c selecting option 2 as compared to the sum of probabilities of all response options that the individual could select. Second, the intercept parameter is freely estimated for each response option in each item (e.g., $\lambda_{0,i,1}$, $\lambda_{0,i,2}$, and $\lambda_{0,i,3}$). However, what is different is that the $\boldsymbol{\lambda}_{i,2}^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)$ in Equation 2 is replaced by $2 \times \boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha_c},\mathbf{q}_i\right)$ in Equation 7. It should be clear now that the proposed ORDM can be expressed as a constrained version of the NRDM, analogous to how the GPCM can be formulated as a constrained version of the NRM.

The MORDM is defined the same as the ORDM in Equation 6, except that the $\lambda_{0,i,m}$ is decomposed into general item parameters and shared response option parameters. Before deciding to share response option parameters across all items, we should remember that DCMs are multidimensional models while the NRM is a unidimensional model. Therefore, it would be unwarranted to assume that all items in a DCM can share the same set of response option parameters because those items may measure different traits. Instead, what we can do is to allow response option parameters to be shared within each dimension. As introduced above, DCMs are confirmatory latent class models, which means that the dimensions in DCMs can be represented through latent classes (i.e., attribute profiles). We express the relationship between items and attribute profiles in an item-by-attribute-profile incidence matrix called the W-matrix (Liu and Jiang, submitted), where an entry $w_{iv} = 1$ when item i measures attribute set v, and 0 otherwise. By definition, each column corresponds to a unique attribute profile; each row has only one entry of 1 and all others of 0. Utilizing the W-matrix, we are able to allow response option parameters to be shared within

items that measure the same set of attributes. Subsequently, the $\lambda_{0,i,m}$ in Equation 6 is decomposed into $\lambda_{0,i}$ and $\sum_{v=1}^{V}\lambda_{0,m_v}w_{iv}$ in the MORDM, where the $\sum_{v=1}^{V}\lambda_{0,m_v}w_{iv}$ represents the response option parameters shared across items that measure attribute set v. Now, we can define the MORDM as

$$P(X_i=m|\boldsymbol{\alpha}_c)=\frac{\exp\sum_{m=0}^{m}\left[\lambda_{0,i}+\sum_{v=1}^{V}\lambda_{0,m_v}w_{iv}+\boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]}{\sum_{s}^{M-1}\exp\sum_{m=0}^{s}\left[\lambda_{0,i}+\sum_{v=1}^{V}\lambda_{0,m_v}w_{iv}+\boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]}. \tag{8}$$

The constraints we impose on the MORDM is the same as those on the ORDM, except that the third constraint (i.e., for the intercept parameters) needs to be adapted to the MORDM. In the MORDM, we impose this constraint:

$$\sum_{v=1}^{V}\lambda_{0,m_v}w_{iv}\leq\sum_{v=1}^{V}\lambda_{0,m-1_v}w_{iv}\forall\ v,m.$$

to make sure that individuals without the measured attributes have a smaller probability of selecting a higher response option.

Let us continue the example of selecting response option 2 on an item with options 0,1,2, and 3. The MORDM in such case is expressed as

$$P\left(X_i=2|\boldsymbol{\alpha}_c\right)=\frac{\exp\left[0+\left[\lambda_{0,i}+\sum_{v=1}^{V}\lambda_{0,m=1_v}w_{iv}+\boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]+\left[\lambda_{0,i}+\sum_{v=1}^{V}\lambda_{0,m=2_v}w_{iv}+\boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]\right]}{\begin{array}{c}\exp\left(0\right)+\exp\left[0+\left[\lambda_{0,i}+\sum_{v=1}^{V}\lambda_{0,m=1_v}w_{iv}+\boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]\right]+\\ \exp\left[0+\left[\lambda_{0,i}+\sum_{v=1}^{V}\lambda_{0,m=1_v}w_{iv}+\boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]+\left[\lambda_{0,i}+\sum_{v=1}^{V}\lambda_{0,m=2_v}w_{iv}+\boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]\right]+\\ \exp\left[0+\left[\lambda_{0,i}+\sum_{v=1}^{V}\lambda_{0,m=1_v}w_{iv}+\boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]+\left[\lambda_{0,i}+\sum_{v=1}^{V}\lambda_{0,m=2_v}w_{iv}+\boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]\right]\\ +\left[\lambda_{0,i}+\sum_{v=1}^{V}\lambda_{0,m=3_v}w_{iv}+\boldsymbol{\lambda}_i^T\mathbf{h}\left(\boldsymbol{\alpha}_c,\mathbf{q}_i\right)\right]\Big]\end{array}} \tag{9}$$

Equation 9 can be viewed as a constrained version of Equation 7 where the intercept parameters are decomposed. To summarize, one can constrain the main effect parameters of the NRDM to arrive at the ORDM, and further constrain the intercept parameters of the ORDM to arrive at the MORDM.

OPERATIONAL STUDY

The purpose of this operational study is to compare the performance of the ORDM and the MORDM with the NRDM through fitting these three models to an ordinal item response dataset. The motivating research question was: can the more parsimonious ORDM and/or the MORDM perform similarly to the NRDM? To answer this question, we looked into the following six types of outcomes: (1) model fit, (2) profile prevalence estimates, (3) item parameter estimates, (4) conditional response option probabilities, (5) attribute and profile classification agreement rates, and (6) individual continuous scores.

Data

The dataset used in this study came from a survey of 8th grade students in Austria. We obtained this dataset from the "CDM" (Robitzsch et al., 2018) *R* package alongside the permission to use this dataset from the authors. In the survey, there were four questions asking about respondents' self-concept in math, and four questions asking about how much they enjoy studying math. Therefore, two attributes were specified: "math self-concept" and "math joy." Each of the eight questions has four response options: 0 (low), 1 (mid-low), 2 (mid-high) and 3 (high). We randomly selected 500 individuals' responses from the entire dataset because we are interested in the model performance under small and attainable sample size conditions. We display the item-trait relationship and frequencies of each response option on each item in **Table 2**. A brief look of **Table 2** reveals that the response data is positively skewed for items 1–4 (i.e., measuring math self-concept) with more individuals selecting options 0 and 1, while it is negatively skewed for items 5–8 (i.e., measuring math joy) with more individuals selecting options 2 and 3.

TABLE 2 | Item data information.

Item	Dimension	0 (Low)	1 (Mid-low)	2 (Mid-high)	3 (High)
1	Math Self-concept	154 (30.8%)	233 (46.6%)	94 (18.8%)	19 (3.8%)
2	Math Self-concept	203 (40.6%)	178 (35.6%)	92 (18.4%)	27 (5.4%)
3	Math Self-concept	237 (47.4%)	153 (30.6%)	65 (13.0%)	45 (9.0%)
4	Math Self-concept	105 (21.0%)	197 (39.4%)	145 (29.0%)	53 (10.6%)
5	Math Joy	13 (2.6%)	67 (13.4%)	196 (39.2%)	224 (44.8%)
6	Math Joy	31 (6.2%)	136 (27.2%)	191 (38.2%)	142 (28.4%)
7	Math Joy	97 (19.4%)	160 (32.0%)	147 (29.4%)	96 (19.2%)
8	Math Joy	73 (14.6%)	160 (32.0%)	155 (31.0%)	112 (22.4%)

Analysis

Parameters were estimated through implementing Hamiltonian Monte Carlo (HMC) algorithms in Stan (Carpenter et al., 2016). HMC has been acclaimed for its estimation efficiency compared to Gibbs sampler and the Metropolis algorithm especially when complex models including DCMs are involved (e.g., Girolami and Calderhead, 2011; da Silva et al., 2017; Jiang and Skorupski, 2017; Jiang and Templin, 2018; Luo and Jiao, 2018). The Stan codes used in this study for estimating the ORDM and MORDM are provided in the **Supplementary Material**.

We used less informative priors in the HMC algorithms with $N\left(0,20\right)$ for each item parameter and *Dirichlet*(2) for each attribute profile. The priors are considered less informative because a large standard deviation (i.e., 20) produces a relatively

flat-shaped normal distribution, and a conjugate Dirichlet distribution with all equivalent parameter values (e.g., 2,2,2,2) is approximately a uniform distribution. Using less informative priors are recommended in similar DCM studies such as Chen et al. (2018), Culpepper and Hudson (2018), and Jiang and Carter (2018).

For each model, we ran two Markov chains with random starting values. The total length of the HMC sample was 6,000, for which the first 2,000 iterations were discarded as burn-in. To assess whether the Markov Chains converged to a stationary distribution the same as a posterior distribution, we computed the multivariate Gelman-Rubin convergence statistic $\hat{R}$ proposed by Brooks and Gelman (1998). $\hat{R}$ smaller than 1.1 for each parameter is usually considered convergence (Gelman and Rubin, 1992; Junker et al., 2016). For each of the three models, we obtained all the $\hat{R}$ smaller than 1.1.

We successfully applied the constraints designed for the ORDM to both the NRDM and the ORDM, and applied the constraints for the MORDM to itself through specifying pseudo response option parameters such that

$$\lambda_{z,m=m} = \lambda_{z,m=1} + \lambda'_{z,m=2} + \ldots + \lambda'_{z,m=m} \; \forall \, z, m. \qquad (10)$$

with the constraint $\lambda'_{0,m} \leq 0 \; \forall \; m$ and $\lambda'_{z,m} \geq 0 \; \forall \; z \geq 1, m$.

For model fit assessment, we used the leave-one-out (LOO) cross-validation approach for Bayesian estimation to compute the expected log predictive density (ELPD) and LOO information criterion (LOOIC) for each model. As suggested in Gelman et al. (2014), Vehtari et al. (2017) and Yao et al. (2018), the LOOIC is preferred over traditional simpler indices such as the Akaike information criterion (AIC), Bayesian information criterion (BIC) and deviance information criterion (DIC). Note that research has been lacking on the performance of the LOOIC for assessing DCM model fit. Regarding other latent variable models, Revuelta and Ximénez (2017) found that the LOOIC perform poorly with multidimensional continuous latent variable models, despite its fully Bayesian nature and excellent performance with unidimensional IRT models (e.g., Luo and Al-Harbi, 2017).

Results

We estimated 48, 32, and 22 item parameters for the NRDM, the ORDM and the MORDM, respectively. For this dataset, the ORDM was 33% smaller than the NRDM, and the MORDM was 54% smaller than the NRDM. For each parameter, we report the mean of the posterior distribution as the point estimate and the standard deviation of the posterior distribution to indicate the uncertainty around the mean estimate. We first examined the results on model fit indices and listed the ELPD and LOOIC estimates and standard errors for each model in **Table 3**. For each index, smaller values indicate better fit. Although both indices suggested better fit for the ORDM than the other two models, their differences relative to the scale of the standard error indicate that the three models did not fit significantly different from each other. In practice, one would probably either select the most parsimonious MORDM or the best fitting ORDM for further interpretations.

TABLE 3 | Model fit information in the operational study.

	NRDM		ORDM		MORDM	
	Estimate	Se	Estimate	Se	Estimate	Se
ELPD	−9.5	2.1	−9.3	2.0	−9.7	1.6
LOOIC	19.1	4.2	18.7	3.9	19.3	3.3

TABLE 4 | Profile prevalence estimates and standard deviations under the NRDM, the ORDM and the MORDM in the operational study.

Profile	NRDM	ORDM	MORDM
(0,0)	0.346 (0.029)	0.351 (0.027)	0.351 (0.026)
(1,0)	0.084 (0.021)	0.074 (0.016)	0.105 (0.019)
(0,1)	0.147 (0.024)	0.156 (0.022)	0.125 (0.021)
(1,1)	0.424 (0.028)	0.419 (0.025)	0.418 (0.024)

Examining profile prevalence estimates provides further evidence about the similar performance of the three models. **Table 4** lists the estimates and standard deviations of the profile prevalence. Each estimate represents the probability of an individual having an attribute profile at large. The estimates for the NRDM were very similar to the ORDM as the point-estimate differences between the models were smaller than 0.01 for every profile. The point-estimate differences between the NRDM and the MORDM were all smaller than 0.02 for every profile.

We could also look into the similarities of the item parameter estimates. **Tables 5**–**7** display the item parameter estimates and their standard deviations for the NRDM, the ORDM, and the MORDM, respectively. Remember that the estimated pseudo parameters can be transformed to real parameters using Equation 10. For example, the intercept parameter for response option 2 of item 1 under the MORDM can be obtained through $\lambda_{0,i} + \lambda_{0,m=1} + \lambda_{0,m=2}{}' = 5.834 - 6.204 - 2.871 = -3.241$. Results show that the parameter estimates were similar across the three models. For example, the intercept estimates for response option 1 of item 1 were −0.423, −0.390, and −0.370, respectively for the NRDM, ORDM and MORDM.

Such similarities can be more revealing through computing probabilities of selecting each response option for individuals with and without the measured attribute. We selected items 1 and 4 (measuring math self-concept) to display their response option curves (ROCs) in **Figure 1**. The three ROCs for each item were similar to each other although those under the NRDM and the ORDM were even more alike. Also made clear by the ROCs is that the response option parameters in the MORDM are not unique to each item; instead, the first four items share the same set of response option parameters. Hence, the ROCs under the MORDM depart a bit more from those ROCs under the NRDM and the ORDM. The location of each intersection between the two curves on the x-axis in each graph represents the minimum response option where individuals with the attribute start to have higher probabilities to select than individuals without the attribute. For example, for items 1 and 4, individuals

TABLE 5 | NRDM: item parameter estimates and standard deviations in the operational study.

	$\lambda_{0,i,m=1}$	$\lambda'_{0,i,m=2}$	$\lambda'_{0,i,m=3}$	$\lambda_{1,i,m=1}$	$\lambda'_{1,i,m=2}$	$\lambda'_{1,i,m=3}$
Item 1	−0.423	−10.710	−12.578	3.619	10.339	10.981
	(0.187)	(4.389)	(4.827)	(0.601)	(4.380)	(4.817)
Item 2	−0.934	−2.265	−1.927	2.149	1.995	0.744
	(0.195)	(0.439)	(0.294)	(0.310)	(0.482)	(0.686)
Item 3	−1.557	−6.900	−10.818	2.714	6.354	10.481
	(0.148)	(1.610)	(4.956)	(0.347)	(2.600)	(4.949)
Item 4	0.269	−4.540	−3.491	3.886	5.319	2.507
	(0.064)	(1.582)	(1.854)	(1.815)	(1.790)	(1.846)
Item 5	1.986	−0.036	−1.241	16.908	2.756	2.130
	(0.202)	(0.003)	(0.216)	(5.130)	(0.501)	(0.247)
Item 6	1.347	−0.716	−2.305	17.927	2.780	2.322
	(0.210)	(0.166)	(0.471)	(5.469)	(0.345)	(0.481)
Item 7	0.296	−1.850	−2.255	0.926	2.857	1.940
	(0.156)	(0.210)	(0.835)	(0.340)	(0.362)	(0.858)
Item 8	0.642	−10.306	−11.038	18.139	12.357	10.716
	(0.145)	(4.748)	(4.195)	(5.770)	(4.731)	(5.182)

TABLE 6 | ORDM: item parameter estimates and standard deviations in the operational study.

	$\lambda_{0,i,m=1}$	$\lambda'_{0,i,m=2}$	$\lambda'_{0,i,m=3}$	$\lambda_{1,i}$
Item 1	−0.390	−4.088	−5.321	3.735
	(0.154)	(0.500)	(0.562)	(0.496)
Item 2	−0.834	−2.181	−3.127	1.971
	(0.144)	(0.282)	(0.338)	(0.236)
Item 3	−1.533	−3.377	−3.194	2.841
	(0.200)	(0.335)	(0.334)	(0.274)
Item 4	0.289	−2.180	−3.784	2.902
	(0.144)	(0.327)	(0.540)	(0.454)
Item 5	1.991	−0.029	−1.352	2.290
	(0.214)	(0.020)	(0.224)	(0.216)
Item 6	1.352	−0.670	−2.586	2.626
	(0.211)	(0.146)	(0.269)	(0.229)
Item 7	0.071	−1.370	−2.297	2.039
	(0.144)	(0.185)	(0.235)	(0.181)
Item 8	0.646	−10.395	−12.733	12.427
	(0.141)	(3.543)	(4.500)	(4.498)

TABLE 7 | MORDM: item parameter estimates and standard deviations in the operational study.

	$\lambda_{0,i}$	$\lambda_{0,m=1}$	$\lambda'_{0,m=2}$	$\lambda'_{0,m=3}$	$\lambda_{1,i}$
Item 1	5.834	−6.204	−2.871	−3.781	2.472
	(2.232)	(3.227)	(0.182)	(0.185)	(0.148)
Item 2	5.141	*	*	*	2.512
	(2.238)				(0.154)
Item 3	4.491	*	*	*	2.732
	(2.245)				(0.145)
Item 4	6.424	*	*	*	3.200
	(2.236)				(0.162)
Item 5	10.313	−7.893	−0.527	−2.111	3.262
	(4.794)	(4.766)	(0.091)	(0.120)	(0.181)
Item 6	9.339	*	*	*	2.348
	(4.756)				(0.139)
Item 7	7.948	*	*	*	1.622
	(4.770)				(0.111)
Item 8	8.285	*	*	*	2.033
	(4.808)				(0.117)

The cells with ""indicates that it shares the same parameter with the cell above it.*

with the math self-concept have higher probabilities selecting response option 1 and above than those without the math self-concept.

Ultimately, the three models can be concluded to have similar performance if individuals have received similar categorical and continuous scores. The categorical scores include individuals' attribute and profile classifications. **Table 8** cross-tabulates the attribute classification agreement between each pair of models. The agreement rates between the NRDM and the ORDM were very high: 99.0% and 99.8% for each attribute, respectively. The agreement rates were all over 99.0% on the math self-concept attribute for each pair of models, and the lowest agreement rate was on the math joy attribute: 92.6% between the ORDM and the MORDM. **Table 9** cross-tabulates the profile classification agreement between each pair of models. Agreement rates between each pair were also very high. For example, only 6 out of the 500 individuals were classified into different profiles under the NRDM and the ORDM. The continuous scores are individuals' marginal probabilities of possessing each attribute (Liu et al., 2018). We display the continuous scores for all individuals between each pair of models in **Figure 2**. As expected, most individuals had scores close to either 0 or 1 under each model. For the pair of the NRDM and the ORDM, individuals' scores almost fit into a linear $y=x$ line, meaning that both models produce very similar continuous scores. For other pairs, most scores can still fit into a linear line with only a few cases where scores differed substantially. To quantify the score differences, we computed the root-mean-square deviation (RMSD) for scores between each pair of models. For α_1, the RMSD values were 0.04, 0.06 and 0.06 for NRDM/ORDM, NRDM/MORDM, and ORDM/MORDM, respectively. For α_2, the RMSD values were 0.03, 0.15, and 0.16 for NRDM/ORDM, NRDM/MORDM,

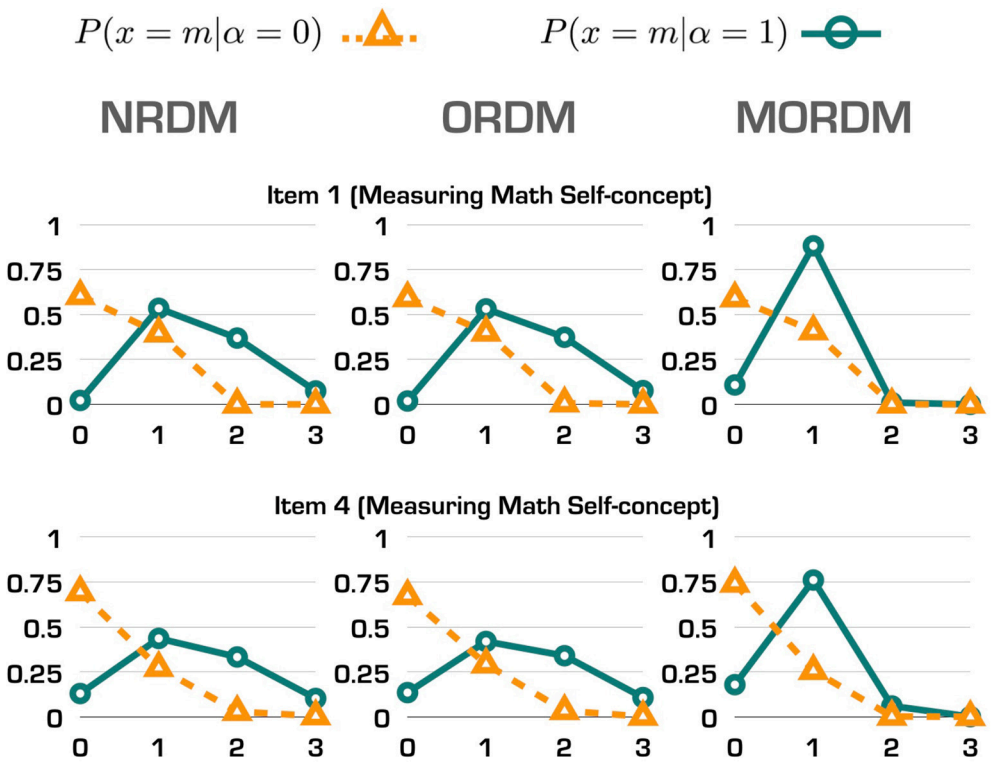

FIGURE 1 | Response option curves for two items in the operational study.

and ORDM/MORDM, respectively. To summarize, results show that the score differences were very small between the models, and we conclude that the three models performed similarly.

SIMULATION STUDY

Methods

The purpose of this simulation study was to investigate whether the proposed ORDM and MORDM can provide unbiased parameter estimate and accurate attribute classification under the operational study condition. In order to do this, we used the parameters obtained from the operational study and generated 100 datasets in R (R Core Team, 2018) for each model. In each dataset, 500 individuals were simulated from a multinomial distribution of (0.351, 0.074, 0.156, 0.419) for each of the four attribute profiles: (0,0), (1,0), (0,1), and (1,1), respectively. We used the item parameters listed in **Tables 6**, **7** to generate item response for the ORDM and MORDM, respectively. We then fit the ORDM to its 100 datasets and the MORDM to its 100 datasets using the same HMC specifications in the operational study. Similar to what we did in the operational study to assess convergence, we obtained the multivariate Gelman-Rubin convergence statistic $\hat{R}$ and found that all the $\hat{R}$ values were between 0.97 and 1.01, indicating that the Markov chains have converged.

To assess parameter recovery, we computed the bias and root mean square error (RMSE) for each item parameter and attribute prevalence estimate. Bias and RMSE for parameter x were computed as:

$$Bias(x) = \frac{\sum_{r=1}^{R} [\hat{e}_r(x) - e(x)]}{R}, \tag{11}$$

$$RMSE(x) = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R} [\hat{e}_r(x) - e(x)]^2}, \tag{12}$$

where $e(x)$ is the true value of parameter x, $\hat{e}_r(x)$ is the rth replicate estimate of parameter x among $R = 100$ datasets. To assess classification accuracy, we explored the agreement between true and estimated classifications on each attribute and provided descriptive statistics on the agreement rates across the 100 datasets.

Results

Tables 10, **11** list the bias and RMSE for the item parameter estimates in the ORDM and MORDM, respectively. Of interest is that most item parameter estimates list bias close to 0 and RMSE smaller than 0.5. We also observed that some of the biases and RMSEs are larger than others. For example, in **Table 10**, the bias and RMSE for $\lambda_{0,5,m=1}$

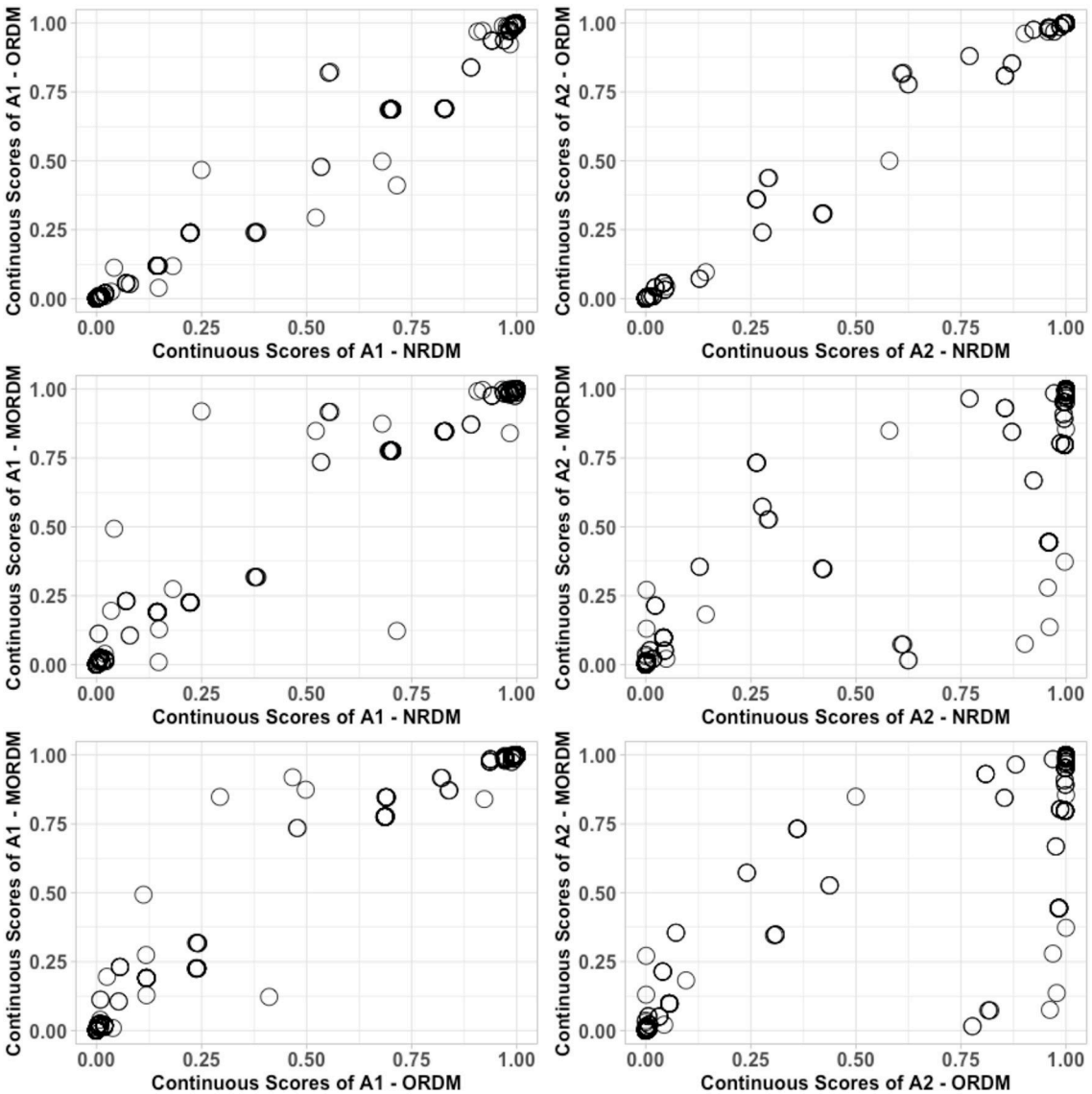

FIGURE 2 | Comparison of continuous scores for each pair of models in the operational study.

seems larger than the $\lambda_{0,i,m=1}$ parameter for other items under the ORDM. We hypothesize that the unbalanced class membership probability and the uniqueness of the original distribution of examinee scores could both contribute to the larger bias and RMSE. A quick revisit of **Table 2** reveals that the response option distribution of item 5 is negatively skewed, which sets itself apart from the other three items that measure the same attribute "math joy." However, this is our initial hypothesis which may be test through a more robust simulation in the future.

Table 12 displays the bias and RMSE for the attribute prevalence estimates in the ORDM and MORDM, respectively. All the bias and RMSE values in this table are smaller than 0.02. **Table 13** contains the descriptive statistics for classification accuracy results. The classification accuracy for each attribute under both models are mostly above 0.99. Results show that both models can correctly recover parameters and provide accurate attribute classifications.

DISCUSSION

Scoring items in an ordinal fashion is common in educational and psychological tests. For example, an essay can be scored on a 0–6 scale, a two-step math question can be partially scored for responses on each step, and a questionnaire can have Likert-type items with eight response options. DCMs

TABLE 8 | Attribute possession agreement between each pair of models in the operational study.

NRDM	ORDM	
	$\alpha_1 = 0$	$\alpha_1 = 0$
$\alpha_1 = 0$	238 (47.6%)	0
$\alpha_1 = 1$	5 (1.0%)	257 (51.4%)
	$\alpha_2 = 0$	$\alpha_2 = 1$
$\alpha_2 = 0$	220 (44.0%)	0
$\alpha_2 = 1$	1 (0.2%)	279 (55.8%)

The total number of agreements between the two models for α_1 and α_2 was 495 (99.0%) and 499 (99.8%), respectively. Cohen's Kappa for α_1 and α_2 were 0.98, and 1.00, respectively.

NRDM	MORDM	
	$\alpha_1 = 0$	$\alpha_1 = 0$
$\alpha_1 = 0$	237 (47.4%)	1 (0.2%)
$\alpha_1 = 1$	1 (0.2%)	261 (52.2%)
	$\alpha_2 = 0$	$\alpha_2 = 1$
$\alpha_2 = 0$	208 (41.6%)	12 (2.4%)
$\alpha_2 = 1$	24 (4.8%)	256 (51.2%)

The total number of agreements between the two models for α_1 and α_2 was 498 (99.6%) and 464 (92.8%), respectively. Cohen's Kappa for α_1 and α_2 were 0.99, and 0.86, respectively.

ORDM	MORDM	
	$\alpha_1 = 0$	$\alpha_1 = 0$
$\alpha_1 = 0$	238 (47.6%)	5 (1.0%)
$\alpha_1 = 1$	0	257 (51.4%)
	$\alpha_2 = 0$	$\alpha_2 = 1$
$\alpha_2 = 0$	208 (41.6%)	13 (2.6%)
$\alpha_2 = 1$	24 (4.8%)	255 (51.0%)

The total number of agreements between the two models for α_1 and α_2 was 495 (99.0%) and 463 (92.6%), respectively. Cohen's Kappa for α_1 and α_2 were 0.98, and 0.85, respectively.

TABLE 9 | Profile possession agreement between each pair of models in the operational study.

NRDM	ORDM			
	(0,0)	**(1,0)**	**(0,1)**	**(1,1)**
(0,0)	160 (32.0%)	0	0	0
(1,0)	3 (0.6%)	57 (11.4%)	0	0
(0,1)	1 (0.2%)	0	77 (15.4%)	0
(1,1)	0	0	2 (0.4%)	200 (40.0%)

The total number of profile agreements between the two models was 494 (98.8%), with a Cohen's Kappa of 0.98.

NRDM	MORDM			
	(0,0)	**(1,0)**	**(0,1)**	**(1,1)**
(0,0)	151 (30.2%)	1 (0.2%)	8 (1.6%)	0
(1,0)	0	56 (11.2%)	0	4 (0.8)
(0,1)	8 (1.6%)	0	70 (14.0%)	0
(1,1)	0	16 (3.2%)	1 (0.2%)	185 (37.0%)

The total number of profile agreements between the two models was 462 (92.4%), with a Cohen's Kappa of 0.89.

ORDM	MORDM			
	(0,0)	**(1,0)**	**(0,1)**	**(1,1)**
(0,0)	151 (30.2%)	4 (0.8%)	9 (1.8%)	0
(1,0)	0	53 (10.6%)	0	4 (0.8%)
(0,1)	8 (1.6%)	0	70 (14.0%)	1 (0.2%)
(1,1)	0	16 (3.2%)	0	184 (36.8%)

The total number of profile agreements between the two models was 458 (91.6%), with a Cohen's Kappa of 0.88.

are psychometric models that aim to classify individuals into groups according to their estimated possession status of the measured attributes. Up to this point, polytomous DCMs, such as the NRDM and its special cases and extensions, are designed for nominal (i.e., unordered) responses. Although those DCMs can accommodate ordered response data, they ignore the monotonicity of response option probabilities and require a very large sample size to estimate. The ORDM and the MORDM were introduced in this paper to constrain the NRDM to situations where items are scored on an ordinal scale. Because the ORDM and the MORDM are polytomous extensions of the binary LCDM core, one could easily constrain the ORDM and the MORDM to arrive at other polytomous DCMs. For example, one could replace the LCDM core with the DINA model to arrive at the (modified) polytomous response DINA model.

The analysis of the survey data demonstrated that the proposed models perform similarly to the NRDM but with much fewer parameters to estimate. With four response options in this dataset, the ORDM was 33% smaller than the NRDM. The ORDM will show more comparative advantages if the number of response options increases. If there are seven response options, the ORDM requires estimations of 56 parameters, which is 42% smaller than the NRDM. The MORDM was 54% smaller than the NRDM in this dataset, and it will require only 29 item parameters if there are seven response options, which is 70% smaller than the NRDM. The smaller model sizes of the ORDM and the MORDM comparing to traditional polytomous models allow them to accommodate much smaller sample sizes and thus prove useful in many small-scale testing scenarios.

In addition to their smaller model sizes, the ORDM and the MORDM offer information that can easily capture item characteristics in addition to response option characteristics. In the NRDM, each type of parameters (i.e., intercept, main effects and interactions) is freely estimated for each response option on each item. As a result, it would be

TABLE 10 | ORDM: bias and RMSE of estimated item parameters in the simulation study.

		$\lambda_{0,i,m=1}$	$\lambda'_{0,i,m=2}$	$\lambda'_{0,i,m=3}$	$\lambda_{1,i}$
Bias	Item 1	0.009	−0.307	−0.344	0.308
	Item 2	−0.004	−0.021	−0.048	0.011
	Item 3	−0.047	−0.098	−0.090	0.098
	Item 4	−0.014	−0.115	−0.161	0.132
	Item 5	0.162	−0.101	−0.044	0.089
	Item 6	0.021	−0.020	−0.060	0.045
	Item 7	0.011	−0.015	−0.050	0.046
	Item 8	0.006	−0.431	−0.201	0.289
RMSE	Item 1	0.134	0.576	0.501	0.567
	Item 2	0.122	0.254	0.315	0.225
	Item 3	0.174	0.382	0.407	0.329
	Item 4	0.126	0.325	0.410	0.341
	Item 5	0.381	0.115	0.220	0.249
	Item 6	0.200	0.138	0.269	0.228
	Item 7	0.121	0.203	0.292	0.225
	Item 8	0.149	0.762	0.719	0.715

TABLE 11 | MORDM: bias and RMSE of estimated item parameters in the simulation study.

		$\lambda_{0,i}$	$\lambda_{0,m=1}$	$\lambda'_{0,m=2}$	$\lambda'_{0,m=3}$	$\lambda_{1,i}$
Bias	Item 1	−0.176	0.168	−0.030	−0.039	0.033
	Item 2	−0.163	*	*	*	0.030
	Item 3	−0.207	*	*	*	0.031
	Item 4	−0.154	*	*	*	0.017
	Item 5	−0.172	0.312	0.009	−0.020	0.033
	Item 6	−0.202	*	*	*	0.019
	Item 7	−0.297	*	*	*	0.000
	Item 8	−0.269	*	*	*	0.009
RMSE	Item 1	0.468	0.471	0.162	0.178	0.152
	Item 2	0.491	*	*	*	0.159
	Item 3	0.509	*	*	*	0.156
	Item 4	0.488	*	*	*	0.176
	Item 5	0.175	0.450	0.086	0.127	0.164
	Item 6	0.458	*	*	*	0.137
	Item 7	0.437	*	*	*	0.109
	Item 8	0.410	*	*	*	0.134

The cells with "" indicates that it shares the same parameter with the cell above it.*

easier to discuss the quality of each response option than that of the whole item. In the ORDM, we only have one main effect parameter for each measured attribute representing its effect on the whole item. In the MORDM, we estimate a general intercept parameter: $\lambda_{0,i}$ for each item, representing the general item difficulty. Such item-level information can be helpful for item selection, revision, and reporting.

TABLE 12 | Bias and RMSE of estimated attribute prevalence for the ORDM and MORDM in the simulation study.

Profile	ORDM		MORDM	
	Bias	RMSE	Bias	RMSE
(0,0)	−0.006	0.008	−0.004	0.009
(1,0)	0.000	0.005	0.000	0.007
(0,1)	0.020	0.021	0.018	0.022
(1,1)	−0.014	0.016	−0.014	0.017

We consider the study as one of the first steps incorporating the ordinal response option characteristics into DCMs. A major limitation of this study is that the findings are couched within the particular data used for this study. For future research, we encourage a more robust simulation study examining the performance of the ORDM and the MORDM under a wide range of factors. For example, one could examine the impact of sample sizes on the performance of the new models. We expect that both models can accommodate even smaller sample sizes than the dataset we used in this paper because DCMs, different from multidimensional item response theory models (e.g., Reckase, 1997), do not aim to precisely locate individuals on multiple continua. But this is unknown until tested. We also encourage researchers to investigate the impact of the Q-matrix complexity on the models' performance. Although the increase of Q-matrix complexity generally reduces model performance (e.g., Madison and Bradshaw, 2015; Liu et al., 2017), its impact on the ORDM and the MORDM remains unknown. In addition, we did not assume an ordered sequence on the possession of attributes in this study, although attribute structures can be found in educational and psychological assessment representing the presence of certain attributes given the presence/absence of other attributes (Leighton et al., 2004; Liu and Huggins-Manley, 2016; Liu, 2018). Examining the impact of different attribute structures on the model performance would be of interest. Finally, we used a fully Bayesian approach to estimate the model parameters. Alternatively, one could estimate the parameters via parametric approaches such as the expectation maximization (EM; e.g., Templin and Hoffman, 2013) and the differential evolution optimization (DEoptim; e.g., Jiang and Ma, 2018).

To summarize, the ORDM and the MORDM are psychometric models that can score ordinal item data to classify individuals into latent groups. They are much smaller and thus easier to implement than DCMs for nominal responses. They also offer useful item-level information in addition to option-level information. With the active research and practice in the area of diagnostic measurement, we anticipate that the proposed models will be useful for scoring polytomous item responses in a wide range of educational and psychological assessments.

TABLE 13 | Descriptive statistics of attribute classification accuracy for the ORDM and MORDM in the simulation study.

	ORDM				MORDM			
	Min	**Mean**	**Max**	**SD**	**Min**	**Mean**	**Max**	**SD**
α_1	0.992	0.998	1.000	0.002	0.981	0.995	1.000	0.005
α_2	0.990	0.998	1.000	0.003	0.979	0.993	1.000	0.006

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02512/full#supplementary-material

REFERENCES

Almond, R. G., DiBello, L. V., Jenkins, F., Senturk, D., Mislevy, R. J., Steinberg, L. S., et al. (2001). "Models for conditional probability tables in educational assessment," in *Artificial Intelligence and Statistics 2001: Proceedings of the Eighth International Workshop* (San Francisco, CA: Morgan Kaufmann).

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., and Williamson, D. M. (2015). *Bayesian Networks in Educational Assessment*. New York, NY: Springer.

Almond, R. G., Mulder, J., Hemat, L. A., and Yan, D. (2009). Bayesian network models for local dependence among observable outcome variables. *J. Educ. Behav. Statist.* 34, 491–521. doi: 10.3102/1076998609332751

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika* 43, 561–573.

Blackwell, T., and Case, J. (2008). Test review - strong interest inventory, revised edition. *Rehabil. Couns. Bull.* 51, 122–126. doi: 10.1177/0034355207311350

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 37, 29–51. doi: 10.1007/BF02291411

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* 7, 434–455.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., et al. (2016). Stan: a probabilistic programming language. *J. Statist. Softw.* 20, 1–37. doi: 10.18637/jss.v076.i01

Chen, J., and de la Torre, J. (2018). Introducing the general polytomous diagnosis modeling framework. *Front. Psychol.* 9:1474. doi: 10.3389/fpsyg.2018.01474

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

Culpepper, S. A., and Hudson, A. (2018). An improved strategy for bayesian estimation of the reduced reparameterized unified model. *Applied psychological measurement,* 42, 99–115. doi: 10.1177/0146621617707511

da Silva, M. A., de Oliveira, E. S. B., von Davier, A. A., and Bazán, J. L. (2017). Estimating the DINA model parameters using the No-U-Turn Sampler. *Biom. J.* 60, 352–368. doi: 10.1002/bimj.201600225

de la Torre, J. (2010). The Partial-Credit DINA Model. *Paper Presented at the International Meeting of the Psychometric Society* (Athens, GA).

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

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statist. Comput.* 24, 997–1016. doi: 10.1007/s11222-013-9416-2

Gelman, A., and Rubin, D. B. (1992), Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–511.

Girolami, M., and Calderhead, B. (2011). Riemann manifold langevin and hamiltonian monte carlo methods. *J. R. Statist. Soc.* 73, 123–214. doi: 10.1111/j.1467-9868.2010.00765.x

Hansen, M. (2013). *Hierarchical Item Response Models for Cognitive Diagnosis.* Unpublished doctoral dissertation. University of California at Los Angeles.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74, 191–210. doi: 10.1007/s11336-008-9089-5

Jiang, Z., and Carter, R. (2018). Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnosis model via Stan. *Behav. Res. Methods*, 1–12. doi: 10.3758/s13428-018-1069-9

Jiang, Z., and Ma, W. (2018). Integrating differential evolution optimization to cognitive diagnosis model estimation. *Front. Psychol.* 9:2142. doi: 10.3389/fpsyg.2018.02142

Jiang, Z., and Skorupski, W. (2017). A Bayesian approach to estimating variance components within a multivariate generalizability theory framework. *Behav. Res. Method* 50, 1–22. doi: 10.3758/s13428-017-0986-3

Jiang, Z., and Templin, J. (2018). Constructing Gibbs Samplers for Bayesian Logistic Item Response Models. *Multivar. Behav. Res.* 53:132. doi: 10.1080/00273171.2017.1404897

Junker, B. W., Patz, R. J., and Van Houdnos, N. M. (2016). "Markov chain Monte Carlo for item response models," in *Handbook of Item Response Theory, Volume 2: Statistical Tools,* ed. W. J. van der Linden, 271–325. Available online at: https://www.crcpress.com/Handbook-of-Item-Response-Theory-Volume-Two-Statistical-Tools/Linden/p/book/9781466514324

Leighton, J. P., Gierl, M. J., and Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: a variation on Tatsuoka's rule-space approach. *J. Educ. Meas.* 41, 205–237. doi: 10.1111/j.1745-3984.2004.tb01163.x

Levy, R., and Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *Int. J. Test.* 4, 333–369. doi: 10.1207/s15327574ijt0404_3

Liu, R. (2018). Misspecification of attribute structure in diagnostic measurement. *Educ. Psychol. Meas.* 78, 605–634. doi: 10.1177/0013164417702458

Liu, R., and Huggins-Manley, A. C. (2016). "The specification of attribute structures and its effects on classification accuracy in diagnostic test design," in *Quantitative Psychology Research,* eds. L. A. van der Ark, D. M. Bolt, W. -C. Wang, J. A. Douglas, and M. Wiberg (New York, NY: Springer), 243–254.

Liu, R., Huggins-Manley, A. C., and Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educ. Psychol. Meas.* 77, 220–240. doi: 10.1177/0013164416645636

Liu, R., Qian, H., Luo, X., and Woo, A. (2018). Relative diagnostic profile: a subscore reporting framework. *Educ. Psychol. Meas.* 78, 1072–1088. doi: 10.1177/0013164417740170

Luo, Y., and Al-Harbi, K. (2017). Performances of LOO and WAIC as IRT model selection methods. *Psychol. Test Assess. Model.* 59:183. Available online at: https://search.proquest.com/openview/689245ac87d24c062121748c85998b5b/1?pq-origsite=gscholar&cbl=43472

Luo, Y., and Jiao, H. (2018). Using the Stan program for Bayesian item response theory. *Educ. Psychol. Meas.* 78, 384–408. doi: 10.1177/0013164417693666

Ma, W., and de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *Br. J. Math. Statist. Psychol.* 69, 253–275. doi: 10.1111/bmsp.12070

Madison, M. J., and Bradshaw, L. P. (2015). The effects of Q-Matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educ. Psychol. Meas.,* 75, 491–511. doi: 10.1177/0013164414539162

McGhee, R. L., Ehrler, D. J., and Buckhalt, J. A. (2007). *FFPI-C: Five-factor Personality Inventory-Children*. Pro-Ed. Available online at: https://psycentre.apps01.yorku.ca/wp/five-factor-personality-inventory-children-ffpi-c/

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Appl. Psychol. Meas.* 16, 159–176. doi: 10.1177/014662169201600206

Prince, J. R. (1998). Interpreting the strong interest inventory: a case study. *Career Dev. Q.* 46, 339–346.

R Core Team (2018). *R (Version 3.5) [Computer Software]*. Vienna: R Foundation for Statistical Computing.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Appl. Psychol. Meas,* 21, 25–36.

Revuelta, J., and Ximénez, C. (2017). Bayesian dimensionality assessment for the multidimensional nominal response model. *Front. Psychol.* 8:961. doi: 10.3389/fpsyg.2017.00961

Robitzsch, A., Kiefer, T., George, A. C., and Uenlue, A. (2018). *CDM: Cognitive Diagnosis Modeling*. R package version 6.1. Available Online at: http://CRAN.R-project.org/package=CDM

Rupp, A. A., Templin, J., and Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: Guilford Press.

Rupp, A. A., and Templin, J. L. (2008). Unique characteristics of diagnostic classification models: a comprehensive review of the current state-of-the-art. *Measurement* 6, 219–262. doi: 10.1080/15366360802490866

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometr. Monogr.* 17, 1–169. doi: 10.1007/BF03372160

Staggs, G. D. (2004). *Meta-analyses of Interest-Personality Convergence Using the Strong Interest Inventory and the Multidimensional Personality Questionnaire*. Unpublished doctoral dissertation, Iowa State University.

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.* 20, 345–354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Templin, J., and Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educ. Meas.* 32, 37–50. doi: 10.1111/emip.12010

Templin, J. L., Henson, R. A., Rupp, A. A., Jang, E., and Ahmed, M. (2008). Cognitive Diagnosis Models For Nominal Response Data. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education* (New York, NY).

Thissen, D. (1991). *MULTILOG, 6.0*. Chicago, IL: Scientific Software.

Tu, D., Zheng, C., Cai, Y., Gao, X., and Wang, D. (2017). A polytomous model of cognitive diagnostic assessment for graded data. *Int. J. Test*. 18, 1–21. doi: 10.1080/15305058.2017.1396465

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statist. Comput*. 27, 1413–1432. doi: 10.1007/s11222-016-9696-4

von Davier, M. (2008). A general diagnostic model applied to language testing data. *Br. J. Math. Statist. Psychol.* 61, 287–307. doi: 10.1348/000711007X193957

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Using stacking to average Bayesian predictive distributions. *Bayesian Anal.* 13, 917–1007.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.