British Journal of Mathematical and Statistical Psychology (2016), 69, 253–275

The British Psychological Society

www.wileyonlinelibrary.com

A sequential cognitive diagnosis model for polytomous responses

Wenchao Ma* and Jimmy de la Torre

Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA

This paper proposes a general polytomous cognitive diagnosis model for a special type of graded responses, where item categories are attained in a sequential manner, and associated with some attributes explicitly. To relate categories to attributes, a category-level Q-matrix is used. When the attribute and category association is specified *a priori*, the proposed model has the flexibility to allow different cognitive processes (e.g., conjunctive, disjunctive) to be modelled at different categories within a single item. This model can be extended for items where categories cannot be explicitly linked to attributes, and for items with unordered categories. The feasibility of the proposed model is examined using simulated data. The proposed model is illustrated using the data from the Trends in International Mathematics and Science Study 2007 assessment.

1. Introduction

Cognitive diagnosis models (CDMs) have recently received increasing attention. Their aim is to classify examinees into different latent classes with unique attribute patterns indicating mastery or non-mastery of a number of skills or attributes of interest. Students with the same total score according to item response theory (IRT) or classical test theory (CTT) can have different attribute patterns, which offers additional information about students' strengths and weaknesses, thus informing instruction and remediation.

A host of CDMs can be found in the literature (for reviews, see, DiBello, Roussos, & Stout, 2007; Rupp & Templin, 2008), and many of them are developed based on strong cognitive assumptions about the processes involved in problem-solving. For example, the deterministic inputs, noisy 'AND' gate (DINA; Haertel, 1989) model assumes that examinees are expected to answer an item correctly only when they possess all required attributes; whereas the deterministic inputs, noisy 'OR' gate (DINO; Templin & Henson, 2006) model assumes that, in principle, examinees are able to answer an item successfully as long as they master at least one required attribute. Some general CDM frameworks subsuming a number of commonly used CDMs have also been developed, such as the generalized DINA (G-DINA; de la Torre, 2011) model, the log-linear CDM (LCDM; Henson, Templin, & Willse, 2009) and the general diagnostic model (GDM; von Davier, 2008). Although developed from different perspectives, the G-DINA model and the LCDM are equivalent in their saturated forms, both of which are special cases of the GDM.

Despite the number of CDMs available, most are targeted for dichotomous responses that stemmed primarily from multiple-choice items. However, the importance of

*Correspondence should be addressed to Wenchao Ma, Rutgers, The State University of New Jersey, 10 Seminary Place, New Brunswick, NJ 08901, USA (email: wenchao.ma@rutgers.edu).

DOI:10.1111/bmsp.12070

constructed-response items has recently been largely overlooked in cognitive diagnostic assessments. Theoretically, the constructed-response items should be able to provide more evidence to support inference about examinees' attribute patterns because they require examinees to explicitly show their problem-solving procedures. The merits of constructed-response items have also been empirically recognized. For example, Birenbaum and Tatsuoka (1987) administered a fraction addition test using open-ended and multiple-choice formats to diagnose students' misconceptions and found that open-ended items were more appropriate for the diagnostic purpose according to various criteria, such as the number of identified students' error types and diagnosis of students' sources of misconceptions. This conclusion has been further examined and verified by Birenbaum, Tatsuoka and Gutvirtz (1992), who also found that students used different cognitive processes when responding to items with different formats. For example, students may not really solve the problem in multiple-choice format as expected, but try to utilize the information in alternatives, which sometimes makes them more likely to obtain an incorrect solution.

Typically, although not always, constructed-response items are scored polytomously, yielding graded response data with ordered categories. To calibrate this type of data, one commonly used strategy is to dichotomize them so that they can be analysed using existing dichotomous CDMs (Johnson *et al.*, 2013; Su, 2013). However, the process of dichotomization often results in loss of information. To deal with polytomously scored items more appropriately, a few polytomous CDMs have been developed, such as the partial credit DINA (PC-DINA; de la Torre, 2010) model, the GDM for graded responses (pGDM; von Davier, 2008), the nominal response diagnostic model (NRDM; Templin, Henson, Rupp, Jang, & Ahmed, 2008), and the polytomous LCDM (Hansen, 2013). However, none of these polytomous CDMs consider the possible relation between attributes and response categories. Unlike polytomous IRT models where the latent trait has an impact on students' responses to all categories, in CDMs different categories could measure different attributes, as shown in an example in the next section. To take this information into account, in this paper a general polytomous CDM for graded responses has been developed. This model is referred to as the *sequential process model* to emphasize that a series of attributes is involved in the problem-solving process.

2. Attribute and category association

Suppose that solving an item consists of a finite number of sequential steps, each of which involves some attributes. Also, suppose that students are scored according to how many successive steps they have successfully performed. Specifically, a student falls into the zero category if s/he fails the first step; the first category if s/he performs the first step correctly but fails the second step; and so forth. In doing so, responses to items with H steps have $H + 1$ ordered categories, namely, category 0 to category H.

Take $4\frac{1}{8} - \frac{3}{8} = ?$ as an example. To solve this item, three steps may be involved: first $4\frac{1}{8}$ is transformed to $3\frac{9}{8}$ to allow fraction subtraction; then, by subtracting the numerators of two fractions, $3\frac{6}{8}$ can be obtained; and finally, $3\frac{6}{8}$ is simplified to $3\frac{3}{4}$. According to the attributes identified by Tatsuoka (1990), students need to know (A1) borrow from whole number part, (A2) subtract numerators, and (A3) reduce answers to the simplest form to succeed in step 1, 2 and 3, respectively. This example is for illustrative purposes only, and items in practice can be more complex. For example, some steps could consist of multiple substeps that are not sequential and some substeps may need multiple attributes.

Additionally, although response categories are assumed to be attained sequentially, different categories do not have to measure different attributes, nor must the attributes show any particular structure. For example, the attributes measured by lower categories do not have to be prerequisites to those required by higher categories.

To relate attributes to categories, the traditional Q-matrix (Tatsuoka, 1983) has been modified. The traditional Q-matrix is a $J \times K$ binary matrix specifying whether an attribute is measured by an item, where J is the test length and K is the number of attributes. The element q_{jk} in row j and column k is equal to 1 if attribute k is needed by item j, and 0 otherwise. For graded responses, a category-level Q-matrix is developed in this paper, referred to as Q_C-matrix, where subscript C is used to denote *category*. Throughout this paper, item j is assumed to have $H_j + 1$ categories (i.e., 0, 1, ..., H_j). The attribute and category association for item j is placed in H_j rows of the Q_C-matrix because category 0 does not require any attribute. Each of H_j rows has K elements indicating which attributes are required by the category. In particular, element 1 indicates that the attribute is required by this category, and 0 indicates that the attribute is not. The Q_C-matrix is a $\sum_{j=1}^{J} H_j \times K$ binary matrix, and if all items are scored dichotomously, the Q_C-matrix is equivalent to the traditional Q-matrix.

Table 1 gives the Q_C-matrix for the item $4\frac{1}{8} - \frac{3}{8} = ?$. The attribute and category association is specified in three rows to account for four categories. The required attributes for a category refer to the attributes required for the step that examinees need to solve to answer this category correctly after they have completed all previous steps successfully. For example, although the first two attributes are also indispensable to achieve category 3, it is not necessary to specify [1 1 1] because after examinees have already achieved category 2, only the third attribute is needed to perform category 3 correctly. The Q_C-matrix defined in this way is referred to as the *restricted* Q_C-matrix.

To create the restricted Q_C-matrix, the attribute and category association must be known *a priori*. However, this information may not be available, especially when CDMs are retrofitted to existing assessments. If so, it is reasonable to assume that all attributes required by an item are needed by each category of this item. The Q_C-matrix defined in this way is called the *unrestricted* Q_C-matrix. For the previous example, the unrestricted Q_C-matrix is given in Table 2.

3. Sequential process model

When a test measures K attributes, examinees can be grouped into 2^K latent classes, each having unique attribute pattern, that is, $\boldsymbol{\alpha}_c = (\alpha_{c1}, \ldots, \alpha_{cK})$, where $c = 1, \ldots, 2^K$. $\alpha_{ck} = 1$ indicates that attribute k is mastered by examinees in latent class c, and $\alpha_{ck} = 0$

Table 1. Restricted Q_C-matrix for $4\frac{1}{8} - \frac{3}{8} = ?$

		Attributes		
Step	Category	A1	A2	A3
$3\frac{9}{8} - \frac{3}{8}$	1	1	0	0
$3\frac{6}{8}$	2	0	1	0
$3\frac{3}{4}$	3	0	0	1

Notes. A1, borrow from whole number part; A2, subtract numerators; A3, reduce answers to the simplest form.

Table 2. Unrestricted Q_C-matrix for $4\frac{1}{8} - \frac{3}{8} = ?$

		Attributes		
Step	Category	A1	A2	A3
$3\frac{9}{8} - \frac{3}{8}$	1	1	1	1
$3\frac{6}{8}$	2	1	1	1
$3\frac{3}{4}$	3	1	1	1

Notes. A1, borrow from whole number part; A2, subtract numerators; A3, reduce answers to the simplest form.

indicates attribute k is not mastered by examinees in latent class c. Similar to Samejima (1995), we define the probability of examinees with attribute pattern $\boldsymbol{\alpha}_c$ answering category h of item j correctly provided that they have already completed category $h - 1$ successfully as the *processing function* of category h, denoted by $S_j(h|\boldsymbol{\alpha}_c)$, and we can reasonably assume that

$$S_j(h|\boldsymbol{\alpha}_c) = \begin{cases} 1, & \text{if } h = 0 \\ 0, & \text{if } h = H_j + 1, \end{cases}$$

because examinees can always achieve category 0, but never achieve category $H_j + 1$. Students score h if and only if they answer categories $1, \ldots, h$ correctly, and if h is not the highest category, category $h + 1$ incorrectly; therefore, the category response function for item j can be expressed as

$$P(X_j = h|\boldsymbol{\alpha}_c) = [1 - S_j(h+1|\boldsymbol{\alpha}_c)] \prod_{x=0}^{h} S_j(x|\boldsymbol{\alpha}_c), \tag{1}$$

subject to the constraints

$$\sum_{h=0}^{H_j} P(X_j = h|\boldsymbol{\alpha}_c) = 1 \quad \forall c,$$

where $h = 0, \ldots, H_j$, and $P(X_j = h|\boldsymbol{\alpha}_c)$ is the probability of examinees with attribute pattern $\boldsymbol{\alpha}_c$ scoring h on item j. It is reasonable to assume that the processing function $S_j(h|\boldsymbol{\alpha}_c)$ is a function of examinees' attribute patterns and the required attributes for category h of item j. The processing function is the kernel of the sequential process model, and can be formulated using most dichotomous CDMs. For example, if solving a step entails the possession of all required attributes, the DINA model can be used as the processing function. By parametrizing each category separately, the sequential process model allows different cognitive processes to be modelled at different categories within a single item.

3.1. Sequential G-DINA model

In this paper, the G-DINA model (de la Torre, 2011) is used as the processing function because it offers a general framework subsuming several widely used CDMs. The resulting model is referred to as the *sequential G-DINA model*.

Like the G-DINA model, for item j, 2^K latent classes can be collapsed into $2^{K_j^*}$ latent groups with unique probabilities of success, where K_j^* is the number of required attributes for item j. For category h, $2^{K_j^*}$ latent groups can be further collapsed into $2^{K_{jh}^*}$ latent groups, where K_{jh}^* is the number of required attributes for category h of item j. Let $\boldsymbol{\alpha}_{ljh}^*$ be the reduced attribute vector for category h of item j consisting of the required attributes for this category only, where $l = 1, \ldots, 2^{K_{jh}^*}$. Without loss of generality, we can assume that the first K_{jh}^* attributes are required for category h of item j, that is, $\boldsymbol{\alpha}_{ljh}^* = [\alpha_{l1}, \ldots, \alpha_{lk}, \ldots, \alpha_{lK_{jh}^*}]$. The processing function $S_j(h|\boldsymbol{\alpha}_c)$ can be written as $S_j(h|\boldsymbol{\alpha}_{ljh}^*)$, and formulated using the identity link G-DINA model:

$$S_j(h|\boldsymbol{\alpha}_{ljh}^*) = \phi_{jh0} + \sum_{k=1}^{K_{jh}^*} \phi_{jhk}\alpha_{lk} + \sum_{k'=k+1}^{K_{jh}^*} \sum_{k=1}^{K_{jh}^*-1} \phi_{jhkk'}\alpha_{lk}\alpha_{lk'} + \ldots + \phi_{jh12\ldots K_{jh}^*} \prod_{k=1}^{K_{jh}^*} \alpha_{lk}, \tag{2}$$

where ϕ_{jh0} is the intercept, ϕ_{jhk} is the main effect due to α_{lk}, $\phi_{jhkk'}$ is the two-way interaction effect due to α_{lk} and $\alpha_{lk'}$, and $\phi_{jh12\ldots K_{jh}^*}$ is K_{jh}^*-way interaction effect due to α_{lk} to $\alpha_{lK_{jh}^*}$. ϕ_{jh0} represents the processing function of category h for examinees who master none of required attributes, ϕ_{jhk} is the change of processing function of category h due to the mastery of attribute k, and interaction coefficients represent the change in the processing function of category h due to the mastery of all relevant attributes that is over and above all impact of lower-order effects. Like the G-DINA model, the processing function can also be defined using a log or logit link function. For category h of item j, there are $2^{K_{jh}^*}$ item parameters, as in, $\boldsymbol{\phi}_{jh} = \{\phi_{jh0}, \phi_{jh1}, \ldots, \phi_{jh12\ldots K_{jh}^*}\}$. By defining processing functions $\boldsymbol{S}_j(h|\boldsymbol{\alpha}_{jh}^*) = \{S_j(h|\boldsymbol{\alpha}_{ljh}^*)\}$, $\boldsymbol{\phi}_{jh}$ can be derived from $\boldsymbol{S}_j(h|\boldsymbol{\alpha}_{jh}^*)$ directly because equation (2) can be expressed as $\boldsymbol{S}_j(h|\boldsymbol{\alpha}_{jh}^*) = \boldsymbol{M}_{jh}\boldsymbol{\phi}_{jh}$, where $\boldsymbol{M}_{jh}$ is an invertible design matrix of dimension $2^{K_{jh}^*} \times 2^{K_{jh}^*}$ (see de la Torre, 2011; for details about the design matrix). This implies that processing functions can also be viewed as item parameters, though this is not true if constraints are added to the processing functions.

As shown by de la Torre (2011), by setting appropriate constraints in the G-DINA model, the DINA model, DINO model, *A*-CDM, linear logistic model (LLM; Maris, 1999) and reduced reparametrized unified model (R-RUM; Hartz, 2002) can be obtained. Those models can also be specified as the processing functions using similar constraints in the sequential G-DINA model. See de la Torre (2011) for details about the appropriate constraints.

The sequential G-DINA model can use either a restricted or unrestricted Q_C-matrix. For notational convenience, the sequential G-DINA model using a restricted Q_C-matrix is called the restricted sequential G-DINA (RS-GDINA) model, while the sequential G-DINA model using an unrestricted Q_C-matrix is called the unrestricted sequential G-DINA (US-GDINA) model. The use of a restricted Q_C-matrix allows us to model different underlying processes in different response categories. The use of an unrestricted Q_C-matrix, on the other hand, provides a possible solution to account for the uncertainty in the attribute and category association. When the attribute and category association is available, the RS-GDINA model may be preferred theoretically because it usually estimates fewer item parameters than the US-GDINA model. Regarding the aforementioned example, the RS-GDINA model has six item parameters but the US-GDINA model has 24, which implies that additional 18 parameters for this single item need to be estimated when using the

unrestricted Q_C-matrix. Nevertheless, the practical consequence of estimating extra parameters needs further empirical examination.

3.2. Parameter estimation

Item parameters of the sequential G-DINA model can be estimated using the marginal maximum likelihood estimation approach via expectation maximization (MMLE/EM) algorithm (Bock & Aitkin, 1981). Let $\boldsymbol{\alpha}_{lj}^*$ be the reduced attribute pattern for the lth collapsed latent group for item j, where $l = 1, \ldots, 2^{K_j^*}$. Also, let X_{ij} be the response of examinee i to item j, where $i = 1, \ldots, N$. Under the assumption of local independence, the conditional probability of the response vector $\mathbf{X}_i$ can be written as

$$P(\mathbf{X}_i|\boldsymbol{\alpha}_{lj}^*) = \prod_{j=1}^{J}\prod_{h=0}^{H_j} P(X_j = h|\boldsymbol{\alpha}_{lj}^*)^{I(X_{ij}=h)},$$

where $I(X_{ij} = h)$ is an indicator variable evaluating whether X_{ij} is equal to h. The MMLE/EM algorithm implements the E-step and M-step iteratively item by item until convergence. In particular, for item j, based on the provisional item parameter estimates and the distribution of reduced latent classes $p(\boldsymbol{\alpha}_{lj}^*)$, the E-step calculates the expected number of examinees with attribute pattern $\boldsymbol{\alpha}_{lj}^*$ scoring in category h, that is,

$$\bar{r}_{ljh} = \sum_{i=1}^{N} I(X_{ij} = h)P(\boldsymbol{\alpha}_{lj}^*|\mathbf{X}_i),$$

where $P(\boldsymbol{\alpha}_{lj}^*|\mathbf{X}_i)$ is the posterior probability of examinee i having reduced attribute pattern $\boldsymbol{\alpha}_{lj}^*$, and can be calculated by

$$P(\boldsymbol{\alpha}_{lj}^*|\mathbf{X}_i) = \frac{P(\mathbf{X}_i|\boldsymbol{\alpha}_{lj}^*)p(\boldsymbol{\alpha}_{lj}^*)}{\sum_{l=1}^{2^{K_j^*}} P(\mathbf{X}_i|\boldsymbol{\alpha}_{lj}^*)p(\boldsymbol{\alpha}_{lj}^*)}.$$

In the M-step, the objective function

$$\ell = \sum_{l=1}^{2^{K_j^*}}\sum_{h=0}^{H_j} \bar{r}_{ljh} \log\left[\hat{P}(X_j = h|\boldsymbol{\alpha}_{lj}^*)\right]$$

needs to be maximized with respect to item parameters $\boldsymbol{\phi}_j$, which is a vector of length $\sum_{h=1}^{H_j} 2^{K_{jh}^*}$ when the step function is the G-DINA model, using some general optimization techniques. The two steps are repeated until convergence. In this study, the Nelder and Mead (1965) simplex method is used for the M-step optimization. It is one of the most popular derivative-free optimization technique applicable for multidimensional non-linear problems. After generating a geometric simplex, its convergence is guided by moving the simplex appropriately (Nelder & Mead, 1965). It should be noted that although the Nelder and Mead method is robust and the default method for the optim function in R (R Core Team, 2015), other optimization techniques such as quasi-Newton methods can also be employed as alternatives. For estimating the joint attribute

distribution, an empirical Bayes method (Carlin & Louis, 2000) is adopted. Specifically, the prior distribution of latent classes is uniform at the beginning, and then updated after each EM iteration based on the posterior distribution, as in $p(\boldsymbol{\alpha}_c) = \sum_{i=1}^{N} P(\boldsymbol{\alpha}_c|\mathbf{X}_i)/N$.

Note that the above MMLE/EM algorithm is suitable for both the RS-GDINA and US-GDINA models. However, for the US-GDINA model using the G-DINA model as the processing function, item parameter estimates in the M-step can be obtained via closed-form solutions; therefore, the general optimization routine is not necessary. After a few algebraic manipulations and simplifications, we have

$$\hat{P}(X_j = h|\boldsymbol{\alpha}_{lj}^*) = \frac{\sum_{i=1}^{N} I(X_{ij} = h) P(\boldsymbol{\alpha}_{lj}^*|\mathbf{X}_i)}{\sum_{i=1}^{N} P(\boldsymbol{\alpha}_{lj}^*|\mathbf{X}_i)}.$$

By substituting $\hat{P}(X_j = h|\boldsymbol{\alpha}_{lj}^*)$ into equation (1), the marginal likelihood estimates of $S_j(h|\boldsymbol{\alpha}_{lj}^*)$ can be obtained. Then, item parameter $\boldsymbol{\phi}$ can be estimated via the least-squares method as introduced by de la Torre (2011). After estimating item parameters, expected a posteriori (EAP) can be used to estimate individuals' attribute patterns.

3.3. Relations with existing polytomous CDMs

Although the US-GDINA model was originally developed for ordered responses with unknown category and attribute association, it has been found to be suitable for nominal response data as well. In particular, the US-GDINA model can be shown to be equivalent to the NRDM (Templin *et al.*, 2008) and the PC-DINA model (de la Torre, 2010) when the processing function is the G-DINA and DINA model, respectively. The equivalence becomes evident when we view all of them as the CDM counterparts of Bock's (1972) nominal response model involving direct estimation of category response functions. For instance, the category response function of the NRDM can be reparametrized using the identity link as

$$P(X_j = h|\boldsymbol{\alpha}_{lj}^*) = \delta_{jh0} + \sum_{k=1}^{K_j^*} \delta_{jhk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jhkk'}\alpha_{lk}\alpha_{lk'} + \ldots + \delta_{jh12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}, \quad (3)$$

with the constraint $\sum_{h=0}^{H_j} P(X_j = h|\boldsymbol{\alpha}_{lj}^*) = 1$. It can be shown that estimating $\boldsymbol{\delta} = \{\delta_{jh0}, \delta_{jh1}, \ldots, \delta_{jh12\ldots K_{jh}^*}\}$ is equivalent to estimating the category response function $P(X_j = h|\boldsymbol{\alpha}_{lj}^*)$ because they can be derived directly from each other. Bearing this in mind, the category response function of the sequential G-DINA model in equation (1) can then act as a bridge between these two models. See the Appendix S1 for more details.

In spite of the equivalence, the importance of the US-GDINA model should not be overlooked. Both RS-GDINA and US-GDINA models are special cases of the sequential G-DINA model, with the only difference in how the Q_C-matrix is constructed to reflect our knowledge (or lack thereof) of the category and attribute association. The development of the US-GDINA model allows us to see that the NRDM and PC-DINA models are special cases of the sequential G-DINA model (i.e., when responses are treated as nominal data). As a result, the proposed sequential G-DINA model can serve as a very general model framework so that researchers are able to calibrate simultaneously a number of different

CDMs for dichotomous, ordered, or unordered polytomous responses with or without specific assumptions about the cognitive processes (e.g., conjunctive, disjunctive or additive) for a single assessment. In addition, when fitted to ordered responses where categories are attained sequentially, the processing functions from the US-GDINA model can provide extra information that the NRDM and PC-DINA models typically do not provide. Finally, the MMLE/EM algorithm developed for the sequential G-DINA model is another contribution of this work in that it provides a much faster alternative to the Markov chain Monte Carlo algorithm originally used for the NRDM (Templin *et al.*, 2008).

4. Simulation study

Two simulation studies were conducted to evaluate the performance of the sequential G-DINA model under various conditions. The processing functions used in the simulation studies were the G-DINA model, unless otherwise stated. Study 1 examined whether parameters of the sequential G-DINA model can be recovered accurately based on the proposed estimation algorithm; whether the sequential G-DINA model can provide more accurate person classifications than the G-DINA model using dichotomized responses; and whether the attribute and category association can be used to improve parameter recovery for the sequential G-DINA model.

The appropriateness of the RS-GDINA model depends upon whether the observed processing functions are in accordance with those predicted by the attribute and category association. If the predicted processing functions deviate dramatically from the observed ones, the US-GDINA model may be more appropriate because it relaxes the assumption about the attribute and category association. Study 2 investigated the impact of the discrepancy between the observed and predicted processing functions on parameter estimation. The likelihood ratio test (LRT), Akaike information criterion (AIC; Akaike, 1974), and Bayesian information criterion (BIC; Schwarz, 1978) have been widely used for model comparison within the CDM context (de la Torre & Lee, 2013; DeCarlo, 2011; Kunina-Habenicht, Rupp, & Wilhelm, 2012). Study 2 also examined whether these indices can be used to select the appropriate sequential G-DINA model under various degrees of discrepancy.

Sixteen polytomous items and five dichotomous items were used for the data simulation. Five attributes were measured by these items. The restricted Q_C-matrix is given in Table 3, where all attributes are measured the same number of times.

4.1. Study 1

4.1.1. Design

Sample size and item quality were controlled in this study. $N = 500$, 1,000, 2,000 or 4,000 examinees were drawn from a uniform attribute distribution. Item responses were simulated based on the RS-GDINA model using the restricted Q_C-matrix in Table 3. Item j was of high quality when $S_j(h|\boldsymbol{\alpha}^*_{ljb} = \mathbf{1}) = .9$ and $S_j(h|\boldsymbol{\alpha}^*_{ljb} = \mathbf{0}) = .1$, moderate quality when $S_j(h|\boldsymbol{\alpha}^*_{ljb} = \mathbf{1}) = .8$ and $S_j(h|\boldsymbol{\alpha}^*_{ljb} = \mathbf{0}) = .2$, and low quality when $S_j(h|\boldsymbol{\alpha}^*_{ljb} = \mathbf{1}) = .7$ and $S_j(h|\boldsymbol{\alpha}^*_{ljb} = \mathbf{0}) = .3$, for all categories. When $K^*_{jb} > 1$, the processing functions for latent classes with $\boldsymbol{\alpha}^*_{ljb}$ not equal to $\mathbf{0}$ or $\mathbf{1}$, that is, $S_j(h|\boldsymbol{\alpha}^*_{ljb} \notin \{\mathbf{0}, \mathbf{1}\})$, were drawn from a uniform distribution $U[S_j(h|\boldsymbol{\alpha}^*_{ljb} = \mathbf{0}), S_j(h|\boldsymbol{\alpha}^*_{ljb} = \mathbf{1})]$. The processing functions were simulated with the monotonicity

Table 3. Restricted Q_C-matrix for data simulation

Item	Category	A1	A2	A3	A4	A5	Item	Category	A1	A2	A3	A4	A5
1	1	1	0	0	0	0	11	1	1	1	0	0	0
1	2	0	1	0	0	0	11	2	0	0	0	0	1
2	1	0	0	1	0	0	12	1	1	1	1	0	0
2	2	0	0	0	1	0	12	2	0	0	0	1	1
3	1	0	0	0	0	1	13	1	1	1	0	0	0
3	2	1	0	0	0	0	13	2	0	0	1	1	1
4	1	0	0	0	0	1	14	1	1	0	1	0	0
4	2	0	0	0	1	0	14	2	0	0	0	1	0
5	1	0	0	1	0	0	14	3	0	0	0	0	1
5	2	0	1	0	0	0	15	1	0	0	0	0	1
6	1	1	0	0	0	0	15	2	0	0	1	1	0
6	2	0	1	1	0	0	15	3	0	1	0	0	0
7	1	0	0	1	0	0	16	1	1	0	0	0	0
7	2	0	0	0	1	1	16	2	0	1	0	0	0
8	1	0	0	0	0	1	16	3	0	0	1	1	0
8	2	1	1	0	0	0	17	1	1	0	0	0	0
9	1	0	0	0	1	1	18	1	0	1	0	0	0
9	2	0	0	1	0	0	19	1	0	0	1	0	0
10	1	0	1	0	1	0	20	1	0	0	0	1	0
10	2	1	0	0	0	0	21	1	0	0	0	0	1

constraint that examinees mastering additional attributes would not have a lower processing function. Note that, to easily control item quality, the processing functions are manipulated directly instead of ϕ because either of them can be viewed as item parameters when the G-DINA model is used as the processing function. Based on simulated processing functions, category response functions of item j can be calculated, as in $\boldsymbol{P}_{lj} = [P(X_j = 0|\boldsymbol{\alpha}^*_{lj}), \ldots, P(X_j = H_j|\boldsymbol{\alpha}^*_{lj})]$. Responses of examinees with attribute pattern $\boldsymbol{\alpha}^*_{lj}$ were generated from a Bernoulli and generalized Bernoulli distribution with parameters $\boldsymbol{P}_{lj}$, if item j is scored dichotomously and polytomously, respectively. To reduce Monte Carlo sampling errors, 100 data sets were generated in each condition.

Both the US-GDINA and the RS-GDINA model were fitted to simulated data. To fit the US-GDINA model, the unrestricted Q_C-matrix needs to be constructed from Table 3. Specifically, for each item, attributes required by a category are also assumed to be required by all other categories. Taking item 15 as an example, in the unrestricted Q_C-matrix all three categories require the last four attributes. To fit the G-DINA model, polytomous responses were dichotomized in two ways: for one, partial credit and full marks were converted to 1; for the other, only full marks were converted to 1, and partial credit was transformed to 0. In either case, the q-vector of an item in the traditional Q-matrix is specified to measure all required attributes for each category of this item. For example, the q-vector for item 15 is [0 1 1 1 1]. The code for implementing the MMLE/EM algorithm presented in the previous section was written in R, and can be requested from the first author.

Item parameter recovery was examined using the root mean square error (RMSE) of the estimated category response function for each latent class from the true, that is,

$$\text{RMSE} = \sqrt{\frac{\sum_{r=1}^{R}\sum_{c=1}^{2^K}\sum_{j=1}^{J}\left[\hat{P}^{(r)}(X_j = h|\boldsymbol{\alpha}_c) - P^{(r)}(X_j = h|\boldsymbol{\alpha}_c)\right]^2}{J \times 2^K \times R}},$$

where J, K and R are the number of items, attributes and replications, respectively, and $\hat{P}^{(r)}(X_j = h|\boldsymbol{\alpha}_c)$ and $P^{(r)}(X_j = h|\boldsymbol{\alpha}_c)$ are the estimated and true probability of scoring in category h of item j for examinees with attribute pattern $\boldsymbol{\alpha}_c$ for the rth replication, respectively. Note that the RMSE was only calculated for the sequential G-DINA model.

Person parameter recovery was evaluated using the proportion of correctly classified attribute vectors (PCV), defined as

$$\text{PCV} = \frac{\sum_{r=1}^{R}\sum_{i=1}^{N} I^{(r)}[\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i]}{N \times R},$$

where $I^{(r)}[\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i]$ is an indicator variable evaluating whether the estimated attribute vector matches the true for the rth replication.

4.1.2. Results

Figure 1 gives the RMSEs of the RS-GDINA and US-GDINA models under various conditions. It is worth emphasizing that the data were generated using the RS-GDINA model; therefore, to evaluate item parameter recovery, we only focused on the RMSEs of the RS-GDINA model. The RMSEs of the RS-GDINA model were between .012 and .090, with the largest value occurring when N = 500 and item quality was low. Although, as expected, sample size and item quality have an impact on parameter estimation, the fact that the maximum RMSE was <.1 shows that item parameters can be recovered accurately based on the proposed estimation algorithm.

In addition, the RS-GDINA model always had smaller RMSEs than the US-GDINA model, as shown in Figure 1. The difference in RMSE between these two models can be larger than .1, when item quality was low and sample size was relatively small. For example, the RMSE of the RS-GDINA model was lower than that of the US-GDINA model by .126, when N = 1,000 and item quality was low. This suggests that the attribute and category association can provide important information for accurate item parameter estimation for the sequential G-DINA model, especially under the condition of low item quality and small sample sizes.

Table 4 gives the PCVs for the sequential G-DINA model and the G-DINA model. The results showed that the manner of dichotomization influenced the classification rates of the G-DINA model. Converting both partial credit and full marks to one point (denoted by GDINA1) produced better person classification rates than converting only full marks to one point (denoted by GDINA2) by up to 7.8%, with the maximum difference occurring when N = 4,000 and item quality was moderate. However, this conclusion may not hold when other dichotomous CDMs rather than the G-DINA model are employed, and therefore needs further investigation.

The sequential G-DINA model produced better person classifications than the G-DINA model fitted to dichotomized responses across all conditions. This conclusion holds regardless of the manner of dichotomization. For example, when item quality was moderate and N = 500, the RS-GDINA model outperformed the GDINA1 by 23.7% in

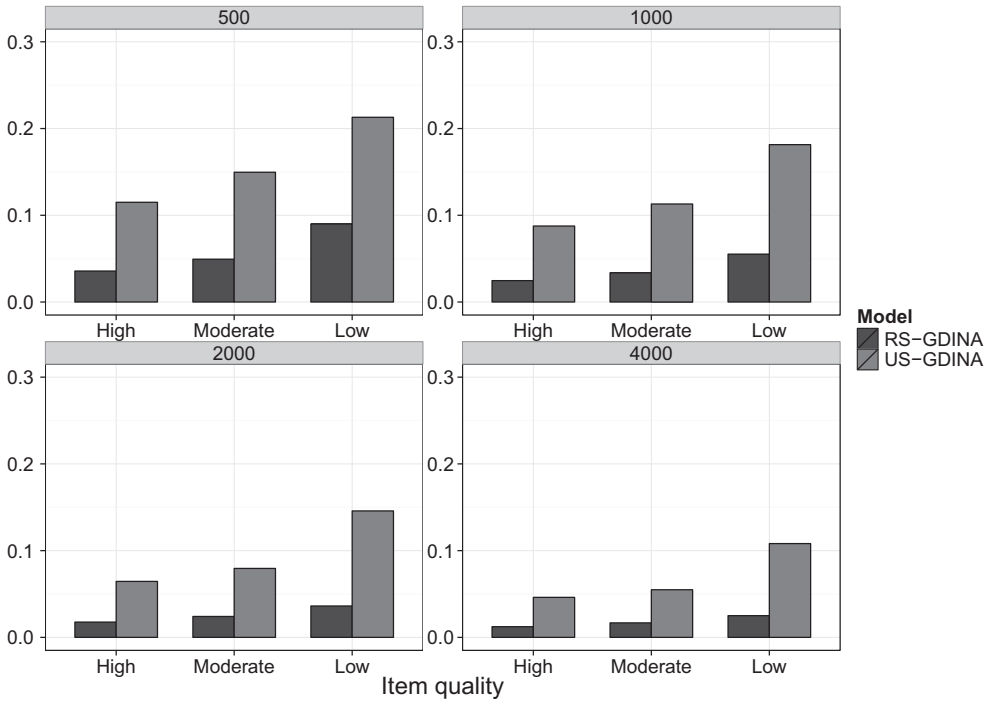


Figure 1. RMSE of the sequential G-DINA models.

terms of the PCV. Additionally, the RS-GDINA model produced better person classifications than the US-GDINA model across all conditions. The difference in PCV between these two models was noticeable when item quality was low and sample size was small. For example, with items of low quality, the RS-GDINA model outperforms the US-GDINA model by 8.7% and 10.0% when sample sizes were 500 and 1,000, respectively. When item quality was high or sample size was large, however, the difference tended to be negligible. These results imply that when enough information is provided through large sample size and high-quality items, the category and attribute association can offer limited extra information to improve the classification; whereas when sample size is small and item quality is low, the category and attribute association can be very important for accurate attribute estimation.

4.2. Study 2

4.2.1. Design

This study explored whether the LRT, AIC and BIC can be used to choose between the RS-GDINA and US-GDINA models. The factors examined in this study included sample size, item quality, fitted models, and magnitude of disturbances. The settings of sample size and item quality were the same as in the previous study. To quantify the uncertainty in attribute and category association, small or large disturbances were added to the simulated processing functions. Specifically, the processing functions of each item were first simulated based on the RS-GDINA model using the restricted Q_C-matrix in Table 3. Take the first item as an example. When item quality is moderate, the simulated processing functions of the second category are $S(h = 2|\boldsymbol{\alpha}^*_{ljh} = 0) = .2$ and $S(h = 2|\boldsymbol{\alpha}^*_{ljh} = 1) = .8$, or

Table 4. PCVs for the sequential G-DINA models and the G-DINA model

N	Item quality	RS-GDINA	US-GDINA	GDINA1	GDINA2
500	High	.917	.901	.759	.742
	Moderate	.679	.601	.442	.371
	Low	.310	.223	.182	.142
1,000	High	.921	.911	.790	.752
	Moderate	.692	.641	.476	.415
	Low	.354	.254	.199	.147
2,000	High	.923	.918	.799	.761
	Moderate	.697	.674	.501	.441
	Low	.372	.295	.217	.152
4,000	High	.924	.921	.809	.765
	Moderate	.702	.690	.534	.456
	Low	.384	.339	.240	.168

Notes. RS-GDINA, the restricted sequential G-DINA model; US-GDINA, the unrestricted sequential G-DINA model; GDINA1, the G-DINA model (examinees get one point as long as they get partial credit); GDINA2, the G-DINA model (examinees get one point only if they get full marks).

equivalently, $S(h = 2|\boldsymbol{\alpha}_{lj}^{*} = 00) = S(h = 2|\boldsymbol{\alpha}_{lj}^{*} = 10) = .2$ and $S(h = 2|\boldsymbol{\alpha}_{lj}^{*} = 01) = S(h = 2|\boldsymbol{\alpha}_{lj}^{*} = 11) = .8$. Then random disturbances $\boldsymbol{\varepsilon}$ were added to the simulated processing functions of $\boldsymbol{\alpha}_{lj}^{*}$. $\varepsilon \sim U[-0.1, 0.1]$ indicates a small disturbance and $\varepsilon \sim U[-0.2, 0.2]$ a large disturbance, where U represents the uniform distribution. Large disturbance implies a large discrepancy between the data and the RS-GDINA model. For notational simplicity, small and large disturbance are referred to as $\varepsilon = 0.1$ and $\varepsilon = 0.2$, respectively. If the processing function is >1 or <0 after adding the disturbance, it is set to .99 and .01, respectively. Study 1 can be viewed as a condition where $\varepsilon = 0$. In each condition, 100 data sets were generated and fitted by both the RS-GDINA and US-GDINA models. Note that adding random disturbances in simulating data based on the RS-GDINA model is equivalent to simulating data from the US-GDINA model. However, the size of disturbances can help quantify how 'wrong' the prespecified attribute and category association is.

The AIC and BIC were used for model comparison. The LRT was also employed as the RS-GDINA model is nested within the US-GDINA model. Specifically, $\Delta\chi^2$ can be calculated as the difference in -2 times the log-likelihood of two models. There are $\sum_{j=1}^{J}\sum_{h=1}^{H_j} 2^{K_{jh}^{*}}$ item parameters and $2^K - 1$ latent class parameters, that is, 145 and 449 parameters for the RS-GDINA and US-GDINA models, respectively, based on the Q_C-matrix in Table 3. Accordingly, $\Delta\chi^2$ follows a χ^2 distribution with 304 degrees of freedom. LRTs were conducted at the .05 significant level. To understand the properties of the LRT, AIC, and BIC, the proportion choosing the US-GDINA model for each statistic was examined. The PCV based on the models selected by the LRT, AIC, and BIC was calculated as well.

4.2.2. Results

Table 5 gives the proportion choosing the US-GDINA model for the LRT, AIC, and BIC. The results for $\varepsilon = 0$ represent Type I errors, which were obtained by reanalysing the data in Study 1. When $\varepsilon = 0$, the AIC and BIC correctly chose the RS-GDINA model for all replications across all conditions; whereas the LRT yielded inflated Type I errors (ranging from .11 to 1) when item quality was moderate or low.

Table 5. Proportion choosing the US-GDINA model

		ε = 0			ε = 0.1			ε = 0.2		
N	Item quality	LRT	AIC	BIC	LRT	AIC	BIC	LRT	AIC	BIC
500	High	.14	.00	.00	1.00	.01	.00	1.00	1.00	.00
	Moderate	.90	.00	.00	1.00	.00	.00	1.00	.99	.00
	Low	1.00	.00	.00	1.00	.00	.00	1.00	.31	.00
1,000	High	.08	.00	.00	1.00	1.00	.00	1.00	1.00	.00
	Moderate	.57	.00	.00	1.00	.02	.00	1.00	1.00	.00
	Low	1.00	.00	.00	1.00	.01	.00	1.00	.98	.00
2,000	High	.02	.00	.00	1.00	1.00	.00	1.00	1.00	.98
	Moderate	.24	.00	.00	1.00	.90	.00	1.00	1.00	.14
	Low	1.00	.00	.00	1.00	.03	.00	1.00	1.00	.00
4,000	High	.06	.00	.00	1.00	1.00	.12	1.00	1.00	1.00
	Moderate	.11	.00	.00	1.00	1.00	.00	1.00	1.00	1.00
	Low	.96	.00	.00	1.00	.67	.00	1.00	1.00	.00

Notes. LRT, likelihood ratio test; AIC, Akaike information criterion (Akaike, 1974); BIC, Bayesian information criterion (Schwarz, 1978).

When noise was added (i.e., ε = 0.1 and ε = 0.2), the LRT was always able to identify this deviation from the RS-GDINA model and chose the US-GDINA model under all conditions. The BIC, however, consistently chose the RS-GDINA model when ε = 0.1, with only one exception occurring when item quality was high and sample size was 4,000. When ε = 0.2, the BIC still tended to select the RS-GDINA model, especially when sample size was relative small or items quality was low.

When ε = 0.2, the proportion choosing the US-GDINA model for the AIC was >98% in all conditions, except when sample size was 500 and item quality was low, where the proportion was 31%. With small disturbances, the AIC preferred the US-GDINA model when items were of high quality and samples were of relatively large sizes. For example, when N = 500 and 4,000, the proportions choosing US-GDINA model for the AIC were <1% and >67%, respectively. When N = 1,000 or 2,000, the proportion choosing US-GDINA model for the AIC increased as item quality improved.

Table 6 gives the PCVs for the RS-GDINA model, the US-GDINA model and selected models using LRT, AIC, and BIC when disturbances were added. The PCVs for RS-GDINA model and US-GDINA model when ε = 0 were given in Table 4. Given that both the AIC and BIC always chose the RS-GDINA model when ε = 0, the PCV results for ε = 0 are omitted from Table 6. When ε = 0.1, the RS-GDINA model produced comparable or even better classification rates than the US-GDINA model, especially when sample sizes were relatively small and item quality was low. With large disturbances, the US-GDINA model outperformed the RS-GDINA model in all conditions, with one exception occurring when N = 500 and item quality was low. The difference in PCV between these two models was up to 8.5%, with the maximum value occurring when N = 4,000 and items were of low quality.

To compare the LRT, AIC, and BIC, the highest and lowest PCV values of the RS-GDINA and US-GDINA models are used as the upper and lower benchmarks, respectively. Across all conditions, selected models based on these three statistics yielded comparable or higher PCV values than the lower benchmark, which implies that all of them are useful for model selection. The PCVs of models selected using the LRT can be lower than the upper

Table 6. PCVs of the sequential G-DINA models and selected models using the LRT, AIC, and BIC

		ε = 0.1					ε = 0.2				
N	Item quality	LRT	AIC	BIC	RS-GDINA	US-GDINA	LRT	AIC	BIC	RS-GDINA	US-GDINA
500	High	.927	.926	.926	.926	.927	.935	.935	.910	.910	.935
	Moderate	.644	.690	.690	.690	.644	.757	.757	.719	.719	.757
	Low	.252	.324	.324	.324	.252	.330	.347	.347	.347	.330
1,000	High	.934	.934	.928	.928	.934	.947	.947	.921	.921	.947
	Moderate	.676	.698	.698	.698	.676	.790	.790	.726	.726	.790
	Low	.286	.359	.360	.360	.286	.403	.403	.387	.387	.403
2,000	High	.940	.940	.930	.930	.940	.951	.951	.950	.923	.951
	Moderate	.709	.709	.707	.707	.709	.806	.806	.744	.733	.806
	Low	.335	.383	.384	.384	.335	.468	.468	.412	.412	.468
4,000	High	.943	.943	.934	.933	.943	.952	.952	.952	.925	.952
	Moderate	.720	.720	.708	.708	.720	.814	.814	.814	.734	.814
	Low	.384	.389	.396	.396	.384	.501	.501	.416	.416	.501

Notes. LRT, likelihood ratio test; AIC, Akaike information criterion (Akaike, 1974); BIC, Bayesian information criterion (Schwarz, 1978); RS-GDINA, restricted sequential G-DINA model; US-GDINA, unrestricted sequential G-DINA model.

benchmark by up to 10%, with the value of 10% occurring when $\varepsilon = 0$, $N = 1{,}000$, and item quality was low (which is not given in Table 6). Note that, in this condition, the LRT always chose the US-GDINA model incorrectly for all replications, as shown in Table 5. For the BIC, the maximum difference in PCV between the selected models and the upper benchmark is 8.5%, which occurred when $\varepsilon = 0.2$, $N = 4{,}000$ and items were of low quality. Finally, the AIC selected the optimal models when $\varepsilon = 0.2$, which yielded the same PCV as the upper benchmark. When $\varepsilon = 0.1$, AIC also produced almost optimal model selections: the maximum difference in PCV between the selected models and the upper benchmark is 0.7%, occurring when $N = 4{,}000$ and item quality is low. These results suggest that, compared with the LRT and BIC, models selected by AIC yielded desirable person classification rates in terms of the PCV under all conditions.

5. Real data illustration

5.1. Data

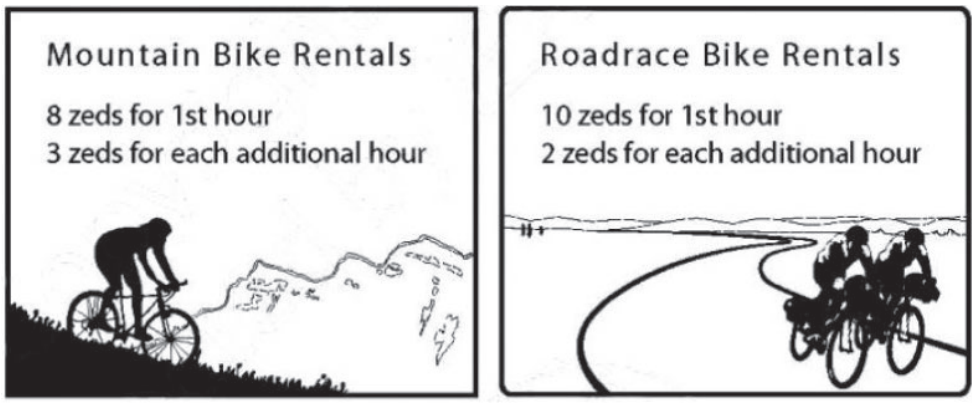
The data for this illustration were a subset of the data originally used by Lee, Park, and Taylan (2011), and were taken from booklets 4 and 5 of the Trends in International Mathematics and Science Study (TIMSS) 2007 fourth-grade mathematics assessment. The responses of 823 students to 12 of 25 items involving eight of the original 15 attributes identified by Lee *et al.* (2011) were used in the current study. The definitions of the attributes are given in Table 7. Two of the five constructed-response items (items 3 and 9) were scored polytomously with three ordered response categories (0, 1 and 2). Items M031242A and M031242B, referred to as items 7a and 7b respectively, are related to a common stimulus as shown in Figure 2. The former requires students to complete tables using the information in two posters, whereas the latter requires one of the correct answers: '3 (as long as does not contradict Part A [i.e., item 7a] including table empty or incomplete)' or 'number(s) correct according to a complete but erroneous table in Part A *or* indicates no match according to a complete but erroneous table in Part A' (Foy & Olson, 2009, p. 95). The scoring rule for the latter item implies a heavy dependence between the two items, which has also been found by Hansen (2013) when examining testlet effects.

Table 7. Attribute definitions for TIMSS 2007 data

A1	Representing, comparing, and ordering whole numbers as well as demonstrating knowledge of place value
A2	Recognizing multiples, computing with whole numbers using the four operations, and estimating computations
A3	Solving problems, including those set in real-life contexts (e.g., measurement and money problems)
A4	Finding the missing number or operation and modelling simple situations involving unknowns in number sentence or expression
A5	Describing relationships in patterns and their extensions; generating pairs of whole numbers by a given rule and identifying a rule for every relationship given pairs of whole numbers
A6	Reading data from tables, pictographs, bar graphs, and pie charts
A7	Comparing and understanding how to use information from data.
A8	Understanding different representations and organizing data using tables, pictographs, and bar graphs

Source: Modified from Lee *et al.* (2011).

Posters for two sports clubs that rent bikes are shown below.



A. Use the information in the posters to complete the tables.

Mountain Bike rental	
Hours	Cost (zeds)
1	8
2	11
3	
4	
5	
6	

Roadrace Bike Rentals	
Hours	Cost (zeds)
1	10
2	12
3	
4	
5	
6	

B. For what number of hours are the rental costs the same at the two clubs?

Answer: _____

Figure 2. Items M031242A and M031242B from the TIMSS 2007 assessment.

Although it is possible for students to solve item 7b independently, it is more straightforward and thus more likely for them to obtain the answer directly by reading from the tables completed in item 7a. A further examination of students' responses showed that only three out of 823 students answered item 7b correctly, but not item 7a. This suggests that we can consider the two items as a single polytomous item to handle the testlet effect, and at the same time, to allow for answering item 7a successfully as a

prerequisite to answering item 7b correctly for most, if not all, students. After removing the responses of three students who answer item 7b correctly, but not item 7a, and combining items 7a and 7b as a single polytomous item, the responses of 820 students to 11 items were analysed.

The sequential G-DINA model was fitted to all items. Note that for dichotomously scored items, the sequential G-DINA model is equivalent to the G-DINA model. We used a and b to denote category 1 and 2, respectively, for polytomously scored items 3 and 9. For example, 9a and 9b represent the first and second categories of item 9, respectively. To fit the model, the q-vector for each category of items 3, 7 and 9 need to be derived from the original item-level q-vector developed by Lee *et al.* (2011). Item 3 requires students to complete a bar graph by drawing two bars based on the information in a table. Students can get a score of 1 if only one bar is drawn correctly, and 2 if both bars are drawn correctly (Foy & Olson, 2009, p. 85). Therefore, it is clear that category 3a is a prerequisite for 3b. Because both categories require the same operation, they measure the same set of attributes, as in A1, A6 and A8. When viewed as an independent item, item 7b requires three attributes (i.e., A2, A3 and A7; Lee *et al.*, 2011); however, when items 7a and 7b are assumed to be attained sequentially, only A7 (i.e., comparing and understanding how to use information from data) is necessary for item 7b. Finally, item 9 requires students to solve a problem in a real-life context. Students get a score of 1 if they provide a correct problem-solving procedure, but with computational errors, or if they provide the correct final solution without showing their work. If both the solution and procedure are correctly presented, students get a score of 2 (Foy & Olson, 2009, p. 98). Although this scoring rubric implies that category 9a is a prerequisite for 9b, it is not clear that which attributes are involved in each category. In particular, it is difficult to determine which attributes are used if only the final solution is provided without showing their work, and which attributes are involved when computational errors occur. Therefore, we used the unrestricted Q_C-matrix for item 9, and assumed that both categories required A2, A3 and A4. The Q_C-matrix is given in Table 8.

Based on the Q_C-matrix, there were $\sum_{j=1}^{11}\sum_{b=1}^{H_j} 2^{K^*_{jb}} = 102$ item parameters and $2^8 - 1 = 255$ latent class parameters. $S(h|\boldsymbol{\alpha}^*_{ljb})$ was constrained to be equal to or greater than $S(h|\boldsymbol{\alpha}^*_{l'jb})$ whenever $\boldsymbol{\alpha}^*_{ljb} \succ \boldsymbol{\alpha}^*_{l'jb}$, similar to de la Torre (2011). Also, the lower and upper bounds for processing functions were set to .001 and .999, respectively. The MMLE/EM algorithm described in the previous section with a convergence criterion of .001 was used for this analysis. It should be noted that, due to the relatively small number of examinees and items, large number of attributes, and possible misspecifications in the Q_C-matrix, the results should be interpreted with caution.

5.2. Results

Table 9 shows the estimated processing functions of the 11 items for each of the specific reduced attribute patterns. There are $2^{K^*_{jb}}$ processing functions for category b of item j associated with the reduced attribute patterns given at the top of the table. It should be noted that the same reduced attribute pattern for different items may not represent the same set of attributes. For example, items 5 and 6 each have four reduced attribute patterns, but they refer to different attributes (i.e., A2 and A3 for item 5, and A2 and A4 for item 6).

For item 7, students who have mastered all the required attributes for category 1 (i.e., A2, A3 and A5) have a 95.5% chance of answering this category correctly; however, those who lack the required attributes have only 8.8% chance of being correct. After completing

Table 8. Q_C-matrix for TIMSS 2007 data

			Attributes							
Item	TIMSS item no.	Category	A1	A2	A3	A4	A5	A6	A7	A8
1	M041052	1	1	1	0	0	0	0	0	0
2	M041281	1	0	1	1	0	1	0	0	0
3a	M041275	1	1	0	0	0	0	1	0	1
3b	M041275	2	1	0	0	0	0	1	0	1
4	M031303	1	0	1	1	0	0	0	0	0
5	M031309	1	0	1	1	0	0	0	0	0
6	M031245	1	0	1	0	1	0	0	0	0
7a	M031242A	1	0	1	1	0	1	0	0	0
7b	M031242B	2	0	0	0	0	0	0	1	0
8	M031242C	1	0	1	1	0	1	0	1	0
9a	M031247	1	0	1	1	1	0	0	0	0
9b	M031247	2	0	1	1	1	0	0	0	0
10	M031173	1	0	1	1	0	0	0	0	0
11	M031172	1	1	1	0	0	0	1	0	1

Notes. Polytomous items are shown in bold. This Q_C-matrix is modified from Lee *et al.* (2011).

category 1 correctly, students who have mastered A7 and those who have yet to master the attribute have a 99.9% and 0.1% chance of being correct on category 2, respectively. Furthermore, students' responses to category 1 do not appear to satisfy the conjunctive assumption that lacking one of the required attributes produces identical processing functions. For example, the processing function for students who have mastered A5 but not A2 and A3 is approximately twice as high as that for students who have mastered A3 only, both of which are much higher than that for students who have not mastered any required attribute. It can also be noted that the conjunctive assumption may not hold well for most categories requiring two or more attributes. Fitting the DINA model to the data in this way, as in Hansen (2013) and Lee *et al.* (2011), might be an oversimplification.

A close scrutiny of the processing functions reveals that for most categories, students who mastered all the required attributes have very high probabilities of success (e.g., the processing functions are >.95 for 12 out of 14 categories); whereas those who mastered none of the required attributes have low probabilities of success (e.g., the processing functions are <.15 for 9 out of 14 categories). Similarly to de la Torre (2008), $S_j(h|\mathbf{1}) - S_j(h|\mathbf{0})$ can be defined as a category discrimination index for category h of item j. For 14 categories of 11 items in this data, the discrimination indices range from .488 to .998, with mean .79. This means that, overall, most categories can be considered very discriminating.

In addition, compared with fitting the NRDM to item 9, fitting the US-GDINA model provided more information about response categories. For instance, mastering A4 contributed considerably to the processing functions for 9b, but not for 9a, which implies a possible misspecification in the q-vector of 9a because A4 does not seem to be necessary for this category. Similar findings can be observed for items 1 and 6. For example, A2 has a trivial contribution to the success probability for item 6, and therefore may not be necessary. These, however, need to be investigated further.

Finally, although the processing functions can be interpreted in a straightforward manner as above, the category response functions can also be derived easily and

Table 9. Estimates of processing functions for TIMSS 2007 data analysis

		Attribute pattern															
		0	1														
		00	10	01	11												
		000	100	010	001	110	101	011	111								
Item	Category	0000	1000	0100	0010	0001	1100	1010	1001	0110	0101	0011	1110	1101	1011	0111	1111
1	1	.511	.941	.511	.999												
2	1	.260	.908	.444	.487	.908	.908	.617	.999								
3a	1	.002	.732	.999	.999	.999	.999	.999	.999								
3b	2	.013	.013	.540	.535	.999	.999	.999	.999								
4	1	.452	.868	.781	.973												
5	1	.122	.865	.305	.961												
6	1	.001	.001	.914	.999												
7a	1	.088	.621	.427	.854	.770	.899	.891	.955								
7b	2	.001	.999														
8	1	.257	.257	.377	.646	.347	.377	.999	.347	.988	.457	.646	.999	.718	.999	.999	.999
9a	1	.072	.145	.452	.072	.646	.145	.470	.646								
9b	2	.001	.645	.532	.532	.715	.762	.532	.762								
10	1	.095	.700	.761	.999												
11	1	.355	.835	.835	.355	.355	.835	.835	.835	.835	.835	.415	.835	.835	.835	.999	.999

Note. Polytomous items are shown in bold.

interpreted accordingly. For example, students who mastered all the required attributes for item 3 have a 99.8% chance of getting a score of 2; whereas those who only mastered A1 have a 72.2% chance of getting a score of 1, but only a 1% chance of getting a score of 2.

6. Summary and discussion

In this paper we have developed a new polytomous CDM for graded responses, the sequential G-DINA model. Unlike other existing polytomous CDMs, by taking the attribute and category association into account, the sequential G-DINA model is able to model different cognitive processes for different response categories when these categories are completed in a sequential manner. Although initially developed for graded responses, the sequential G-DINA model is also suitable for unordered categorical responses when the unrestricted Q_C-matrix is used. The simulation study shows that the proposed estimation algorithm can produce accurate item and person parameter recovery.

The RS-GDINA model and the US-GDINA model were distinguished in this paper to account for possible uncertainty in attribute and category association. The LRT, AIC, and BIC were used to compare the RS-GDINA and US-GDINA models empirically. Based on the simulation study, selected models based on AIC can produce almost optimal person classifications in all simulated conditions. The LRT and BIC yielded worse results in some conditions.

The development of the sequential G-DINA model has important practical implications in that it opens up the possibility of relating response categories to attributes of interest. In particular, when writing polytomous items for cognitive diagnostic assessment, item writers may consider whether it is possible to link categories with attributes. In doing so, more diagnostic information may be extracted, which in turn can lead to more accurate person classifications.

Despite promising results, only the psychometric framework has been developed in this study. This offers researchers and practitioners a flexible tool to analyse polytomous items, but it is only a beginning of exploiting the value of polytomous items in cognitive diagnostic assessment. Additional research along these lines is needed. For example, this study only used one Q_C-matrix, and in future studies researchers could consider various Q_C-matrices to examine the impact of the Q_C-matrix on parameter recovery. Additionally, like most other CDMs, the sequential G-DINA model is a single-strategy model assuming that all examinees use the same strategy, which, nevertheless, may not necessary be the case in practice. For example, to solve $4\frac{1}{8} - \frac{3}{8} = ?$, another strategy is to convert the mixed number to an improper fraction before further operations. Multiple-strategy issues have been considered in the context of dichotomous response data. For example, Mislevy (1996) considered a mixture model for estimating the strategy being used. De la Torre and Douglas (2008) developed a multiple-strategy DINA model which allows students to use different strategies for different items. A major difference for graded response data stemming from constructed-response items is that the strategy used by each student for each item is probably observable if students show their work explicitly, and therefore estimating the strategy being used is not needed. It would be straightforward to incorporate multiple Q_C-matrices into the sequential G-DINA model, in conjunction with indicator variables showing the strategies being used by each student for each item.

Although the sequential G-DINA model does not make any assumption about the attribute structures, if attributes are structured, it seems intuitively more reasonable to assess lower-level attributes in lower-level categories. How the attribute structures can be

incorporated in the sequential G-DINA model would be an interesting topic to examine in the future. Also, in the simulation studies, five single-attribute items were included to ensure the Q_C-matrix is complete (Chiu, Douglas, & Li, 2009). However, it would be worthwhile to investigate further whether the completeness in the sequential G-DINA model can be achieved using single-attribute specifications at the category rather than item level. Furthermore, although the authors have noted the relationship between the sequential G-DINA model and the NRDM (Templin *et al.*, 2008) and PC-DINA model (de la Torre, 2010), it is still not clear how the proposed model relates to other polytomous CDMs, such as the pGDM (von Davier, 2008) and polytomous LCDM (Hansen, 2013). This needs further investigation. Finally, because of the sequential mechanism underlying the proposed model, it is appropriate for items with analytic scoring rubrics. At this point, it is not clear if the sequential G-DINA model is applicable to items that are scored using holistic rubrics.

Acknowledgements

The authors would like to thank the editor and the anonymous reviewers for insightful comments and valuable suggestions. This research was partially supported by National Science Foundation Grant DRL-0744486.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. doi:10.1109/TAC.1974.1100705

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats – It does make a difference for diagnostic purposes. *Applied Psychological Measurement*, *11*, 385–395. doi:10.1177/014662168701100404

Birenbaum, M., Tatsuoka, K. K., & Gutvirtz, Y. (1992). Effects of response format on diagnostic assessment of scholastic achievement. *Applied Psychological Measurement*, *16*, 353–363. doi:10.1177/014662169201600406

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459. doi:10.1007/BF02293801

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633–665. doi:10.1007/S11336-009-9125-0

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343–362. doi:10.1111/j.1745-3984.2008.00069.x

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. doi:10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*, 595–624. doi:10.1007/s11336-008-9063-2

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

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

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). A review of cognitively diagnostic assessment and a summary of psychometric models. In R. Rao, & S. Sinharay, *Handbook of statistics: Vol. 26. Psychometrics* (pp. 979–1030). Amsterdam, the Netherlands: Elsevier.

Foy, P., & Olson, J. F. (2009). *TIMSS 2007 user guide for the international database*. Chestnut Hill, MA: International Study Center, Boston College.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321. doi:10.1111/j.1745-3984.1989.tb00336.x

Hansen, M. (2013). *Hierarchical item response models for cognitive diagnosis*. Unpublished doctoral dissertation. University of California at Los Angeles.

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

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. doi:10.1007/s11336-008-9089-5

Johnson, M., Lee, Y.-S., Sachdeva, R. J., Zhang, J., Waldman, M., & Park, J. Y. (2013, April). *Examination of gender differences using the multiple groups DINA model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59–81. doi:10.1111/j.1745-3984.2011.00160.x

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*, 144–177. doi:10.1080/15305058.2010.534571

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187–212. doi:10.1007/BF02294535

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*, 379–416. doi:10.1111/j.1745-3984.1996.tb00498.x

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*, 308–313. doi:10.1093/comjnl/7.4.308

R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

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. doi:10.1080/15366360802490866

Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika*, *60*, 549–572. doi:10.1007/BF02294328

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. doi:10.1214/aos/1176344136

Su, Y.-L. (2013). *Cognitive diagnostic analysis using hierarchically structured skills*. Unpublished doctoral dissertation. University of Iowa.

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

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N., Fredrickson, R. L., Glaser, A. M., Lesgold & M. G., Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

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

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.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307. doi:10.1348/000711007X193957

Received 4 December 2015; revised version received 20 April 2016

Supporting Information

The following supporting information may be found in the online edition of the article:

Appendix S1. Relations between the sequential G-DINA model and existing polytomous CDMs.