Behavior Research Methods (2020) 52:422–439
https://doi.org/10.3758/s13428-019-01239-9

A general diagnostic classification model for rating scales

Ren Liu[1] · Zhehan Jiang[2]

Published online: 25 April 2019

Abstract
This study proposes and evaluates a general diagnostic classification model (DCM) for rating scales. We applied the proposed model to a dataset to compare its performance with traditional DCMs for polytomous items. We also conducted a simulation study based on the applied study condition in order to evaluate the parameter recovery of the proposed model. The findings suggest that the proposed model shows promise for (1) accommodating much smaller sample sizes by reducing a large number of parameters for estimation; (2) obtaining item category response probabilities and individual scores very similar to those from a traditional saturated model; and (3) providing general item information that is not available in traditional DCMs for polytomous items.

Keywords Diagnostic classification model · Rating scales · Psychological assessments · Polytomous item responses · Nominal response diagnostic model

Understanding whether an individual possesses certain characteristics or has developed certain behaviors is a common objective in psychological testing. Recently, the diagnostic classification model (DCM), a newer type of model that aims to group individuals according to their possession/nonpossession of multiple latent traits, has been proposed in the literature (Rupp, Templin, & Henson, 2010). DCMs have been mostly discussed and used in the context of dichotomous items. Example dichotomous DCMs include the deterministic inputs, noisy "AND" gate (DINA; Haertel, 1989) model, the generalized DINA model (G-DINA; de la Torre, 2011), and the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009). Only a few DCMs that have been developed for scoring polytomous items. Examples include the nominal response diagnostic model (NRDM; Templin, Henson, Rupp, Jang, & Ahmed, 2008) and its special case: the partial-credit DINA (PC-DINA; de la Torre, 2010) model. These polytomous DCMs can score Likert-scale items on a personality test, constructed-response items on a writing exam, or a multiple-step item on a math test. However, one major limitation of the current polytomous DCMs is that those models usually contain a large number of parameters, and thus require a large sample size in order to provide stable estimates. For example, the NRDM was recommended to fit with at least 5,000 individuals in Templin et al. (2008). This type of sample size is often hard to achieve for a classroom test and small-scale psychological assessments. If polytomous models cannot provide stable estimates with smaller sample sizes, a common compromise is to recode the polytomous item responses into multiple dichotomous responses (e.g., Drasgow, 1995; Thorpe & Favia, 2012). The results of such recoding include loss of information and violation of the original score interpretation in the principled assessment framework.

The purpose of this study was to introduce a new polytomous DCM, called the *rating scale diagnostic model (RSDM)*, that contains fewer parameters and thus can accommodate smaller sample sizes. Traditional polytomous DCMs require free estimation of the parameters associated with each response category of each item. In contrast, the most important feature of the RSDM is that it allows category threshold parameters to be shared across items that measure the same set of attributes. In this article, we first explain the concept of shared category threshold parameters and the development of the RSDM. We then fit the model to an operational dataset, to compare the parameter estimates and examinee classifications to those of the NRDM. Next, we conduct a small simulation study based on the operational study condition, to assess the parameter recovery of the model. Finally, we discuss the application of the model and future research recommendations.

✉ Ren Liu
rliu45@ucmerced.edu

Zhehan Jiang
zjiang17@ua.edu

[1] Psychological Sciences, University of California, Merced, CA, USA

[2] University Libraries, University of Alabama, Montgomery, AL, USA

Background and model development

The model introduced in this study was developed on the basis of (1) the concept of shared category threshold parameters, demonstrated in earlier item response theory (IRT) models (Andrich, 1978; Muraki, 1990), and (2) the approach of representing multidimensionality through latent classes in DCMs (Rupp & Templin, 2008; Templin & Hoffman, 2013).

Shared category threshold parameters

Among polytomous IRT models, the modified graded response model (MGRM; Muraki, 1990) is an example in which the category threshold parameters are shared across items. The MGRM is a special case of the graded response model (GRM; Samejima, 1969). This study was inspired by the relationship between these two models.

The GRM is defined as

$$P(X_i \geq m|\theta_e) = \frac{exp[d_i(\theta_e - b_{im})]}{1 + exp[d_i(\theta_e - b_{im})]}, \quad (1)$$

with $P(X_i \geq 0|\theta_e) = 1$, where $i = 1, \ldots, I$ index items, $m = 1, \ldots, M-1$ index category thresholds for M response options, d_i is a discrimination parameter for item i, and b_{im} is a category threshold parameter for option m of item i. The MGRM is defined as

$$P(X_i \geq m|\theta_e) = \frac{exp[d_i(\theta_e - b_i + t_m)]}{1 + exp[d_i(\theta_e - b_i + t_m)]}, \quad (2)$$

where t_m is a category threshold parameter for option m across all items. Comparing Eq. 1 to Eq. 2, the category threshold parameter b_{im} is decomposed into two components: b_i (the item general location parameter) and t_m (the threshold parameter for category m common to all items). In other words, instead of freely estimating one parameter at the intersection of each item and each threshold, the MGRM has only one parameter for each threshold, and that set of threshold parameters is applied to all items. This assumes that the relative difficulty between steps is held constant across items. For example, Likert scales with the same response categories across items (e.g., *strongly disagree*, *disagree*, *agree*, *strongly agree*) may naturally have this feature, while other item types may not.

One benefit of using models with shared threshold parameters is the requirement for smaller sample sizes, because such models require estimation of fewer parameters than their equivalent models with freely estimated threshold parameters. For a 20-item test with five response options, the GRM requires the free estimation of $20 \times 4 = 80$ threshold parameters, whereas the MGRM requires only $20 + 4 = 24$ threshold parameters.

In addition to the MGRM, other polytomous IRT models have the feature of shared threshold parameters. For example, the relationship between the partial-credit model (PCM; Masters, 1982) and the rating scale model (RSM; Andrich, 1978) is similar to that between the GRM and the MGRM. With the PCM, the threshold parameters may vary across items. In contrast, the threshold parameters in the RSM are shared across all items.

Model development

Here we will first introduce the current polytomous DCMs, and then we will present how we apply constraints of the shared thresholds in the NRDM.

The NRDM, PC-DINA, and the SG-DINA can accommodate both nominal and ordered response items. All three models can be equivalent when certain constraints apply (Ma & de la Torre, 2016). The PC-DINA builds on the binary DINA model, whereas both the NRDM and SG-DINA build on the binary LCDM. SG-DINA adds a processing function to the NRDM, so that each step category within an item can measure different latent traits, whereas the NRDM cannot. We developed the RSDM, in which parameter constraints are applied to the NRDM, although the processing function in SG-DINA can also easily be applied to the RSDM, if desired.

Let $k = 1, \ldots, K$ index latent traits, commonly called *attributes* in DCM studies. The individuals in latent class c have an attribute profile $\boldsymbol{\alpha}_c = \{\alpha_1, \ldots, \alpha_K\}$. The NRDM is defined as

$$P(X_i = m|\boldsymbol{\alpha}_c) = \frac{exp\left[\lambda_{0,i,m} + \boldsymbol{\lambda}_{i,m}^T \mathbf{h}(\boldsymbol{\alpha}_c, \mathbf{q}_i)\right]}{\sum_{m=0}^{M-1} exp\left[\lambda_{0,i,m} + \boldsymbol{\lambda}_{i,m}^T \mathbf{h}(\boldsymbol{\alpha}_c, \mathbf{q}_i)\right]}, \quad (3)$$

where $\boldsymbol{\lambda}_{i,m}^T \mathbf{h}(\boldsymbol{\alpha}_c, \mathbf{q}_i) = \sum_{k=1}^{K} \lambda_{1,i,k,m}(\alpha_{c,k} q_{i,k}) + \sum_{k=1}^{K-1} \sum_{k'=K+1}^{K} \lambda_{2,i,k,k',m}\left(\alpha_{c,k}\alpha_{c,k'} q_{i,k} q_{i,k'}\right) + \ldots$. In Eq. 3, $\lambda_{0,\,i,\,m}$ is the intercept for category m in item i; $\boldsymbol{\lambda}_{i,\,m}$ is a vector of coefficients representing the attribute effects on responding to category m in item i; and $\mathbf{h}(\boldsymbol{\alpha}_c, \mathbf{q}_i)$ is a set of linear combination of $\boldsymbol{\alpha}_c$ and $\mathbf{q}_i$ defined by the **Q**-matrix (Tatsuoka, 1983). The **Q**-matrix is an item-by-attribute incidence matrix in which an entry $q_{i,\,k}$ equals 1 if item i measures attribute k, and 0 otherwise. Note that each parameter has a subscript "i, m," meaning that the parameter is freely estimated at the intersection of each item and each category threshold, for the intercept, main effects, and higher order effects.

We propose the RSDM, which can be expressed as a constrained version of the NRDM, analogous to the description of the MRGM as a constrained version of the GRM. However, because the NRDM is a multidimensional model, we cannot assume that one set of shared category parameters

can be applied to all the items. Instead, we propose to fix the category threshold parameter (also known as the step parameter) for each dimension. The dimensions in DCMs are represented by latent classes (i.e., different attribute combinations). In other words, we propose to constrain the step parameters of items measuring the same set of attributes.

Table 1 presents an example **Q**-matrix capturing the relationship between 12 items and three attributes. In this example, Items 1 and 2 are both measuring α_1, and Items 7 and 8 are measuring the same set of attributes α_1 and α_2. In the RSDM, the step parameters for Items 1 and 2 are constrained to be the same, and the step parameters for Items 7 and 8 are constrained to be the same. Let $v = 1, \ldots, V$ index attribute combinations. We can capture the information of which items measure which attribute combinations in an item-by-attribute-set matrix, called a **W**-matrix. Table 2 is a **W**-matrix obtained from Table 1. In a **W**-matrix, w_{iv} equals 1 if item i measures attribute combination v, and 0 otherwise. By definition, there is only one entry of "1" in each row for each item i. The benefit of utilizing the **W**-matrix is that we can assign step parameters to each attribute set (v) instead of each item (i). Utilizing the **W**-matrix, the RSDM is defined as

$$P(X_i = m|\boldsymbol{\alpha}_c) = \frac{exp\left[\lambda_{0,i} + \sum_{v=1}^{V}\lambda_{0,m,v}w_{iv} + \left(\boldsymbol{\lambda}_i + \sum_{v=1}^{V}\boldsymbol{\lambda}_{i,m,v}w_{iv}\right)^T \mathbf{h}(\boldsymbol{\alpha}_c, \mathbf{q}_i)\right]}{\sum_{m=0}^{M-1} exp\left[\lambda_{0,i} + \sum_{v=1}^{V}\lambda_{0,m,v}w_{iv} + \left(\boldsymbol{\lambda}_i + \sum_{v=1}^{V}\boldsymbol{\lambda}_{i,m,v}w_{iv}\right)^T \mathbf{h}(\boldsymbol{\alpha}_c, \mathbf{q}_i)\right]}, \quad (4)$$

where $\left(\boldsymbol{\lambda}_i + \sum_{v=1}^{V}\boldsymbol{\lambda}_{i,m,v}w_{iv}\right)^T \mathbf{h}(\boldsymbol{\alpha}_c, \mathbf{q}_i) = \sum_{k=1}^{K}\left(\lambda_{1,i,\ k} + \sum_{v=1}^{V}\lambda_{1,m,v}w_{iv}\right)\left(\alpha_{c,k}q_{i,k}\right) + \sum_{k=1}^{K-1}\sum_{k'=K+1}^{K}\left(\lambda_{2,i,k,k'} + \sum_{v=1}^{V}\lambda_{2,m,v}w_{iv}\right)(\alpha_c\ , k\alpha_{c,k'}q_{i,k}q_{i,k'}) \ldots$. In the RSDM, we set four constraints for identifiability purposes. First, we avoid an infinite number of solutions by adopting Thissen's (1991) approach, in which the parameters for the first category in an item are fixed to 0, such that

$$\lambda_{0,i} + \sum_{v=1}^{V}\lambda_{0,m=0,v}w_{iv} = 0 \forall i,$$
$$\lambda_{1,i,k} + \sum_{v=1}^{V}\lambda_{1,m=0,v}w_{iv} = 0 \forall i,k,$$
$$\lambda_{2,i,k,k'} + \sum_{v=1}^{V}\lambda_{2,m=0,v}w_{iv} = 0 \forall i,k,k',$$

and for all higher-order interactions. Second, the possession of more attributes monotonically increases the log-odds of the probability of a higher response category. In the RSDM, we constrain the main and interaction effect parameters to be greater than 0, such that

$$\lambda_{1,i,k} + \sum_{v=1}^{V}\lambda_{1,m,v}w_{iv} \geq 0 \forall k,m,v,$$
$$\lambda_{2,i,k,k'} + \sum_{v=1}^{V}\lambda_{2,m,v}w_{iv} \geq 0 \forall k,k',m,v,$$

and for all higher-order interactions. Third, individuals without associated attributes have a higher or at least an equal log-odds of the probability selecting a lower rather than a higher response category. We achieve this by constraining the intercept parameters of a lower category larger than its adjacent higher category for a given attribute set v, such that

$$\sum_{v=1}^{V}\lambda_{0,m,v}w_{iv} \geq \sum_{v=1}^{V}\lambda_{0,m+1,v}w_{iv} \forall v.$$

Fourth, when an individual possesses the associated attributes, selecting a higher response category on the items should increase or at least be equal to the log-odds of the probability of selecting its adjacent lower response category, such that

$$\sum_{v=1}^{V}\lambda_{z,m,v}w_{iv} \geq \sum_{v=1}^{V}\lambda_{z,m-1,v}w_{iv} \forall v, z \geq 1.$$

where z indicates the level of effect (i.e., 0 for an intercept, 1 for a main effect, 2 for two-way interactions, etc.).

Comparing the RSDM in Eq. 4 to the NRDM in Eq. 3, we can see that the intercept parameter $\lambda_{0,\,i,\,m}$ is decomposed into

Table 1 An example **Q**-matrix

Item	α_1	α_2	α_3
1	1	0	0
2	1	0	0
3	0	1	0
4	0	1	0
5	0	0	1
6	0	0	1
7	1	1	0
8	1	1	0
9	0	1	1
10	0	1	1
11	1	0	1
12	1	0	1

Table 2 The **W**-matrix developed from the **Q**-matrix in Table 1

Item	v_1	v_2	v_3	v_{1*2}	v_{2*3}	v_{1*3}
1	1	0	0	0	0	0
2	1	0	0	0	0	0
3	0	1	0	0	0	0
4	0	1	0	0	0	0
5	0	0	1	0	0	0
6	0	0	1	0	0	0
7	0	0	0	1	0	0
8	0	0	0	1	0	0
9	0	0	0	0	1	0
10	0	0	0	0	1	0
11	0	0	0	0	0	1
12	0	0	0	0	0	1

$\lambda_{0,i}$ (an item general location parameter for the intercept) and $\sum_{v=1}^{V}\lambda_{0,m,v}w_{iv}$ (a threshold parameter for category m common to the items measuring attribute set v). The parameters for the main effects and the interaction effects are also similarly decomposed into those two parts.

Let us discuss the parameter more closely by supposing that each item has five response options (e.g., *never*, *rarely*, *sometimes*, *often*, and *always*). The NRDM requires the estimation of a total of 180 parameters, although some constraints are applied to the model estimation. The breakdown of each type of parameter (i.e., the intercept, main effect, and interaction effect) is shown in Table 3. In contrast, The RSDM requires the estimation of a total of 108 parameters, whose breakdown is shown in Table 4. Comparing Table 4 to Table 3, the major difference is that even-numbered items share the category threshold parameters with the odd-numbered items that measure the same attribute set. For example, Items 1 and 2 share eight parameters: four threshold parameters for the intercept, and four for the main effect. Similarly, Items 7 and 8 share 16 parameters: four threshold parameters for the intercept, eight for the main effect, and four for the interaction effect. This approach reduces the total number of parameters by 40%. In this example, the same attribute set is only measured by two items. Therefore, each set of category threshold parameters is only shared between two items. If more items are added, a set of category threshold parameters can be shared across more than two items, which saves even more parameters as compared to the NRDM.

In practice, the RSDM is most useful when the sample size is limited, because it is a simpler model than the NRDM, while containing all the possible attribute effects. One could also replace its binary core (the LCDM) with other models, such as the compensatory reparametrized unified model (Hartz, 2002; by removing all the interactions) or the DINA model (by keeping only higher-order interactions). However, like earlier IRT models with similar shared category parameter constraints, the RSDM is only applicable when the response categories are identical across items (e.g., Likert scales).

Table 3 Number of parameters in the NRDM: Assigning step parameters to each item

Item	Intercept	Main Effect	Interaction	Total
1	5	5	0	10
2	5	5	0	10
3	5	5	0	10
4	5	5	0	10
5	5	5	0	10
6	5	5	0	10
7	5	10	5	20
8	5	10	5	20
9	5	10	5	20
10	5	10	5	20
11	5	10	5	20
12	5	10	5	20
Total	60	90	30	180

Table 4 Number of parameters in the RSDM: Assigning step parameters to each attribute set

Item	Intercept		Main Effect		Interaction		Total
	G	T	G	T	G	T	
1	1	4	1	4	0	0	12
2	1	*	1	*	0	0	
3	1	4	1	4	0	0	12
4	1	*	1	*	0	0	
5	1	4	1	4	0	0	12
6	1	*	1	*	0	0	
7	1	4	2	8	1	4	24
8	1	*	2	*	1	*	
9	1	4	2	8	1	4	24
10	1	*	2	*	1	*	
11	1	4	2	8	1	4	24
12	1	*	2	*	1	*	
Total	36		54		18		108

"G" represents general item location parameters, and "T" represents category threshold parameters. "*" indicates that the cell shares the same set of parameters with the cell in the row above

Operational study

The purpose of this operational study was to compare the calibrations of the same dataset under the RSDM and the NRDM. We compared parameter estimates, attribute-level and profile-level individual classifications, and the marginal probability of possession of each attribute between the two models.

Data

The dataset used in this study came from the International Critical Thinking and International Communication Attitudes and Beliefs Survey, accessed via the Learning Without Borders project at the University of Florida. In the questionnaire, four items were intended to measure intercultural critical-thinking competence (referred to hereafter as *critical thinking*), and four items were intended to measure intercultural communication competence (referred to hereafter as *communication*). The item stems asked about the degree to which the respondent believed that he or she could perform certain critical thinking or communication tasks, allowing responses of *Strongly disagree*, *Disagree*, *Neutral*, *Agree*, and *Strongly agree*. We obtained 318 individuals' responses to the eight items, and display the frequencies associated with each

response option on each item in Table 5. The sample size and test length in this study were limited and not perfect, because we purposefully utilized this opportunity to test the boundaries of the data used in DCM modeling. As is discussed in Templin and Bradshaw (2013), DCM models trade the precision of locating individuals on one continuum with binary classifications of multiple traits, which could lead to higher reliability estimates. Similar test length can be seen in studies such as Jurich and Bradshaw (2014) in which four or fewer items were used to measure each attribute.

Analysis

We used Markov chain Monte Carlo (MCMC) algorithms to estimate the RSDM and NRDM parameters. The algorithms were implemented in Stan (Carpenter et al., 2016). For each model, we ran two Markov chains with 6,000 iterations in total. The first 1,000 in each chain were used for the burn-in and discarded from the analysis, leaving us with 4,000 draws from the assumed stationary distribution for model parameter inference. To assess the convergence of parameters, we used the multivariate version of the Gelman–Rubin convergence statistic $\hat{R}$ (Brooks & Gelman, 1998; Gelman & Rubin, 1992). A commonly used benchmark is to declare convergence if $\hat{R} < 1.1$ (Junker, Patz, & VanHoudnos, 2016). In both models, all the $\hat{R}$ values were smaller than 1.02, suggesting convergence to the stationary distribution.

For the estimation of the RSDM, we used priors of $N(0, 20)$ for each item parameter and Dirichlet(2) for each attribute profile, similar to Liu and Jiang (2018). These priors are considered less informative and have been recommended in similar DCM studies, such as Chen, Culpepper, Chen, and Douglas (2018) and Jiang and Carter (2018). For the model constraints, we successfully implemented the first constraint as described in the model. For the second constraint, we constrained both the item general location parameter and the category threshold parameter to be positive, such that

$$\lambda_{1,i,k} \geq 0 \forall k,$$

$$\sum_{v=1}^{V} \lambda_{1,m,v} w_{iv} \geq 0 \forall m, v.$$

Note that this is more restrictive than the original second constraint, but we applied this constraint in the present form due to the limitations of the Stan syntax. For the third and fourth constraints, we used pseudo-step parameters in the syntax, such that

$$\lambda_{z,m^*} = \left\{ \begin{array}{c} \lambda_{z,m=1} \forall z, m = 1 \\ \lambda_{z,m=1} + \lambda'_{z,m=2} + \ldots + \lambda'_{z,m=m^*} \forall z, m > 1 \end{array} \right\}, \quad (5)$$

with the constraints $\lambda'_{0,m} \leq 0 \forall m$ and $\lambda'_{z,m} \geq 0 \forall z \geq 1, m$. For estimation of the NRDM, we planned to implement the same set of constraints and succeeded in implementing the first, third, and fourth constraints. However, we failed to apply the second constraint to the NRDM with this dataset. Although the model was able to converge with the second constraint, the parameters were not interpretable and label switching problems were prevalent in the results [e.g., 90% of individuals were classified into the (0, 0) group]. We hypothesized that the estimation problem with the NRDM was associated with the sample size, because the NRDM was recommended for a much larger sample size in Templin et al. (2008). As a result, we fitted the NRDM without the second constraint. Fortunately, the main effects the parameters remained positive. To confirm our results, we obtained parameter estimates of the NRDM in the "CDM" (Robitzsch, Kiefer, George, & Uenlue, 2018) and "GDINA" (Ma & de la Torre, 2018) R packages. The results from the two packages were aligned with our estimation of the NRDM without the second constraint.

For evaluation of the model fit, we used the leave-one-out cross-validation (LOO) method, which is designed for

Table 5 Item data used for the operational study

Item	Dimension	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1. I am able to think critically to interpret global and intercultural issues	Critical Thinking	25 (7.86%)	25 (7.86%)	40 (12.58%)	162 (50.94%)	66 (20.75%)
2. I actively learn about different cultural norms	Critical Thinking	18 (5.66%)	59 (18.55%)	67 (21.07%)	96 (30.19%)	78 (24.53%)
3. I can recognize how different cultures solve problems	Critical Thinking	20 (6.29%)	44 (13.84%)	68 (21.38%)	161 (50.63%)	25 (7.86%)
4. I am able to recognize how members of other cultures make decisions	Critical Thinking	23 (7.23%)	30 (9.43%)	51 (16.04%)	99 (31.13%)	115 (36.16%)
5. I am able to interact effectively with members of other cultures	Communication	12 (3.77%)	50 (15.72%)	118 (37.11%)	121 (38.05%)	17 (5.35%)
6. I can adapt to different cultural environments	Communication	18 (5.66%)	82 (25.79%)	90 (28.30%)	89 (27.99%)	39 (12.26%)
7. I am able to communicate effectively with members of other cultures	Communication	19 (5.97%)	57 (17.92%)	77 (24.21%)	131 (41.19%)	34 (10.69%)
8. I can clearly articulate my point of view to members of other cultures	Communication	16 (5.03%)	70 (22.01%)	122 (38.36%)	86 (27.04%)	24 (7.55%)

evaluating the predictive accuracy of a Bayesian model with simulated parameter values (Gelman, Hwang, & Vehtari, 2014; Vehtari, Gelman, & Gabry, 2017; Yao, Vehtari, Simpson, & Gelman, 2018). As was pointed out in Vehtari et al., the LOO has many advantages over traditional simpler indices, such as the Akaike information criterion (AIC), Bayesian information criterion (BIC), and deviance information criterion (DIC). In this study, we implemented the Pareto-smoothed importance-sampling algorithm (Vehtari et al., 2017) to compute the LOO, and report the expected log predictive density (ELPD) and LOO information criterion (LOOIC) for both the RSDM and the NRDM. The ELPD is computed as

$$EPLD = \sum_{x=1}^{X} \log p(y_x|y_{-x}), \tag{6}$$

where $p(y_x|\, y_{-x}) = \int p(y_x|\, \theta)p(\theta|\, y_{-x})d\theta$ is the leave-one-out predictive density given the dataset without data point x (Vehtari et al., 2017). The LOOIC is $-2 * \text{EPLD}$.

Results

Model fit We estimated 32 item parameters for the RSDM and 64 item parameters for the NRDM. The RSDM displayed a fit with ELPD = – 117.7, LOOIC = 235.5, and the NRDM displayed a fit with ELPD = – 86.3, LOOIC = 172.5. The estimated difference of the expected LOOs between the two models was 31.5, with a standard error of 36.4, indicating that the NRDM did not fit the data significantly better than the RSDM. In practice, applied researchers would most likely consider moving forward with the RSDM, because it is much more parsimonious. For this study, we will continue comparing other aspects between the two models.

Item parameters Table 6 lists the parameter estimates and associated standard errors for the RSDM. Note that the step parameters associated with the same dimension were shared across items. Specifically, Items 1 to 4 shared one set of step parameters; Items 5 to 8 share another set of step parameters. Table 7 lists the parameter estimates and associated standard errors for the NRDM. In both tables, we report pseudo-step parameters, indicated as λ'. One could simply add the pseudo-step parameters in order to obtain the real parameters. For example, the $\lambda_{0,\ m=3}$ in the RSDM equals to $\lambda_{0,m=1} + \lambda'_{0,m=2} + \lambda'_{0,m=3}$, and the $\lambda_{0,i,m=3} = \lambda_{0,i,m=1} + \lambda'_{0,i,m=2} + \lambda'_{0,i,m=3}$. A notable difference between the two models is that we were able to obtain a general intercept parameter $\lambda_{0,\ i}$ and a general main effect parameter $\lambda_{1,\ i}$ for each item under the RSDM. A larger $\lambda_{0,\ i}$ indicates that, at large, a person without associated attributes is more likely to select a higher response category on an item. For example, Item 2 is easier to endorse than Item 1 when an individual does not possess the attribute critical

Table 6 RSDM: Item parameter estimates and standard errors for the operational study

	$\lambda_{0,\ i}$	$\lambda_{0,\ m=1}$	$\lambda'_{0,m=2}$	$\lambda'_{0,m=3}$	$\lambda'_{0,m=4}$	$\lambda_{1,\ i}$	$\lambda_{1,\ m=1}$	$\lambda'_{1,m=2}$	$\lambda'_{1,m=3}$	$\lambda'_{1,m=4}$
Item 1	6.660 (0.415)	– 6.767 (0.416)	– 0.759 (0.006)	– 2.186 (0.025)	– 11.222 (0.301)	16.180 (0.263)	16.589 (0.278)	1.897 (0.009)	3.239 (0.026)	10.630 (0.301)
Item 2	7.190 (0.417)	*	*	*	*	16.117 (0.277)	*	*	*	*
Item 3	7.033 (0.418)	*	*	*	*	15.936 (0.262)	*	*	*	*
Item 4	6.819 (0.413)	*	*	*	*	15.715 (0.260)	*	*	*	*
Item 5	7.771 (0.296)	– 6.692 (0.298)	– 0.728 (0.005)	– 1.712 (0.012)	– 2.810 (0.041)	15.885 (0.266)	16.197 (0.264)	1.710 (0.006)	1.898 (0.012)	1.506 (0.041)
Item 6	7.244 (0.297)	*	*	*	*	16.015 (0.276)	*	*	*	*
Item 7	7.174 (0.298)	*	*	*	*	15.548 (0.258)	*	*	*	*
Item 8	7.417 (0.297)	*	*	*	*	16.131 (0.274)	*	*	*	*

"*" indicates that the values in this cell are equivalent to those in the cell in the row above

Table 7 NRDM: Item parameter estimates and standard errors for the operational study

	$\lambda_{0,\,i,\,m=1}$	$\lambda'_{0,i,m=2}$	$\lambda'_{0,i,m=3}$	$\lambda'_{0,i,m=4}$	$\lambda_{1,\,i,\,m=1}$	$\lambda'_{1,i,m=2}$	$\lambda'_{1,i,m=3}$	$\lambda'_{1,i,m=4}$
Item 1	− 0.062 (0.006)	− 0.540 (0.008)	− 1.634 (0.075)	− 12.724 (0.357)	16.422 (0.265)	4.000 (0.029)	3.452 (0.075)	11.852 (0.357)
Item 2	0.815 (0.007)	− 1.621 (0.013)	− 13.922 (0.293)	− 6.413 (0.402)	17.623 (0.267)	2.787 (0.013)	14.422 (0.294)	6.204 (0.402)
Item 3	0.115 (0.007)	− 0.466 (0.010)	− 0.395 (0.018)	− 12.597 (0.261)	18.281 (0.262)	1.393 (0.013)	1.437 (0.020)	10.779 (0.261)
Item 4	0.148 (0.007)	− 0.456 (0.007)	− 14.694 (0.259)	− 5.611 (0.355)	16.490 (0.262)	7.171 (0.439)	15.783 (0.259)	5.767 (0.355)
Item 5	0.823 (0.008)	− 0.462 (0.011)	− 0.633 (0.016)	− 12.865 (0.347)	17.654 (0.253)	1.945 (0.014)	0.738 (0.017)	10.952 (0.346)
Item 6	0.567 (0.007)	− 0.703 (0.010)	− 14.925 (0.268)	− 6.452 (0.350)	18.391 (0.248)	1.087 (0.014)	15.113 (0.268)	5.621 (0.351)
Item 7	0.620 (0.006)	− 1.167 (0.011)	− 13.173 (0.289)	− 6.081 (0.346)	17.491 (0.262)	2.256 (0.014)	13.867 (0.289)	4.724 (0.346)
Item 8	0.717 (0.008)	− 0.732 (0.009)	− 9.752 (0.274)	− 6.773 (0.272)	17.353 (0.254)	1.766 (0.011)	9.536 (0.273)	8.117 (0.272)

thinking. A larger $\lambda_{1,\,i}$ indicates that possessing the required attributes for an item has a larger effect on selecting a higher response category on that item. In the NRDM, instead, we were only able to obtain parameters for each response category.

Category response probabilities It may be difficult to eyeball the probabilities of responding to each category from Tables 6 and 7 and tell the differences between the two models. Therefore, we plotted the category response curves for all the eight items in Fig. 1. In each plot, the response categories are on the *x*-axis, and the probability of selecting a response option is on the *y*-axis. The results from the RSDM are displayed with dashed lines, and those from the NRDM are displayed with solid lines. Each line with circles represents the probability of selecting a response option for an attribute possessor. Each line with triangles represents the probability of selecting a response option for an attribute nonpossessor. The plots from the RSDM and NRDM look similar overall, meaning that they would produce similar individual classification results.

Profile and attribute possession agreement Table 8 displays estimates of the attribute prevalence of the two models. Each value represents the probability of an individual having an attribute profile at large (Templin & Hoffman, 2013). From Table 8, we can see that the probabilities of having each attribute profile were very similar between the two models. Therefore, we moved on to investigate the actual individual classification agreement between the two models. Table 9 shows a cross-classification count of individuals in each profile assigned by the two models. In all, 93.4% of individuals (i.e., 297 out of 318) were assigned the same attribute profile by both models, with a Cohen's kappa of .87. Table 10 shows the cross-classification agreement between the two models at the attribute level. The agreement was 98.1% (with a Cohen's kappa of .94) and 95.3% (with a Cohen's kappa of .87) for critical thinking and communication, respectively. According to Landis and Koch (1977), Cohen's kappa values higher than .81 suggest almost perfect agreement. For those 21 disagreements, we also examined the individual response patterns and found that most of the classifications provided by the RSDM probably make more sense intuitively. For example, on Items 5–8, which all measure communication, two individuals' response patterns were (2,2,2,4) and (2,2,4,2), respectively. The NRDM assigned both individuals with possession of communication, and the RSDM assigned both individuals a nonpossession status. Considering that this is a five-option scale, the RSDM results are probably intuitively easier to understand.

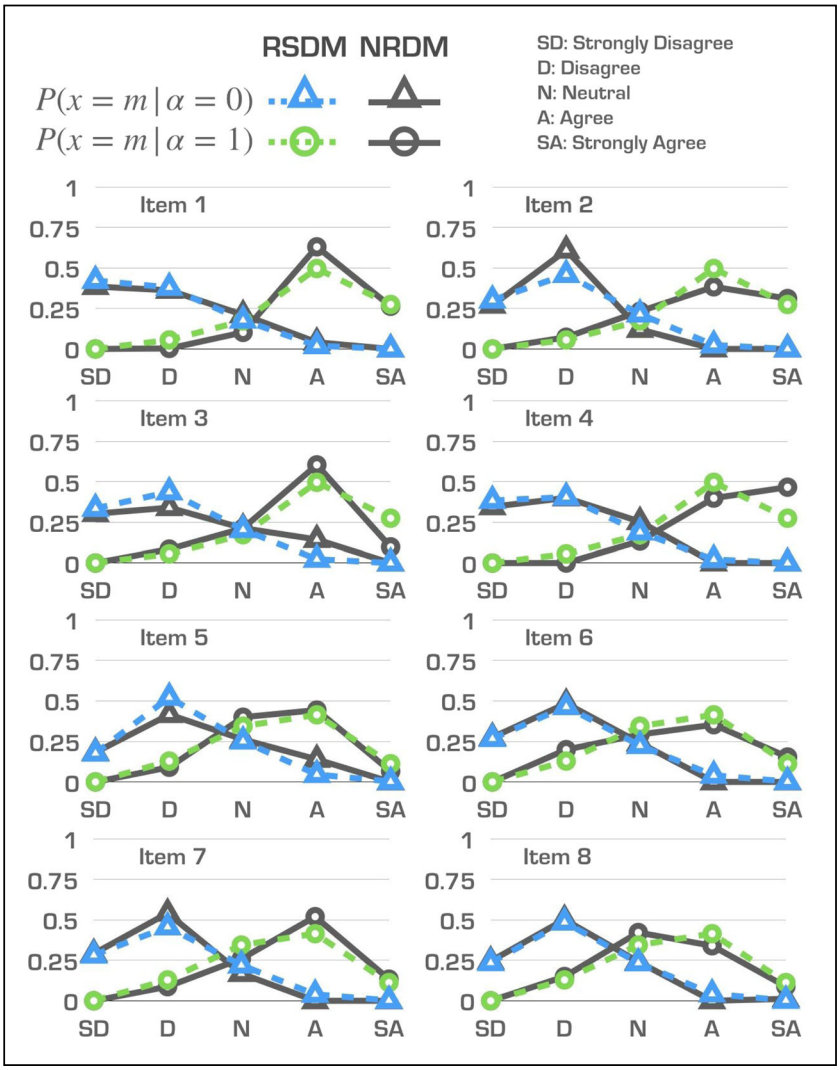

Fig. 1 Category response curves under the RSDM and NRDM

We also examined individuals' marginal probabilities of possession for each attribute between the two models. Those probabilities, known as *continuous scores* (Liu, Qian, Luo, & Woo, 2017), are displayed in Fig. 2. We computed the root-mean-square deviation between the continuous scores from two models and found that the differences were very small: .098 for critical thinking, and .118 for communication.

Table 8 Attribute prevalence estimates under the RSDM and NRDM for the operational study

Profile	RSDM	NRDM
(0, 0)	.137	.135
(1, 0)	.076	.072
(0, 1)	.052	.077
(1, 1)	.734	.717

Simulation study

The preceding operational study informed us that the RSDM performed similarly to the NRDM in terms of model fit, category response probabilities, attribute and profile classifications,

Table 9 Profile possession agreement between the RSDM and NRDM for the operational study

NRDM	RSDM			
	(0, 0)	(1, 0)	(0, 1)	(1, 1)
(0, 0)	42 (13.2%)	0	1 (0.3%)	0
(1, 0)	0	28 (8.8%)	0	5 (1.6%)
(0, 1)	1 (0.3%)	0	21 (6.6%)	6 (1.9%)
(1, 1)	0	8 (2.5%)	0	206 (64.8%)

Note: The total number of profile agreement between the two models was 297 (93.4%)

Table 10 Attribute possession agreement between the RSDM and NRDM for the operational study

NRDM		RSDM
	$\alpha_1 = 0$	$\alpha_1 = 1$
$\alpha_1 = 0$	65 (20.4%)	6 (1.9%)
$\alpha_1 = 1$	0	247 (77.7%)
	$\alpha_2 = 0$	$\alpha_2 = 1$
$\alpha_2 = 0$	70 (22.0%)	6 (1.9%)
$\alpha_2 = 1$	9 (2.8%)	233 (73.3%)

The total agreement scores between the two models for α_1 and α_2 were 312 (98.1%) and 303 (95.3%), respectively

and continuous scores. However, the RSDM is much more parsimonious than the NRDM and offers additional information about the overall item properties. The purpose of the simulation study was to investigate whether the RSDM can produce unbiased parameter estimates and correctly classify individuals under the applied study condition.

Method

We generated 100 datasets in R (R Core Team, 2018) using the parameter values obtained from the applied study. Specifically, the data-generating item parameters are listed in Table 6. A total of 500 individuals were generated for each dataset from a multinomial distribution with probabilities of (.137, .076, .052, .734) for the four attribute profiles (0, 0), (1, 0), (0, 1), and (1, 1), respectively. This distribution was obtained from the attribute prevalence in the applied study. The item and person parameters were submitted to the RSDM in order to compute the probability of scoring each response category on each item for each individual. We then drew a random number from the multinomial distribution of the response category probabilities for each item and individual to serve as the individual's item response. Then we examined the average proportion of individuals in each response category across the 100 datasets and ensured that our simulated datasets mirror the applied dataset.

We then fit the RSDM to each dataset using the same Stan code and MCMC specifications as for the applied study. To assess parameter recovery, we computed the bias and root-mean square error (RMSE) for each item parameter and attribute prevalence estimate. The formulas for computing bias and RMSE are presented as follows, where x is a parameter, $e(x)$ is the true (simulated) value of parameter x, and $\hat{e}_r(x)$ is the rth replicate estimate of parameter x among $R = 100$ replications:

$$Bias(x) = \frac{\sum_{r=1}^{R}\left[\hat{e}_r(x) - e(x)\right]}{R}, \tag{7}$$

$$RMSE(x) = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left[\hat{e}_r(x) - e(x)\right]^2}. \tag{8}$$

To assess classification accuracy, we computed the percentage of agreement between the true and estimated classifications on each attribute in each dataset.

Results

Table 11 displays the bias and RMSE of each item parameter. The majority of the parameters displayed bias around 0 and RMSE below .9. However, we found larger bias and RMSE on the intercept and main effect parameters of the $m = 3$ and $m = 4$ categories of the first attribute and of the $m = 4$ category of the second attribute. We hypothesized that two things might relate to the larger bias and RMSE of the step parameters at extreme locations. First, the way we specified the pseudo-step parameters could have added uncertainty to the estimation of parameters at a higher threshold. For example, $\lambda_{1,m=4,v} = \lambda_{1,m=1,v} + \lambda'_{1,m=2,v} + \lambda'_{1,m=3,v} + \lambda'_{1,m=4,v} \forall v$. Unlike scoring $m = 1$, where the logit is a direct function of $\lambda_{1,m=1,v}$, the logit of scoring $m = 4$ is a function of the four components listed above. We would expect that estimation of

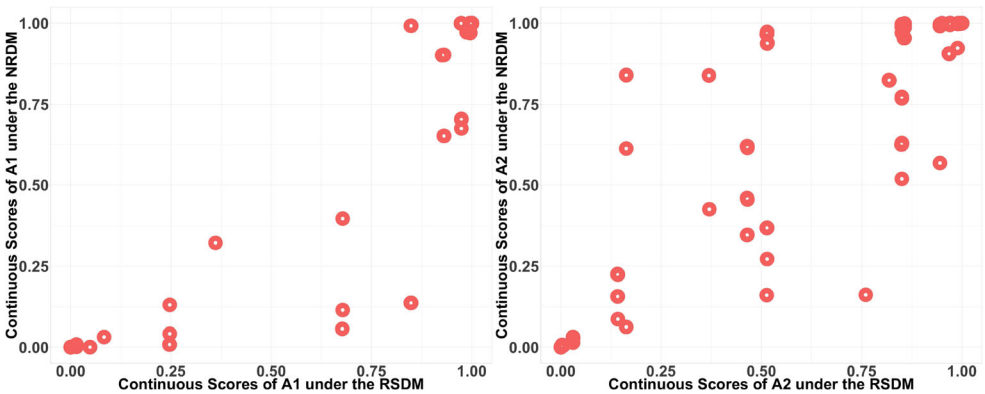

Fig. 2 Continuous scores under the RSDM and NRDM

Table 11 Bias and RMSE of estimated item parameters for the RSDM in the simulation study

	Item-Specific General Intercept Parameters							
	$\lambda_{0,1}$	$\lambda_{0,2}$	$\lambda_{0,3}$	$\lambda_{0,4}$	$\lambda_{0,5}$	$\lambda_{0,6}$	$\lambda_{0,7}$	$\lambda_{0,8}$
Bias	– 0.040	– 0.072	– 0.062	– 0.076	0.098	0.038	0.056	0.029
RMSE	0.632	0.778	0.726	0.772	0.541	0.360	0.471	0.221
	Item-Specific General Main Effect Parameters							
	$\lambda_{1,1}$	$\lambda_{1,2}$	$\lambda_{1,3}$	$\lambda_{1,4}$	$\lambda_{1,5}$	$\lambda_{1,6}$	$\lambda_{1,7}$	$\lambda_{1,8}$
Bias	0.024	– 0.115	– 0.077	0.086	0.065	– 0.052	0.077	– 0.114
RMSE	0.179	0.705	0.777	0.716	0.611	0.430	0.681	0.807
	Category Threshold Intercept Parameters							
	$\lambda_{0,m=1,v_1}$	$\lambda'_{0,m=2,v_1}$	$\lambda'_{0,m=3,v_1}$	$\lambda'_{0,m=4,v_1}$	$\lambda_{0,m=1,v_2}$	$\lambda'_{0,m=2,v_2}$	$\lambda'_{0,m=3,v_2}$	$\lambda'_{0,m=4,v_2}$
Bias	– 0.041	– 0.004	0.553	0.834	0.195	0.014	0.028	0.837
RMSE	0.775	0.175	1.144	1.129	0.920	0.151	0.334	1.603
	Category Threshold Main Effect Parameters							
	$\lambda_{1,m=1,v_1}$	$\lambda'_{1,m=2,v_1}$	$\lambda'_{1,m=3,v_1}$	$\lambda'_{1,m=4,v_1}$	$\lambda_{1,m=1,v_2}$	$\lambda'_{1,m=2,v_2}$	$\lambda'_{1,m=3,v_2}$	$\lambda'_{1,m=4,v_2}$
Bias	– 0.015	0.024	0.551	0.836	– 0.168	0.021	0.024	0.841
RMSE	0.442	0.231	1.139	1.129	0.412	0.194	0.351	1.602

parameters at a higher threshold might have less precision. Second, estimation of the step parameters of extreme categories might also be less precise, similar to the "outward bias" issue documented with estimations of unidimensional rating scale models (Huggins-Manley, Algina, & Zhou, 2018; Samejima, 1973). Both of these studies showed that "outward bias" directly relates to the free estimations of item discrimination parameters associated with the unidimensional models. In the RSDM, the main effect parameters, which are conceptually similar to discrimination parameters, are freely estimated. Therefore, we hypothesized that this might also be a contributing factor to the larger bias and RMSE for the step parameters of extreme categories.

Table 12 shows the bias and RMSE of estimated attribute prevalence. The bias and RMSE were both .01 or smaller. This means that the estimated distribution of the individual attribute profiles was almost identical to the true distribution. Table 13 presents the classification accuracy results at the attribute and profile levels. The means for the rows are all around .99, meaning that the RSDM was able to accurately recover individuals' attributes under the conditions in this study.

Table 12 Bias and RMSE of estimated attribute prevalence for the RSDM in the simulation study

	(0, 0)	(1, 0)	(0, 1)	(1, 1)
Bias	– .002	.001	.008	– .007
RMSE	.006	.009	.010	.010

Discussion

DCMs can be very useful in psychological rating scales. For example, responses on a Myers–Briggs type indicator (MBTI; Myers & Myers, 1980) questionnaire could be scored using a DCM with four dichotomous attributes through which individuals are assigned with one of the 16 possible attribute profiles (e.g., ENTJ: extraversion, intuition, thinking, and judging) through a retrofitting process (Liu, Huggins-Manley, & Bulut, 2018). However, traditional DCMs for polytomous item responses require a very large sample size that is hardly attainable for either research or assessment practice in daily life. The RSDM proposed in this study requires a much smaller sample size, making the application of DCMs to rating scales in commonplace tests possible.

The RSDM was developed as a constrained version of the NRDM, built on the practices of sharing category threshold parameters (Andrich, 1978; Muraki, 1990) and representing dimensions with latent classes in DCMs (Rupp & Templin, 2008; Templin & Hoffman, 2013). The simulation study demonstrated that the RSDM was able to recover parameters and

Table 13 Descriptive statistics of attribute and profile classification accuracy for the RSDM in the simulation study

	Min	Mean	Max	*SD*
Attribute 1	.983	.996	1.000	.003
Attribute 2	.973	.990	1.000	.005
Profile	.970	.987	1.000	.006

Table 14 Relationship between current DCMs for polytomous item data

		Polytomous Features		
Type of item data		Both nominal and ordinal	Ordinal only	
Model extension method		Divide-by-total approach	Cumulative approach	Divide-by-total approach
Corresponding reduced DCM for binary item data	The LCDM	NRDM GDM SG-DINA	P-LCDM GPDM	ORDM MORDM RSDM
	The DINA model (a special case of the LCDM)	PC-DINA	DINA-GD	N/A

correctly classify individuals under an applied condition. Although we witnessed larger bias and RMSE for category threshold parameters at extreme locations, the estimated attribute profile distribution and attribute classifications were almost ideal. The applied study showed that the RSDM produced very similar item category response probabilities and individual scores with the NRDM. In addition to its 50% smaller model size, the RSDM also offered general item information through an overall intercept and an overall main effect for each item, which is not available in traditional DCMs for polytomous items. Practitioners may find this information useful for item revision, reporting, and score interpretation.

Through the positive results of both the simulation and applied studies, we believe that the RSDM holds promise for scoring rating scale data in psychological tests. A major limitation of the present study is that the findings are couched within the applied study condition; therefore, we encourage future research to investigate conditions that may impact the performance of the RSDM. The following topics may be of particular interest to the applicability of the RSDM. First, investigating the parameter recovery of the RSDM under various sample size conditions. We fitted the RSDM to an operational dataset with 500 individuals and obtained acceptable parameter recovery and excellent classification accuracy. It would be helpful to know the performance of the RSDM with other smaller and bigger sample sizes. Second, examining the impact of the **Q**-matrix complexity on the RSDM. We fitted the RSDM to a two-dimensional single-structured dataset. However, it is not uncommon to have a complex-structured dataset especially under the educational assessment setting. We hypothesize that if the number of dimensions increases while other things remain equal, the performance of the RSDM is likely to be even closer to the NRDM. This is because increasing the number of dimensions essentially increases the number of v. If each item has its own v, the RSDM will be equivalent to the NRDM in which the category threshold parameters of each item are freely estimated. Third, examining the identifiability of the RSDM. Comparing to scoring dichotomous item responses, scoring polytomous item responses requires an increase in the number of parameters in the models. Although the RSDM reduces the total number of parameters, it would be helpful to investigate the necessary conditions for identifiability, similar to Gu and Xu (2018) and Xu and Zhang (2016). Fourth, incorporating attribute structure (i.e., the possession sequence of attributes) into the RSDM could further simplify the structural part of the model (see Liu, 2018; Liu & Huggins-Manley, 2016; Liu, Huggins-Manley, & Bradshaw, 2017). Fifth, exploring the impact of the agreement/disagreement of the category threshold distances among items that measure the same set of attributes could assist practitioners with future applications of the RSDM. As we described in the Model Development section , the RSDM is built on the concept that the relative distances between steps are held constant across all items within the same dimension. Although we acknowledge that the relative distances cannot be perfectly identical, it is helpful to explore the robustness of the model against the distance discrepancy.

This final recommendation, for future research on the effect of constant category threshold distances, a unique feature of the RSDM, brings up a final point of discussion for this study. We will briefly compare ten DCMs that are currently capable of handling polytomous item responses, to point out where the RSDM sits in the big picture. These models are (1) the NRDM, (2) the PC-DINA, (3) the general diagnostic model (GDM; von Davier, 2005), (4) the polytomous log-linear cognitive diagnosis model (P-LCDM; Hansen, 2013), (5) the sequential G-DINA model (SG-DINA; Ma & de la Torre, 2016), (6) the DINA model for graded data (DINA-GD; Tu et al., 2017), (7) the general polytomous diagnosis model (GPDM; Chen & de la Torre, 2018), (8) the ordinal response diagnostic model (ORDM; Liu & Jiang, 2018), (9) the modified ordinal response diagnostic model (MORDM; Liu & Jiang, 2018), and (10) the RSDM. We categorize the ten models in Table 14, on the basis of their reduced binary models and polytomous extension features. Eight of the models can be reduced to the LCDM if there are only two response options, whereas PC-DINA and DINA-GD can be reduced to the DINA model. Since the DINA model is a special case of the LCDM, all eight of the LCDM-based models can be easily reduced to DINA-based models. Four models (i.e., the NRDM, the GDM, the SG-DINA, and the PC-DINA) can handle both nominal and ordinal item data, whereas the other six models, including the RSDM, incorporate the ordinal feature of item responses into the models, and thus can only handle ordinal item data. Among the six ordinal

DCMs, three of them (i.e., the P-LCDM, the GPDM, and the DINA-GD) utilize the cumulative approach in which we model the probability of selecting a particular response option and all higher ones, given respondents' attribute profile. The other three models (i.e., the ORDM, the MORDM, and the RSDM) utilize the divide-by-total approach, in which we model the ratio of selecting a particular response option divided by the sum of these values across all response options.

To conclude, although polytomous DCMs have been developed and applied to psychological tests (e.g., Templin & Henson, 2006), the majority of model development and applications of DCMs are thriving in educational testing scenarios. We hope that the RSDM can be useful for classifying individuals on psychological rating scales such as personality tests and diagnostics of behavioral/mental disorders.

Appendix: Stan code for the RSDM

```
data{
int Np;
int Ni;
int Nc;
int Ns;
int Y[Np, Ni];
}
parameters{
simplex[Nc] Vc;
real<lower=0> l1_1 ;
real<lower=0> l2_1 ;
real<lower=0> l3_1 ;
real<lower=0> l4_1 ;
real<lower=0> l5_1 ;
real<lower=0> l6_1 ;
real<lower=0> l7_1 ;
real<lower=0> l8_1 ;
real l1_0 ;
real l2_0 ;
real l3_0 ;
real l4_0 ;
real l5_0 ;
real l6_0 ;
real l7_0 ;
real l8_0 ;
real<lower=0> step1D1I;
real<lower=0> step2D1I;
real<lower=0> step3D1I;
real<lower=0> step4D1I;
real<lower=0> step1D2I;
real<lower=0> step2D2I;
real<lower=0> step3D2I;
real<lower=0> step4D2I;
real<lower=0> step1D1M;
real<lower=0> step2D1M;
real<lower=0> step3D1M;
real<lower=0> step4D1M;
real<lower=0> step1D2M;
real<lower=0> step2D2M;
real<lower=0> step3D2M;
real<lower=0> step4D2M;
}
transformed parameters{
vector[Ns] PImat[Ni, Nc];

PImat[1,1,1]=0;
PImat[2,1,1]=0;
PImat[3,1,1]=0;
PImat[4,1,1]=0;
PImat[5,1,1]=0;
PImat[6,1,1]=0;
PImat[7,1,1]=0;
PImat[8,1,1]=0;

PImat[1,2,1]=0;
PImat[2,2,1]=0;
PImat[3,2,1]=0;
PImat[4,2,1]=0;
PImat[5,2,1]=0;
PImat[6,2,1]=0;
PImat[7,2,1]=0;
PImat[8,2,1]=0;

PImat[1,3,1]=0;
PImat[2,3,1]=0;
PImat[3,3,1]=0;
PImat[4,3,1]=0;
PImat[5,3,1]=0;
PImat[6,3,1]=0;
PImat[7,3,1]=0;
PImat[8,3,1]=0;

PImat[1,4,1]=0;
PImat[2,4,1]=0;
PImat[3,4,1]=0;
PImat[4,4,1]=0;
PImat[5,4,1]=0;
PImat[6,4,1]=0;
PImat[7,4,1]=0;
PImat[8,4,1]=0;

PImat[1,1,2]=(l1_0)-step1D1I;
PImat[2,1,2]=(l2_0)-step1D1I;
PImat[3,1,2]=(l3_0)-step1D1I;
PImat[4,1,2]=(l4_0)-step1D1I;
PImat[5,1,2]=(l5_0)-step1D2I;
PImat[6,1,2]=(l6_0)-step1D2I;
PImat[7,1,2]=(l7_0)-step1D2I;
PImat[8,1,2]=(l8_0)-step1D2I;

PImat[1,1,3]=(l1_0)-step1D1I-step2D1I;
PImat[2,1,3]=(l2_0)-step1D1I-step2D1I;
```

```
PImat[3,1,3]=(l3_0)-step1D1I-step2D1I;
PImat[4,1,3]=(l4_0)-step1D1I-step2D1I;
PImat[5,1,3]=(l5_0)-step1D2I-step2D2I;
PImat[6,1,3]=(l6_0)-step1D2I-step2D2I;
PImat[7,1,3]=(l7_0)-step1D2I-step2D2I;
PImat[8,1,3]=(l8_0)-step1D2I-step2D2I;

PImat[1,1,4]=(l1_0)-step1D1I-step2D1I-step3D1I;
PImat[2,1,4]=(l2_0)-step1D1I-step2D1I-step3D1I;
PImat[3,1,4]=(l3_0)-step1D1I-step2D1I-step3D1I;
PImat[4,1,4]=(l4_0)-step1D1I-step2D1I-step3D1I;
PImat[5,1,4]=(l5_0)-step1D2I-step2D2I-step3D2I;
PImat[6,1,4]=(l6_0)-step1D2I-step2D2I-step3D2I;
PImat[7,1,4]=(l7_0)-step1D2I-step2D2I-step3D2I;
PImat[8,1,4]=(l8_0)-step1D2I-step2D2I-step3D2I;

PImat[1,1,5]=(l1_0)-step1D1I-step2D1I-step3D1I-step4D1I;
PImat[2,1,5]=(l2_0)-step1D1I-step2D1I-step3D1I-step4D1I;
PImat[3,1,5]=(l3_0)-step1D1I-step2D1I-step3D1I-step4D1I;
PImat[4,1,5]=(l4_0)-step1D1I-step2D1I-step3D1I-step4D1I;
PImat[5,1,5]=(l5_0)-step1D2I-step2D2I-step3D2I-step4D2I;
PImat[6,1,5]=(l6_0)-step1D2I-step2D2I-step3D2I-step4D2I;
PImat[7,1,5]=(l7_0)-step1D2I-step2D2I-step3D2I-step4D2I;
PImat[8,1,5]=(l8_0)-step1D2I-step2D2I-step3D2I-step4D2I;

PImat[1,2,2]=(l1_0+l1_1)-step1D1I+step1D1M;
PImat[2,2,2]=(l2_0+l2_1)-step1D1I+step1D1M;
PImat[3,2,2]=(l3_0+l3_1)-step1D1I+step1D1M;
PImat[4,2,2]=(l4_0+l4_1)-step1D1I+step1D1M;
PImat[5,2,2]=(l5_0)-step1D2I;
PImat[6,2,2]=(l6_0)-step1D2I;
PImat[7,2,2]=(l7_0)-step1D2I;
PImat[8,2,2]=(l8_0)-step1D2I;

PImat[1,2,3]=(l1_0+l1_1)-step1D1I-step2D1I+step1D1M+step2D1M;
PImat[2,2,3]=(l2_0+l2_1)-step1D1I-step2D1I+step1D1M+step2D1M;
PImat[3,2,3]=(l3_0+l3_1)-step1D1I-step2D1I+step1D1M+step2D1M;
PImat[4,2,3]=(l4_0+l4_1)-step1D1I-step2D1I+step1D1M+step2D1M;
PImat[5,2,3]=(l5_0)-step1D2I-step2D2I;
PImat[6,2,3]=(l6_0)-step1D2I-step2D2I;
PImat[7,2,3]=(l7_0)-step1D2I-step2D2I;
PImat[8,2,3]=(l8_0)-step1D2I-step2D2I;

PImat[1,2,4]=(l1_0+l1_1)-step1D1I-step2D1I-
step3D1I+step1D1M+step2D1M+step3D1M;
PImat[2,2,4]=(l2_0+l2_1)-step1D1I-step2D1I-
step3D1I+step1D1M+step2D1M+step3D1M;
```

```
PImat[3,2,4]=(l3_0+l3_1)-step1D1I-step2D1I-
step3D1I+step1D1M+step2D1M+step3D1M;
PImat[4,2,4]=(l4_0+l4_1)-step1D1I-step2D1I-
step3D1I+step1D1M+step2D1M+step3D1M;
PImat[5,2,4]=(l5_0)-step1D2I-step2D2I-step3D2I;
PImat[6,2,4]=(l6_0)-step1D2I-step2D2I-step3D2I;
PImat[7,2,4]=(l7_0)-step1D2I-step2D2I-step3D2I;
PImat[8,2,4]=(l8_0)-step1D2I-step2D2I-step3D2I;

PImat[1,2,5]=(l1_0+l1_1)-step1D1I-step2D1I-
step3D1I+step1D1M+step2D1M+step3D1M-step4D1I+step4D1M;
PImat[2,2,5]=(l2_0+l2_1)-step1D1I-step2D1I-
step3D1I+step1D1M+step2D1M+step3D1M-step4D1I+step4D1M;
PImat[3,2,5]=(l3_0+l3_1)-step1D1I-step2D1I-
step3D1I+step1D1M+step2D1M+step3D1M-step4D1I+step4D1M;
PImat[4,2,5]=(l4_0+l4_1)-step1D1I-step2D1I-
step3D1I+step1D1M+step2D1M+step3D1M-step4D1I+step4D1M;
PImat[5,2,5]=(l5_0)-step1D2I-step2D2I-step3D2I-step4D2I;
PImat[6,2,5]=(l6_0)-step1D2I-step2D2I-step3D2I-step4D2I;
PImat[7,2,5]=(l7_0)-step1D2I-step2D2I-step3D2I-step4D2I;
PImat[8,2,5]=(l8_0)-step1D2I-step2D2I-step3D2I-step4D2I;

PImat[1,3,2]=(l1_0)-step1D1I;
PImat[2,3,2]=(l2_0)-step1D1I;
PImat[3,3,2]=(l3_0)-step1D1I;
PImat[4,3,2]=(l4_0)-step1D1I;
PImat[5,3,2]=(l5_0+l5_1)-step1D2I+step1D2M;
PImat[6,3,2]=(l6_0+l6_1)-step1D2I+step1D2M;
PImat[7,3,2]=(l7_0+l7_1)-step1D2I+step1D2M;
PImat[8,3,2]=(l8_0+l8_1)-step1D2I+step1D2M;

PImat[1,3,3]=(l1_0)-step1D1I-step2D1I;
PImat[2,3,3]=(l2_0)-step1D1I-step2D1I;
PImat[3,3,3]=(l3_0)-step1D1I-step2D1I;
PImat[4,3,3]=(l4_0)-step1D1I-step2D1I;
PImat[5,3,3]=(l5_0+l5_1)-step1D2I+step1D2M-step2D2I+step2D2M;
PImat[6,3,3]=(l6_0+l6_1)-step1D2I+step1D2M-step2D2I+step2D2M;
PImat[7,3,3]=(l7_0+l7_1)-step1D2I+step1D2M-step2D2I+step2D2M;
PImat[8,3,3]=(l8_0+l8_1)-step1D2I+step1D2M-step2D2I+step2D2M;

PImat[1,3,4]=(l1_0)-step1D1I-step2D1I-step3D1I;
PImat[2,3,4]=(l2_0)-step1D1I-step2D1I-step3D1I;
PImat[3,3,4]=(l3_0)-step1D1I-step2D1I-step3D1I;
PImat[4,3,4]=(l4_0)-step1D1I-step2D1I-step3D1I;
PImat[5,3,4]=(l5_0+l5_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M;
```

```
PImat[6,3,4]=(l6_0+l6_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M;
PImat[7,3,4]=(l7_0+l7_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M;
PImat[8,3,4]=(l8_0+l8_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M;

PImat[1,3,5]=(l1_0)-step1D1I-step2D1I-step3D1I-step4D1I;
PImat[2,3,5]=(l2_0)-step1D1I-step2D1I-step3D1I-step4D1I;
PImat[3,3,5]=(l3_0)-step1D1I-step2D1I-step3D1I-step4D1I;
PImat[4,3,5]=(l4_0)-step1D1I-step2D1I-step3D1I-step4D1I;
PImat[5,3,5]=(l5_0+l5_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M-step4D2I+step4D2M;
PImat[6,3,5]=(l6_0+l6_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M-step4D2I+step4D2M;
PImat[7,3,5]=(l7_0+l7_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M-step4D2I+step4D2M;
PImat[8,3,5]=(l8_0+l8_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M-step4D2I+step4D2M;

PImat[1,4,2]=(l1_0+l1_1)-step1D1I+step1D1M;
PImat[2,4,2]=(l2_0+l2_1)-step1D1I+step1D1M;
PImat[3,4,2]=(l3_0+l3_1)-step1D1I+step1D1M;
PImat[4,4,2]=(l4_0+l4_1)-step1D1I+step1D1M;
PImat[5,4,2]=(l5_0+l5_1)-step1D2I+step1D2M;
PImat[6,4,2]=(l6_0+l6_1)-step1D2I+step1D2M;
PImat[7,4,2]=(l7_0+l7_1)-step1D2I+step1D2M;
PImat[8,4,2]=(l8_0+l8_1)-step1D2I+step1D2M;

PImat[1,4,3]=(l1_0+l1_1)-step1D1I+step1D1M-step2D1I+step2D1M;
PImat[2,4,3]=(l2_0+l2_1)-step1D1I+step1D1M-step2D1I+step2D1M;
PImat[3,4,3]=(l3_0+l3_1)-step1D1I+step1D1M-step2D1I+step2D1M;
PImat[4,4,3]=(l4_0+l4_1)-step1D1I+step1D1M-step2D1I+step2D1M;
PImat[5,4,3]=(l5_0+l5_1)-step1D2I+step1D2M-step2D2I+step2D2M;
PImat[6,4,3]=(l6_0+l6_1)-step1D2I+step1D2M-step2D2I+step2D2M;
PImat[7,4,3]=(l7_0+l7_1)-step1D2I+step1D2M-step2D2I+step2D2M;
PImat[8,4,3]=(l8_0+l8_1)-step1D2I+step1D2M-step2D2I+step2D2M;

PImat[1,4,4]=(l1_0+l1_1)-step1D1I+step1D1M-step2D1I+step2D1M-
step3D1I+step3D1M;
PImat[2,4,4]=(l2_0+l2_1)-step1D1I+step1D1M-step2D1I+step2D1M-
step3D1I+step3D1M;
PImat[3,4,4]=(l3_0+l3_1)-step1D1I+step1D1M-step2D1I+step2D1M-
step3D1I+step3D1M;
PImat[4,4,4]=(l4_0+l4_1)-step1D1I+step1D1M-step2D1I+step2D1M-
step3D1I+step3D1M;
```

```
PImat[5,4,4]=(l5_0+l5_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M;
PImat[6,4,4]=(l6_0+l6_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M;
PImat[7,4,4]=(l7_0+l7_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M;
PImat[8,4,4]=(l8_0+l8_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M;

PImat[1,4,5]=(l1_0+l1_1)-step1D1I+step1D1M-step2D1I+step2D1M-
step3D1I+step3D1M-step4D1I+step4D1M;
PImat[2,4,5]=(l2_0+l2_1)-step1D1I+step1D1M-step2D1I+step2D1M-
step3D1I+step3D1M-step4D1I+step4D1M;
PImat[3,4,5]=(l3_0+l3_1)-step1D1I+step1D1M-step2D1I+step2D1M-
step3D1I+step3D1M-step4D1I+step4D1M;
PImat[4,4,5]=(l4_0+l4_1)-step1D1I+step1D1M-step2D1I+step2D1M-
step3D1I+step3D1M-step4D1I+step4D1M;
PImat[5,4,5]=(l5_0+l5_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M-step4D2I+step4D2M;
PImat[6,4,5]=(l6_0+l6_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M-step4D2I+step4D2M;
PImat[7,4,5]=(l7_0+l7_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M-step4D2I+step4D2M;
PImat[8,4,5]=(l8_0+l8_1)-step1D2I+step1D2M-step2D2I+step2D2M-
step3D2I+step3D2M-step4D2I+step4D2M;
}
model {
vector[Nc] contributionsC;
vector[Ni] contributionsI;
//Prior
l1_1~normal(0,20);
l2_1~normal(0,20);
l3_1~normal(0,20);
l4_1~normal(0,20);
l5_1~normal(0,20);
l6_1~normal(0,20);
l7_1~normal(0,20);
l8_1~normal(0,20);

l1_0~normal(0,20);
l2_0~normal(0,20);
l3_0~normal(0,20);
l4_0~normal(0,20);
l5_0~normal(0,20);
l6_0~normal(0,20);
l7_0~normal(0,20);
l8_0~normal(0,20);
```

```
step1D1I~normal(0,20);
step2D1I~normal(0,20);
step3D1I~normal(0,20);
step4D1I~normal(0,20);
step1D2I~normal(0,20);
step2D2I~normal(0,20);
step3D2I~normal(0,20);
step4D2I~normal(0,20);
step1D1M~normal(0,20);
step2D1M~normal(0,20);
step3D1M~normal(0,20);
step4D1M~normal(0,20);
step1D2M~normal(0,20);
step2D2M~normal(0,20);
step3D2M~normal(0,20);
step4D2M~normal(0,20);
Vc~dirichlet(rep_vector(2.0, Nc));

//Likelihood
for (iterp in 1:Np){
for (iterc in 1:Nc){
for (iteri in 1:Ni){
contributionsI[iteri]= categorical_lpmf(Y[iterp,iteri]+1|
softmax(((PImat[iteri,iterc]))));
}
contributionsC[iterc]=log(Vc[iterc])+sum(contributionsI);
}
target+=log_sum_exp(contributionsC);
}
}
```

References

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

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*, 1–37.

Chen, J., & de la Torre, J. (2018). Introducing the general polytomous diagnosis modeling framework. *Frontiers in Psychology*, 9.

Chen, Y., Culpepper, S. A., Chen, Y., & Douglas, J. (2018). Bayesian Estimation of the DINA **Q**-matrix. *Psychometrika*, *83*, 89–108.

de la Torre, J. (2010). The partial-credit DINA model. Paper presented at the international meeting of the Psychometric Society, Athens.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

Drasgow, F. (1995). Introduction to the polytomous IRT special issue. *Applied Psychological Measurement*, *19*, 1–3.

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

Gelman, A., & Rubin, D. B. (1992), Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511.

Gu, Y., & Xu, G. (2018). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika*. Advance online publication. https://doi.org/10.1007/s11336-018-9619-8

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 301–321.

Hansen, M. (2013). *Hierarchical item response models for cognitive diagnosis* (Unpublished Doctoral dissertation, UCLA).

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation. Champaign, IL: University of Illinois.

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*, 191–210.

Huggins-Manley, A. C., Algina, J., & Zhou, S. (2018). Models for semiordered data to address not applicable responses in scale measurement. *Structural Equation Modeling*, *25*, 230–243.

Jiang, Z., & Carter, R. (2018). Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnosis model via Stan. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-018-1069-9

Junker, B. W., Patz, R. J., & VanHoudnos, N. M. (2016). Markov chain Monte Carlo for item response models. In W. J. van der Linden (Ed.), Handbook of item response theory, Vol. 2: Statistical tools (pp. 271–312). Boca Raton: CRC Press.

Jurich, D. P., & Bradshaw, L. P. (2014). An illustration of diagnostic classification modeling in student learning outcomes assessment. *International Journal of Testing*, *14*, 49–72.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Liu, R. (2018). Misspecification of attribute structure in diagnostic measurement. *Educational and Psychological Measurement*, *78*, 605–634. https://doi.org/10.1177/0013164417702458

Liu, R., & Huggins-Manley, A. C. (2016). The specification of attribute structures and its effects on classification accuracy in diagnostic test design. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A.

Douglas, & M. Wiberg (Eds.), Quantitative psychology research (pp. 243–254). New York: Springer.

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

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement*, *78*, 357–383. https://doi.org/10.1177/0013164416685599

Liu, R., & Jiang, Z. (2018). Diagnostic classification models for ordinal item responses. *Frontiers in Psychology*, *9*, 2512. https://doi.org/10.3389/fpsyg.2018.02512

Liu, R., Qian, H., Luo, X., & Woo, A. (2017). Relative diagnostic profile: A subscore reporting framework. *Educational and Psychological Measurement*, *78*, 1072–1088. https://doi.org/10.1177/0013164417740170

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

Ma, W., & de la Torre, J. (2018). GDINA: The generalized DINA model framework (R package version 2.0). Retrieved from http://CRAN.R-project.org/package=GDINA.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*, 59–71.

Myers, I., & Myers, P. (1980). Gifts differing: Understanding personality type. Mountain View: Davies-Black.

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

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

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*, 219–262.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*,*34*(4, Pt. 2).

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, *38*, 203–219.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*, 251–275.

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

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Templin, J. L., Henson, R. A., Rupp, A. A., Jang, E., & 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.

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

Thorpe, G. L., & Favia, A. (2012). Data analysis using item response theory methodology: An introduction to selected programs and applications (Psychology Faculty Scholarship, Paper 20). Orono: University of Maine.

Tu, D., Zheng, C., Cai, Y., Gao, X., & Wang, D. (2018). A polytomous model of cognitive diagnostic assessment for graded data. *International Journal of Testing*, *18*(3), 231–252.

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

von Davier, M. (2005). A general diagnostic model applied to language testing data (ETS Research Report Series, 2005). Princeton: Educational Testing Service.

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, *81*, 625–649. https://doi.org/10.1007/s11336-015-9471-z

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions. *Bayesian Analysis*, *13*, 917–1007. https://doi.org/10.1214/17-BA1091