Diagnostic Classification Models for Polytomous Item Responses

Ren Liu

University of California, Merced

Zhehan Jiang

University of Alabama

Liu, R., Jiang, Z. (2018). Diagnostic classification models for polytomous item responses. *Manuscript Submitted for Publication and Currently Under Review*.

Abstract

The purpose of this study is to develop and evaluate two diagnostic classification models (DCMs) for scoring ordinal item data. We 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. 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, polytomous item responses, partial credit model, rating scales, nominal response diagnostic model

Diagnostic Classification Models for Polytomous Item Responses

Grouping people into different categories are often of interest in educational and psychological tests. For example, the Five Factor Personality Inventory-Children (McGhee, Ehrler, & Buckhalt, 2007) aims to identify which personalities a child possesses. In another case of career assessment, the Strong Interest Inventory (Blackwell & Case, 2008; Prince, 1998; Staggs, 2004) 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 polytomous 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, Templin, & Henson, 2010). However, most current DCMs for polytomous items consider item response categories as nominal without using the ordered category information (e.g., de la Torre, 2010; Ma & de la Torre, 2016; Templin et al., 2008). 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 polytomous DCMs that are designed to score individuals on an ordered 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. Finally, we discuss the application and advantages of the models and offer future research recommendations.

Review of Current Polytomous DCMs

DCMs are 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 $k = 1, \dots, K$ to index latent traits and $\boldsymbol{\alpha}_c = \{\alpha_1, \dots, \alpha_K\}$ 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.

Up to this point, five DCMs have been developed to score polytomous data: (1) the nominal response diagnostic model (NRDM; Templin et al., 2008), (2) the general diagnostic model (GDM; von Davier, 2008), (3) the partial-credit deterministic input noisy "and" gate (PC-DINA; de la Torre, 2010) model, (4) the sequential generalized DINA model (SG-DINA; Ma & de la Torre, 2016), and (5) the polytomous log-linear cognitive diagnosis model (P-LCDM; Hansen, 2013). The first four models 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. For example, the NRDM defines the probability of individuals in latent class *c* selecting response option *m* on item *i*, such that

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

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

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

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.

The NRDM builds on its dichotomous core: the LCDM, and it shares commonalities with the other four polytomous DCMs introduced above. For example, the polytomous GDM is equivalent to the NRDM, the PC-DINA is a special case of the NRDM where the former extends the dichotomous DINA model, the SG-DINA extends the NRDM through adding a processing function to accommodate option-attribute relationship, and the P-LCDM also builds upon the LCDM and models the differences between cumulative probabilities of adjacent options. Except for the P-LCDM, the first four models accommodate nominal response options so that each response option has its own set of parameters.

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

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

for the NRM,

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

for the GPCM, and

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

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 polytomous 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 Polytomous Response Diagnostic Model (PRDM) and the Modified Polytomous Response Diagnostic Model (MPRDM). The PRDM is defined as

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

where $\boldsymbol{\lambda}_i^T \mathbf{h}(\boldsymbol{\alpha}_c, \mathbf{q}_i) = \sum_{k=1}^{K} \lambda_{1,i,k}(\alpha_{c,k} q_{i,k}) + \sum_{k=1}^{K-1} \sum_{k'=K+1}^{K} \lambda_{2,i,k,k'}\left(\alpha_{c,k}\alpha_{c,k'} q_{i,k} q_{i,k'}\right) + \cdots$. For identifiability purposes, we impose three sets of constraints on the PRDM. 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 greater than 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

$$\lambda_{0,i,m-1} \leq \lambda_{0,i,m} \forall \; i, m \; .$$

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

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

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 $\boldsymbol{\lambda}_{i,2}^{T}\mathbf{h}(\alpha_c, \mathbf{q}_i)$ in Equation 2 is replaced by $2 \times \boldsymbol{\lambda}_i^T\mathbf{h}(\boldsymbol{\alpha_c}, \mathbf{q}_i)$ in Equation 7. It

should be clear now that the proposed PRDM 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 MPRDM is defined the same as the PRDM 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, where an entry $w_{iv} = 1$ when item *i* measures attribute set *v*, and 0 otherwise. By definition, 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 MPRDM, 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 MPRDM as

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

The constraints we impose on the MPRDM is the same as those on the PRDM, except that the third constraint (i.e., for the intercept parameters) needs to be adapted to the MPRDM. In the MPRDM, 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 MPRDM in such case is expressed as

$$P(X_i = 2|\boldsymbol{\alpha}_c) = \tag{9}$$

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

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 PRDM, and further constrain the intercept parameters of the PRDM to arrive at the MPRDM.

Operational Study

The purpose of this operational study is to compare the performance of the PRDM and the MPRDM with the NRDM through fitting these three models to an ordinal item response dataset. The motivating research question was: can the more parsimonious PRDM and/or the MPRDM 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 agreements, 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. 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 1. A brief look of Table 1 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.

Analysis

Before fitting the models to the dataset, we first examined the multidimensionality of the dataset because fitting a DCM won't be helpful if the dataset is mostly unidimensional. We conducted a factor analysis and the results suggested a two-factor solution with two eigenvalues larger than 1 (i.e., 4.29 and 1.35). This confirmation of the multidimensionality of the dataset allows us to continue the analysis.

Parameters were estimated through implementing Markov Chain Monte Carlo (MCMC) algorithms in Stan (Carpenter et al., 2016). For each model, we ran two Markov chains with random starting values. The total length of the MCMC 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 PRDM to both the NRDM and the PRDM, and applied the constraints for the MPRDM to itself through specifying pseudo response option parameters such that

$$\lambda_{z,m=m} = \lambda_{z,m=1} + \lambda'_{z,m=2} + \cdots + \lambda'_{z,m=m} \; \forall \; z, m. \tag{10}$$

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, Hwang, and Vehtari (2014), Vehtari, Gelman, and Gabry (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).

Results

We estimated 48, 32, and 22 item parameters for the NRDM, the PRDM and the MPRDM, respectively. For this dataset, the PRDM was 33% smaller than the NRDM, and the MPRDM was 54% smaller than the NRDM. We first examined the results on model fit indices and listed the ELPD and LOOIC estimates and standard errors for each model in Table 2. For each index, smaller values indicate better fit. Although both indices suggested better fit for the PRDM 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 MPRDM or the best fitting PRDM for further interpretations.

Examining profile prevalence estimates provides further evidence about the similar performance of the three models. Table 3 lists the estimates and standard errors 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 PRDM 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 MPRDM were all smaller than 0.02 for every profile.

We could also look into the similarities of the item parameter estimates. Tables 4, 5, and 6 display the item parameter estimates and their standard errors for the NRDM, the PRDM, and the MPRDM, 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 MPRDM 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, PRDM and MPRDM.

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 2 (measuring math self-concept), and items 5 and 6 (measuring math joy) 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 PRDM were even more alike. Comparing across items, items 1 and 2's ROCs were similar to each other; items 5 and 6's ROCs were similar to each other. 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 2,

individuals with the math self-concept have higher probabilities selecting response option 1 and above than those without the math self-concept. For items 3 and 4, individuals with math joy have higher probabilities selecting response option 2 and above than those without it.

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. Tables 7, 8 and 9 cross-tabulate the attribute classification agreement between each pair of models, respectively. The agreement rates between the NRDM and the PRDM 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 PRDM and the MPRDM. Table 10, 11 and 12 cross-tabulate the profile classification agreement between each pair of models, respectively. 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 PRDM. The continuous scores are individuals' marginal probabilities of possessing each attribute. 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 PRDM, 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 display the root-mean-square deviation (RMSD) for scores between each pair of models in Table 13. Results show that the score differences were very small between the models, and we conclude that the three models perform similarly.

Discussion

Scoring items in a polytomous 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 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 PRDM and the MPRDM were introduced in this paper to constrain the NRDM to situations where items are scored on an ordered scale. Because the PRDM and the MPRDM are polytomous extensions of the binary LCDM core, one could easily constrain the PRDM and the MPRDM 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 PRDM was 33% smaller than the NRDM. The PRDM will show more comparative advantages if the number of response options increases. If there are seven response options, the PRDM requires estimations of 56 parameters, which is 42% smaller than the NRDM. The MPRDM 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 PRDM and the MPRDM 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 PRDM and the MPRDM 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 easier to discuss the quality of each response option than that of the whole item. In the PRDM, we only have one main effect parameter for each measured attribute representing its effect on the whole item. In the MPRDM, 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.

For future research, it would be helpful to 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 & Bradshaw, 2015), its impact on the PRDM and the MPRDM remains unknown.

To summarize, the PRDM and the MPRDM are new 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 1

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%)

Table 2

Model Fit Information

	NRDM		PRDM		MPRDM	
	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 3

Profile Prevalence Estimates and Standard Errors under the NRDM, the PRDM and the MPRDM

Profile	NRDM	PRDM	MPRDM
(0,0)	0.346(0.005)	0.351(0.001)	0.351(0.001)
(1,0)	0.084(0.004)	0.074(0.001)	0.105(0.001)
(0,1)	0.147(0.003)	0.156(0.001)	0.125(0.001)
(1,1)	0.424(0.003)	0.419(0.001)	0.418(0.001)

Table 4

NRDM: Item Parameter Estimates and Standard Errors

	$\lambda_{0,i,m=1}$	$\lambda'_{0,i,m=2}$	$\lambda'_{0,i,m=3}$	$\lambda_{1,i,m=1}$	$\lambda'_{1,i,m=2}$	$\lambda'_{1,i,m=3}$
Item 1	-0.423	-10.710	-12.578	3.619	10.339	10.981
	(0.029)	(0.971)	(0.532)	(0.054)	(0.963)	(0.528)
Item 2	-0.934	-2.265	-1.927	2.149	1.995	0.744
	(0.032)	(0.024)	(0.076)	(0.019)	(0.027)	(0.077)
Item 3	-1.557	-6.900	-10.818	2.714	6.354	10.481
	(0.028)	(1.030)	(0.879)	(0.022)	(1.026)	(0.878)
Item 4	0.269	-4.540	-3.491	3.886	5.319	2.507
	(0.008)	(2.088)	(1.351)	(0.779)	(2.070)	(1.345)
Item 5	1.986	-0.036	-1.241	16.908	2.756	2.130
	(0.015)	(0.002)	(0.013)	(0.607)	(0.025)	(0.015)
Item 6	1.347	-0.716	-2.305	17.927	2.780	2.322
	(0.011)	(0.008)	(0.024)	(0.575)	(0.017)	(0.024)
Item 7	0.296	-1.850	-2.255	0.926	2.857	1.940
	(0.008)	(0.023)	(0.062)	(0.018)	(0.028)	(0.063)
Item 8	0.642	-10.306	-11.038	18.139	12.357	10.716
	(0.007)	(0.717)	(0.696)	(0.589)	(0.714)	(0.695)

Table 5

PRDM: Item Parameter Estimates and Standard Errors

	$\lambda_{0,i,m=1}$	$\lambda'_{0,i,m=2}$	$\lambda'_{0,i,m=3}$	$\lambda_{1,i}$
Item 1	-0.390	-4.088	-5.321	3.735
	(0.008)	(0.048)	(0.056)	(0.071)
Item 2	-0.834	-2.181	-3.127	1.971
	(0.008)	(0.016)	(0.018)	(0.014)
Item 3	-1.533	-3.377	-3.194	2.841
	(0.013)	(0.027)	(0.023)	(0.022)
Item 4	0.289	-2.180	-3.784	2.902
	(0.012)	(0.034)	(0.044)	(0.038)
Item 5	1.991	-0.029	-1.352	2.290
	(0.016)	(0.002)	(0.015)	(0.011)
Item 6	1.352	-0.670	-2.586	2.626
	(0.011)	(0.007)	(0.014)	(0.015)
Item 7	0.071	-1.370	-2.297	2.039
	(0.007)	(0.012)	(0.018)	(0.013)
Item 8	0.646	-10.395	-12.733	12.427
	(0.007)	(0.772)	(0.764)	(0.764)

Table 6

MPRDM: Item Parameter Estimates and Standard Errors

	$\lambda_{0,i}$	$\lambda_{0,m=1}$	$\lambda'_{0,m=2}$	$\lambda'_{0,m=3}$	$\lambda_{1,i}$
Item 1	5.834 (0.493)	-6.204 (0.491)	-2.871 (0.021)	-3.781 (0.021)	2.472 (0.017)
Item 2	5.141 (0.493)	*	*	*	2.512 (0.017)
Item 3	4.491 (0.494)	*	*	*	2.732 (0.015)
Item 4	6.424 (0.493)	*	*	*	3.200 (0.018)
Item 5	10.313 (1.237)	-7.893 (1.233)	-0.527 (0.007)	-2.111 (0.009)	3.262 (0.013)
Item 6	9.339 (1.225)	*	*	*	2.348 (0.009)
Item 7	7.948 (1.229)	*	*	*	1.622 (0.007)
Item 8	8.285 (1.241)	*	*	*	2.033 (0.007)

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

Table 7

Attribute Possession Agreement between the NRDM and the PRDM

NRDM	PRDM	
	$\alpha_1 = 0$	$\alpha_1 = 1$
$\alpha_1 = 0$	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%)

Note: The total number of agreement 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.

Table 8

Attribute Possession Agreement between the NRDM and the MPRDM

NRDM	MPRDM	
	$\alpha_1 = 0$	$\alpha_1 = 1$
$\alpha_1 = 0$	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%)

Note: The total number of agreement 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.

Table 9

Attribute Possession Agreement between the PRDM and the MPRDM

PRDM	MPRDM	
	$\alpha_1 = 0$	$\alpha_1 = 1$
$\alpha_1 = 0$	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%)

Note: The total number of agreement 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 10

Profile Possession Agreement between the NRDM and the PRDM

NRDM	PRDM			
	(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%)

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

Table 11

Profile Possession Agreement between the NRDM and the MPRDM

NRDM	MPRDM			
	(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%)

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

Table 12

Profile Possession Agreement between the PRDM and the MPRDM

PRDM	MPRDM			
	(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%)

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

Table 13

RMSD for Continuous Scores Between Each Pair of Models

	NRDM-PRDM	NRDM-MPRDM	PRDM-MPRDM
α_1	0.04	0.06	0.06
α_2	0.03	0.15	0.16

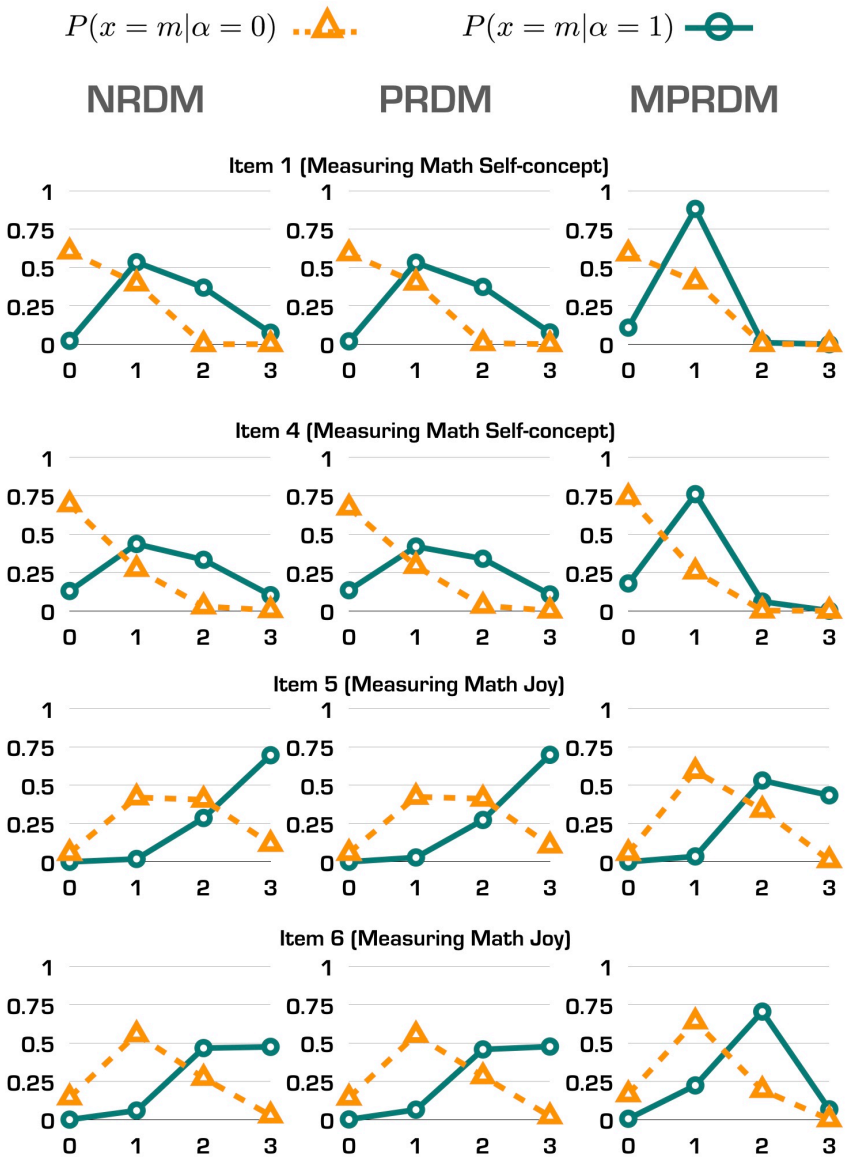

Figure 1. Response option curves for selected items.

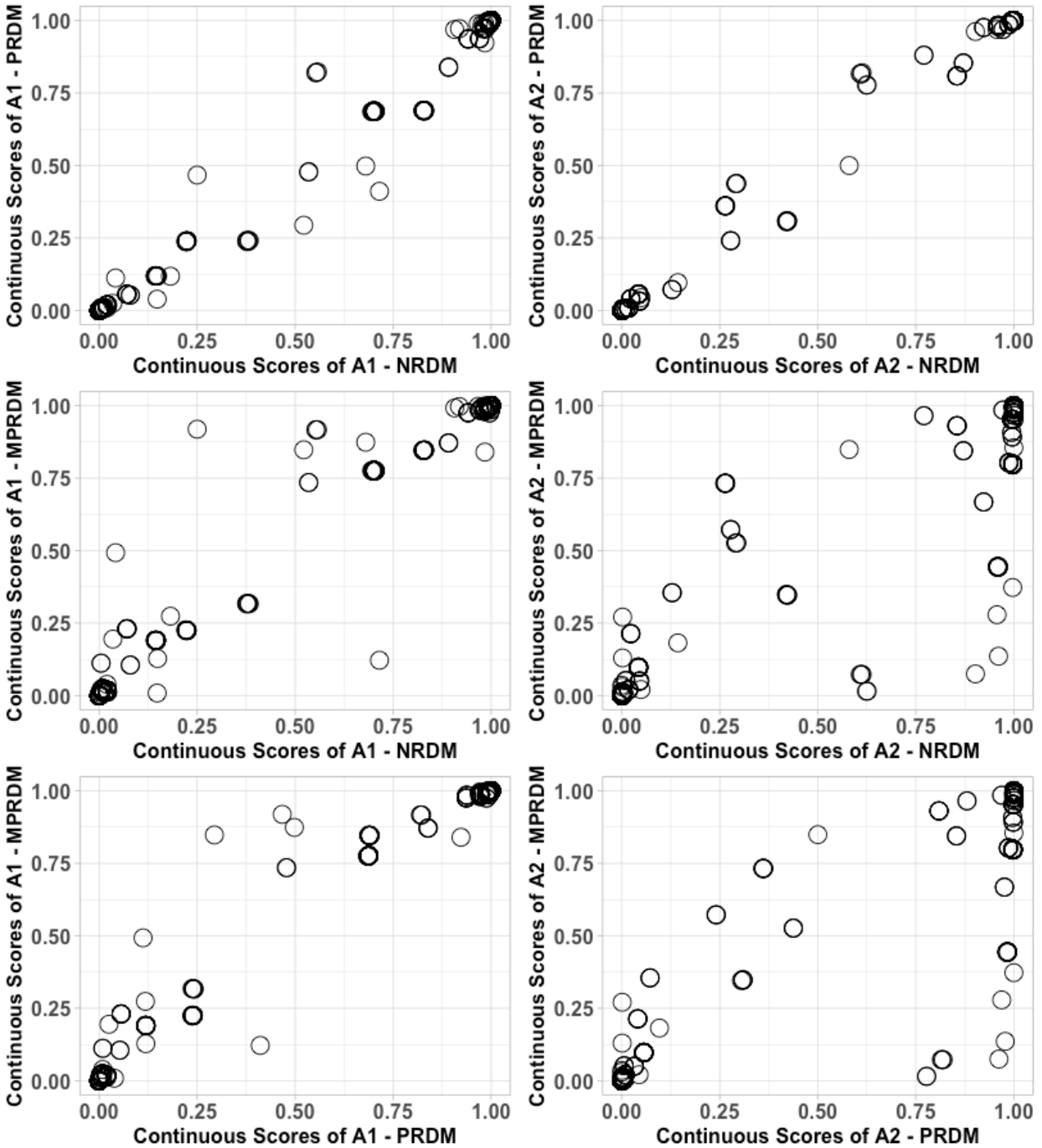

Figure 2. Comparison of continuous scores.

References

Blackwell, T., & Case, J. (2008). "Test Review - Strong Interest Inventory, Revised Edition". *Rehabilitation Counseling Bulletin*. *51* (2): 122–126.

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

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7,* 434–455.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... & Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*(2), 1-37.

de la Torre, J. (2010, July). *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.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, *24*(6), 997-1016.

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., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191-210.

Huggins-Manley, A. C., & Algina, J. (2015). The Partial Credit Model and Generalized Partial Credit Model as Constrained Nominal Response Models, With Applications in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(2), 308-318.

Junker, B. W., Patz, R. J., & Van Houdnos, N. M. (2016). Markov chain Monte Carlo for item response models. *Handbook of Item Response Theory, Volume Two: Statistical Tools*, *21*, 271-325.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 253-275.

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-Matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement, 75(3),* 491-511.

McGhee, R. L., Ehrler, D. J., & Buckhalt, J. A. (2007). *FFPI-C: Five-factor Personality Inventory-Children*. Pro-Ed.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176

Prince, J.R. (1998). "Interpreting the Strong Interest Inventory: A case study". *The Career Development Quarterly*, 46(4), 339–346.

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

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25-36.

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

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 219-262.

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

Staggs, G. D. (2003). *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. *Journal of Educational Measurement*, *20*(4), 345-354.

Templin, J. L., Henson, R. A., Rupp, A. A., Jang, E., & Ahmed, M. (2008, March). *Cognitive diagnosis models for nominal response data*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37-50.

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

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*. 27(5), 1413--1432. doi:10.1007/s11222-016-9696-4

von Davier, M. (2008). A general diagnostic model applied to language testing data. British *Journal of Mathematical and Statistical Psychology, 61*, 287–307.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions. *Bayesian Analysis*.