the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2020), 73, 142–163
© 2019 The British Psychological Society

www.wileyonlinelibrary.com

An empirical Q-matrix validation method for the sequential generalized DINA model

Wenchao Ma[1]* and Jimmy de la Torre[2]

[1]Department of Educational Studies in Psychology, Research Methodology and Counseling, University of Alabama, Tuscaloosa, Alabama, USA
[2]Faculty of Education, University of Hong Kong, Hong Kong

As a core component of most cognitive diagnosis models, the Q-matrix, or item and attribute association matrix, is typically developed by domain experts, and tends to be subjective. It is critical to validate the Q-matrix empirically because a misspecified Q-matrix could result in erroneous attribute estimation. Most existing Q-matrix validation procedures are developed for dichotomous responses. However, in this paper, we propose a method to empirically detect and correct the misspecifications in the Q-matrix for graded response data based on the sequential generalized deterministic inputs, noisy 'and' gate (G-DINA) model. The proposed Q-matrix validation procedure is implemented in a stepwise manner based on the Wald test and an effect size measure. The feasibility of the proposed method is examined using simulation studies. Also, a set of data from the Trends in International Mathematics and Science Study (TIMSS) 2011 mathematics assessment is analysed for illustration.

1. Introduction

Cognitive diagnosis models (CDMs) refer to a set of psychometric models that intend to group individuals into latent classes with distinct skill profiles. A more generic term for skills is 'attributes', which are typically, although not always, assumed to be binary latent variables. An attribute profile indicates which attributes individuals have possessed and which they have not. In educational contexts, CDM analyses could provide diagnostic information about a student's strengths and weaknesses on a set of fine-grained skills to facilitate classroom instruction and learning. CDMs have also shown their potential for application in other fields such as psychological disorder diagnosis and personnel selection (e.g., Sorrel *et al.*, 2016; Templin & Henson, 2006; de la Torre, van der Ark, & Rossi, 2018).

A number of CDMs have been proposed in the literature. Examples include the deterministic inputs, noisy 'and' gate (DINA; Haertel, 1989) model, the deterministic inputs, noisy 'or' gate (DINO; Templin & Henson, 2006) model and the generalized DINA (G-DINA; de la Torre, 2011) model for dichotomous response, and the sequential G-DINA model (Ma & de la Torre, 2016) and the general diagnostic model (von Davier, 2008) for

*Correspondence should be addressed to Wenchao Ma, Department of Educational Studies in Psychology, Research Methodology and Counseling, The University of Alabama, Box 870231, Room 307B, 520 Colonial Drive, Tuscaloosa 35487, AL, USA (email: wenchao.ma@ua.edu).

DOI:10.1111/bmsp.12156

polytomous responses. Regardless of their parametrizations, most CDMs rely on a Q-matrix (Tatsuoka, 1983), which specifies whether an attribute is measured by an item. The importance of the Q-matrix in CDM analyses cannot be overemphasized. It has been widely recognized that a misspecified Q-matrix can degrade item parameter estimation, produce poor model-data fit, and result in erroneous attribute estimation (e.g., Chiu, 2013; Rupp & Templin, 2008).

Some studies have explored how to estimate Q-matrix from data directly without the need for a provisional Q-matrix (Chen, Culpepper, Chen, & Douglas, 2018; Chen, Liu, Xu, & Ying, 2015; Liu, Xu, & Ying, 2012, 2013). For example, Chen *et al.* (2015) proposed to use a regularized method to estimate Q-matrix under various CDMs. However, its performance can be affected by the number of attributes and the complexity of the underlying CDMs. In practice, it is common that a provisional Q-matrix is created by experts during the phase of test development. Although experts might be subjective and some entries in this provisional Q-matrix might not be correct, it can still serve as a useful starting point. At present, researchers have developed various approaches to refining the Q-matrix (Chen, 2017; Chiu, 2013; DeCarlo, 2012; Gu, Liu, Xu, & Ying, 2018; de la Torre, 2008; de la Torre & Chiu, 2016; Wang *et al.*, 2018). For example, in the approach of DeCarlo (2012), some entries are treated as random variables, and estimated along with all other parameters using the Markov chain Monte Carlo (MCMC) method. This method was developed for the DINA model and requires that potential misspecified entries in the Q-matrix be identified *a priori*. In addition, de la Torre and Chiu (2016) proposed a Q-matrix validation procedure that can be used for the G-DINA model and all models subsumed by the G-DINA model. Despite its flexibility, a cut-off needs to be specified in advance.

Although a number of Q-matrix validation procedures are available, none of them is developed for CDMs for ordinal responses. This can impede the use of constructed-response items in diagnostic assessments because constructed-response items typically produce ordinal scores. With this study, we attempt to fill this gap by developing a Q-matrix validation procedure for the sequential G-DINA model (Ma & de la Torre, 2016) for polytomously scored items that can be decomposed into a set of tasks and are scored sequentially. Items of this type are common in educational tests. For example, Masters (1982) identified three sequential steps for $\sqrt{7.5/0.3-16}$, but Tutz (1997) argued that, for items with sequential steps, item response models belonging to the class of the continuation ratio model (CRM; Agresti, 2013) are more interpretable than the partial credit model (Masters, 1982). The sequential G-DINA model is a member of the class of the CRM, and thus it is naturally suitable for items of this type. In addition, unlike other CDMs for polytomous responses, the sequential G-DINA model can account for the fact that different attributes might be involved in different tasks, and thus it has the potential to provide more accurate estimation of students' attribute profiles. The proposed Q-matrix validation procedure attempts to determine which attributes are measured for each task. For dichotomous items, the sequential G-DINA model is equivalent to the G-DINA model, and therefore the proposed method can also be used for dichotomous responses as long as the underlying CDM is subsumed by the G-DINA model. It has been recognized that statistical procedures for empirically validating the Q-matrix, including the method developed in this paper, should not be used with the intention of replacing the domain experts. Instead, the validation results should be used to provide ancillary information for domain experts.

The remainder of this paper is laid out as follows. In Section 2, we provide an overview of the sequential G-DINA model. In Section 3, we introduce a category level discrimination index, which is a straightforward extension of the G-DINA discrimination index (GDI; de la Torre & Chiu, 2016). In Section 4, we describe how the Q-matrix can be validated using the Wald test, and in Section 5 we introduce in detail a Q-matrix validation algorithm using the Wald test and the discrimination index. Two simulation studies are presented in Section 6, followed by the analysis of a set of data from Trends in International Mathematics and Science Study (TIMSS) 2011 mathematics assessment. We conclude in the last section with a brief summary of this study, and a discussion of directions for future research.

2. Overview of the sequential G-DINA model

The sequential G-DINA model is suitable for items that consist of a series of tasks and are scored according to how many tasks have been undertaken successfully. An item of this type for measuring proportion reasoning skills (e.g., Tjoe & de la Torre, 2014) is given as follows.

> Nate and Dale are making s'mores. Nate has four marshmallows and three crackers. Dale has seven marshmallows and five crackers. Whose s'mores have a stronger marshmallow taste (greater marshmallows-to-crackers ratio)?

To answer this item, the first task is to find the marshmallows-to-crackers ratios for Nate and Dale, and the second task is to compare the two ratios. Students are given a score of 0 if they fail the first task, a score of 1 if they perform the first task successfully but fail the second, and a score of 2 if they perform both tasks successfully. Because there is a one-to-one mapping between tasks and response categories, they are used interchangeably in this paper. More formally, suppose K attributes are involved in a test with J items, and to answer item j, H_j tasks need to be performed sequentially. Also, different tasks can measure different attributes. A binary q-vector of length K, $\boldsymbol{q}_{jh}$, is associated with category h of item j, and its kth element $q_{jhk} = 1$ if attribute k is required by task h of item j, and $q_{jhk} = 0$ otherwise. A collection of all q-vectors produces a category level Q-matrix (Ma & de la Torre, 2016), with the dimensions of $\sum_{j=1}^{J} H_j \times K$.

The K binary attributes lead to 2^K latent classes with unique attribute patterns (i.e., $\boldsymbol{\alpha}_c = (\alpha_{c1},\ldots, \alpha_{cK})$), where $c = 1,\ldots, 2^K$. Element $\alpha_{ck} = 1$ if attribute k is mastered by individuals in latent class c, and $\alpha_{ck} = 0$ if attribute k is not mastered by individuals in the same latent class. The probability of individual i with attribute pattern $\boldsymbol{\alpha}_c$ performing task h correctly provided that she or he has already completed task $h-1$ successfully is referred to as the processing function (Samejima, 1997) and it is expressed as

$$s_{jh}(\boldsymbol{\alpha}_c) = P(Y_{ij} \geq h | Y_{ij} \geq h - 1, \boldsymbol{\alpha}_c) = \frac{P(Y_{ij} \geq h | \boldsymbol{\alpha}_c)}{P(Y_{ij} \geq h - 1 | \boldsymbol{\alpha}_c)}. \tag{1}$$

The conditional probability of obtaining a score of h on item j can be written as

$$P(Y_{ij} = h|\boldsymbol{\alpha}_c) = [1 - s_{j,h+1}(\boldsymbol{\alpha}_c)] \prod_{y=0}^{h} s_{jy}(\boldsymbol{\alpha}_c), \tag{2}$$

where $s_{j0}(\boldsymbol{\alpha}_c) \equiv 1$ and $s_{j,\ H_j\ +\ 1}(\boldsymbol{\alpha}_c) \equiv 0$. Note that this model is referred to as the sequential process model (Ma & de la Torre, 2016), although different names have been used to refer to models of this type in different contexts, such as the CRM (Agresti, 2013; Mellenbergh, 1995), the sequential model (Tutz, 1997) and the step model (Verhelst, Glas, & de Vries, 1997).

The processing function can be defined using any dichotomous CDM, and the sequential G-DINA model is obtained when the form of the G-DINA model (de la Torre, 2011) is used. Specifically, for task h of item j, 2^K latent classes can be collapsed into $2^{K^*_{jh}}$ latent groups, where K^*_{jh} is the number of required attributes for this task. Let $\boldsymbol{\alpha}^*_{ljh}$ be the reduced attribute pattern for task h of item j consisting of the required attributes for this task only, where $l = 1, \ldots, 2^{K^*_{jh}}$. Assuming the first K^*_{jh} attributes are required, we have $\boldsymbol{\alpha}^*_{ljh} = (\alpha_{ljh1}, \ldots, \alpha_{ljhk}, \ldots, \alpha_{ljhK^*_{jh}})$. More formally, like Ma and de la Torre (2019), let $\boldsymbol{\alpha}^*_{ljh} = M_{jh}(\boldsymbol{\alpha}_c)$ if latent class with attribute pattern $\boldsymbol{\alpha}_c$ is collapsed into latent group with reduced attribute pattern $\boldsymbol{\alpha}^*_{ljh}$, and denote $C_{ljh} = \{\boldsymbol{\alpha}_c : M_{jh}(\boldsymbol{\alpha}_c) = \boldsymbol{\alpha}^*_{ljh}\}$. If $\boldsymbol{\alpha}_c \in C_{ljh}$, we have $s_{jh}(\boldsymbol{\alpha}_c) = s(\boldsymbol{\alpha}^*_{ljh})$. The processing function for the sequential G-DINA model is given by

$$s(\boldsymbol{\alpha}^*_{ljh}) = \phi_{jh0} + \sum_{k=1}^{K^*_{jh}} \phi_{jhk}\alpha_{ljhk} + \sum_{k'=k+1}^{K^*_{jh}} \sum_{k=1}^{K^*_{jh}-1} \phi_{jhkk'}\alpha_{ljhk}\alpha_{ljhk'} + \ldots + \phi_{jh12\ldots K^*_{jh}} \prod_{k=1}^{K^*_{jh}} \alpha_{ljhk}, \tag{3}$$

where ϕ_{jh0} is the intercept, ϕ_{jhk} is the main effect due to attribute k, $\phi_{jhkk'}$ is the two-way interaction effect due to attributes k and k', and $\phi_{jh12\ldots K^*_{jh}}$ is the K^*_{jh}-way interaction effect due to all required attributes. When all interaction effects are set to zeros, this is equivalent to the additive-CDM (*A*-CDM; de la Torre, 2011). Note that the subscripts of the processing function are dropped for simplicity when used along with the reduced attribute patterns.

3. Category-level G-DINA discrimination index

The GDI was originally proposed by de la Torre and Chiu (2016) for empirically validating the Q-matrix in conjunction with the G-DINA model for dichotomous responses. It can be extended for the sequential G-DINA model and defined in a straightforward manner for each non-zero category of a polytomously scored item. Specifically, the posterior probability of an individual i having attribute pattern $\boldsymbol{\alpha}_c$ given the item response vector $\boldsymbol{Y}_i$ can be calculated by

$$P(\boldsymbol{\alpha}_c|\boldsymbol{Y}_i) = \frac{P(\boldsymbol{Y}_i|\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c)}{\sum\limits_c P(\boldsymbol{Y}_i|\boldsymbol{\alpha}_c)p(\boldsymbol{\alpha}_c)}. \tag{4}$$

For category h of item j, the category level GDI is formulated as

$$\varsigma^2_{jh} = \sum_{l=1}^{2^{K^*_{jh}}} p(\boldsymbol{\alpha}^*_{ljh})[s(\boldsymbol{\alpha}^*_{ljh}) - \bar{s}_{jh}]^2, \tag{5}$$

where

$$p(\boldsymbol{\alpha}^*_{ljh}) = \sum_{i=1}^{N} \sum_{\boldsymbol{\alpha}_c \in C_{ljh}} P(\boldsymbol{\alpha}_c | \boldsymbol{Y}_i) \tag{6}$$

and

$$\bar{s}_{jh} = \sum_{l=1}^{2^{K^*_{jh}}} p(\boldsymbol{\alpha}^*_{ljh}) s(\boldsymbol{\alpha}^*_{ljh}). \tag{7}$$

Note that ς^2_{jh} is the variance of success probabilities for category h of item j for all latent groups given $\boldsymbol{q}_{jh}$, and it measures a category's overall discriminating power. de la Torre and Chiu (2016) have shown that, theoretically, when the correct Q-matrix is used and models fit the data perfectly, the correct q-vector and overspecified q-vectors from the correct one produce the largest GDI. Hence, the q-vector with the largest GDI, but requiring the fewest attributes, is the correct q-vector. In practice, however, overspecified q-vectors from the correct one have a larger GDI than the correct q-vector due to random errors. de la Torre and Chiu (2016) calculated the proportion of variance accounted for (PVAF) by a particular q-vector to the maximum due to the q-vector $\mathbf{1}$ for each item, and the q-vector with a PVAF greater than a certain pre-specified cut-off, but requiring fewest attributes, can be considered correct. This method is very flexible, given that it can be used without any assumption about the form of CDMs. The use of PVAF also provides a way of quantifying the discriminating power of each candidate q-vector. However, this method does not consider the item parameter estimation errors, and determining the cut-off for PVAF *a priori* could be challenging.

4. Attribute validation using the Wald test

The Wald test (Wald, 1943) is a widely used hypothesis test in statistics. In the context of CDMs, it has been used for comparing the G-DINA model and the reduced CDMs that the G-DINA model subsumes (Ma & de la Torre, 2019; Ma, Iaconangelo, & de la Torre, 2016; Sorrel, Abad, Olea, de la Torre, & Barrada, 2017; de la Torre, 2011), and detecting differential item functioning (Hou, de la Torre, & Nandakumar, 2014; Ma, Terzi, Lee, & de la Torre, 2017). In this section, we illustrate how the Wald test can be used to evaluate whether or not an attribute that is assumed to be required is statistically necessary in a q-vector involving two or more ones. Specifically, if changing an element one to zero in a q-vector does not lead to a significantly worse model-data fit, the attribute is said to be unnecessary statistically. This allows us to conduct the Q-matrix validation from a perspective of model comparison.

Again, suppose q_{jhk} is the kth element of the q-vector of category h of item j and the number of required attributes $K^*_{jh} = \sum_{k=1}^{K} q_{jhk} \geq 2$. To test whether an element one can be changed to zero, a $2^{K^*_{jh}-1} \times 2^{K^*_{jh}}$ restriction matrix $\boldsymbol{R}$ is needed for the Wald test so that under the null hypothesis, $\boldsymbol{R} \times \boldsymbol{s}_{jh} = \mathbf{0}$, where $\boldsymbol{s}_{jh}$ is a vector of processing functions for category h of item j. For example, assume $K^*_{jh} = 3$ and $\boldsymbol{q}_{jh} = (1, 1, 1, \ldots)$. To test whether Attribute 1 is required statistically, the null hypothesis is

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \times \begin{bmatrix} s(000) \\ s(100) \\ s(010) \\ s(001) \\ s(110) \\ s(101) \\ s(011) \\ s(111) \end{bmatrix} = \mathbf{0}. \tag{8}$$

The restriction matrices for testing the necessity of Attributes 2 and 3 are

$$\begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}, \tag{9}$$

and

$$\begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}, \tag{10}$$

respectively. The Wald statistic is defined as

$$W = [\boldsymbol{R} \times \boldsymbol{s}_{jh}]'[\boldsymbol{R} \times \boldsymbol{V}_{jh} \times \boldsymbol{R}']^{-1}[\boldsymbol{R} \times \boldsymbol{s}_{jh}]. \tag{11}$$

Here, $\boldsymbol{V}_{jh}$ is a $2^{K_{jh}^*} \times 2^{K_{jh}^*}$ submatrix of the covariance matrix $\boldsymbol{V}(\boldsymbol{s}) = \mathcal{I}^{-1}(\boldsymbol{s})$. The information matrix $\mathcal{I}(\boldsymbol{s})$ is approximated using the outer product of gradient approach as in Ma and de la Torre (2018) with element:

$$\sum_{i=1}^{N} \left[\frac{\partial \log L(\boldsymbol{Y}_i)}{\partial s(\boldsymbol{\alpha}_{ljh}^*)} \frac{\partial \log L(\boldsymbol{Y}_i)}{\partial s(\boldsymbol{\alpha}_{l'j'h'}^*)} \right], \tag{12}$$

where $L(\boldsymbol{Y}_i)$ is the observed marginalized likelihood and

$$\frac{\partial \log L(\boldsymbol{Y}_i)}{\partial s(\boldsymbol{\alpha}_{ljh}^*)} = \sum_{\boldsymbol{\alpha}_c \in C_{ljh}} P(\boldsymbol{\alpha}_c | \boldsymbol{Y}_i) \left[\frac{I(\boldsymbol{Y}_{ij} \geq h)}{s(\boldsymbol{\alpha}_{ljh}^*)} - \frac{I(\boldsymbol{Y}_{ij} = h-1)}{s(\boldsymbol{\alpha}_{ljh}^*)} \right] \tag{13}$$

The Wald statistic W is asymptotically χ^2 distributed with $2^{K_{jh}^*-1}$ degrees of freedom.

5. A Q-matrix validation algorithm

We have shown in the previous section how the Wald test can be used to evaluate whether an attribute that is assumed to be necessary is statistically required or not in a q-vector involving two or more ones. Here, we describe a Q-matrix validation procedure for the sequential G-DINA model using the aforementioned discrimination index and Wald statistic. This procedure is implemented category by category and item by item. Specifically, the first required attribute is chosen based on the PVAF, whereas choosing the next required attributes, if any, is based on both the Wald test and the PVAF. The Wald test serves as a hypothesis test, and the PVAF functions as an effect size measure, which can be critical when more than one attribute is deemed necessary based on the Wald test. More specifically, for category h of item j, the algorithm is conducted as follows.

Step 1. Define $\Omega = \{1,\ldots, K\}$ as a set consisting of the indices for all K attributes. Also, let A be a set consisting of the indices for all the required attributes identified during the validation process, and $B = \Omega\backslash A$. The attributes indexed in set B are called target attributes in that their necessity needs to be examined. Initialize $A = \varnothing$, and thus $B = \{1,\ldots, K\}$. Define a q-vector search bank C consisting of K single-attribute competing q-vectors. Replace the provisional q-vector (i.e., $\boldsymbol{q}_{jh}$ in the Q-matrix) with each of the competing q-vectors in C, and calculate their associated PVAFs. The target attribute required by the competing q-vector producing the largest PVAF is defined as a required attribute. Assume this attribute is attribute k', and update sets A and B: $A = \{k'\}$ and $B = \Omega\backslash A$.

Step 2. Check whether the q-vector requiring the attributes indexed in set A has a PVAF >0.95, which is the same as the cut-off used in de la Torre and Chiu (2016). If yes, the validation process terminates; otherwise, update the search bank C so that each competing q-vector requires all attributes indexed in set A and one target attribute indexed in set B. As a result, there are at least two-ones in each competing q-vector in this step. For example, assume we have three attributes and in the first step, (0, 1, 0) had the largest PVAF compared with (1, 0, 0) and (0, 0, 1). Therefore, $A = \{2\}$, and $B = \{1, 3\}$. The competing q-vectors include (1, 1, 0) and (0, 1, 1), both of which require attribute 2 as it is indexed in set A. Each of the competing q-vectors also requires an target attribute (i.e., attributes 1 and 3 for the first and second competing q-vectors, respectively). The Wald test is used to examine whether or not the target attribute is statistically necessary for each competing q-vector. If none of the target attributes is required, the validation process terminates; if at least one target attribute is required, the one specified in the competing q-vector with the largest PVAF is assumed to be required, and the associated q-vector is the best among all current competing q-vectors. The index of the target attribute in this q-vector is added to set A and removed from set B. The necessity of the required attributes except the target attribute in this competing q-vector is examined using the Wald test as well. If any of them are deemed unnecessary statistically after the target attribute has been included, their indices are removed from set A to set B. Step 2 is repeated until no new index can be added to or removed from sets A and B. The flowchart for this validation procedure is given in Figure 1.

Steps 1 and 2 are implemented for each category of each item. Step 1 aims to determine the first required attribute using the PVAF, and Step 2 attempts to identify other required attributes, if any, using the Wald test, in conjunction with the PVAF when necessary. After Step 2 ends, all attributes indexed in set A are believed to be required for the studied

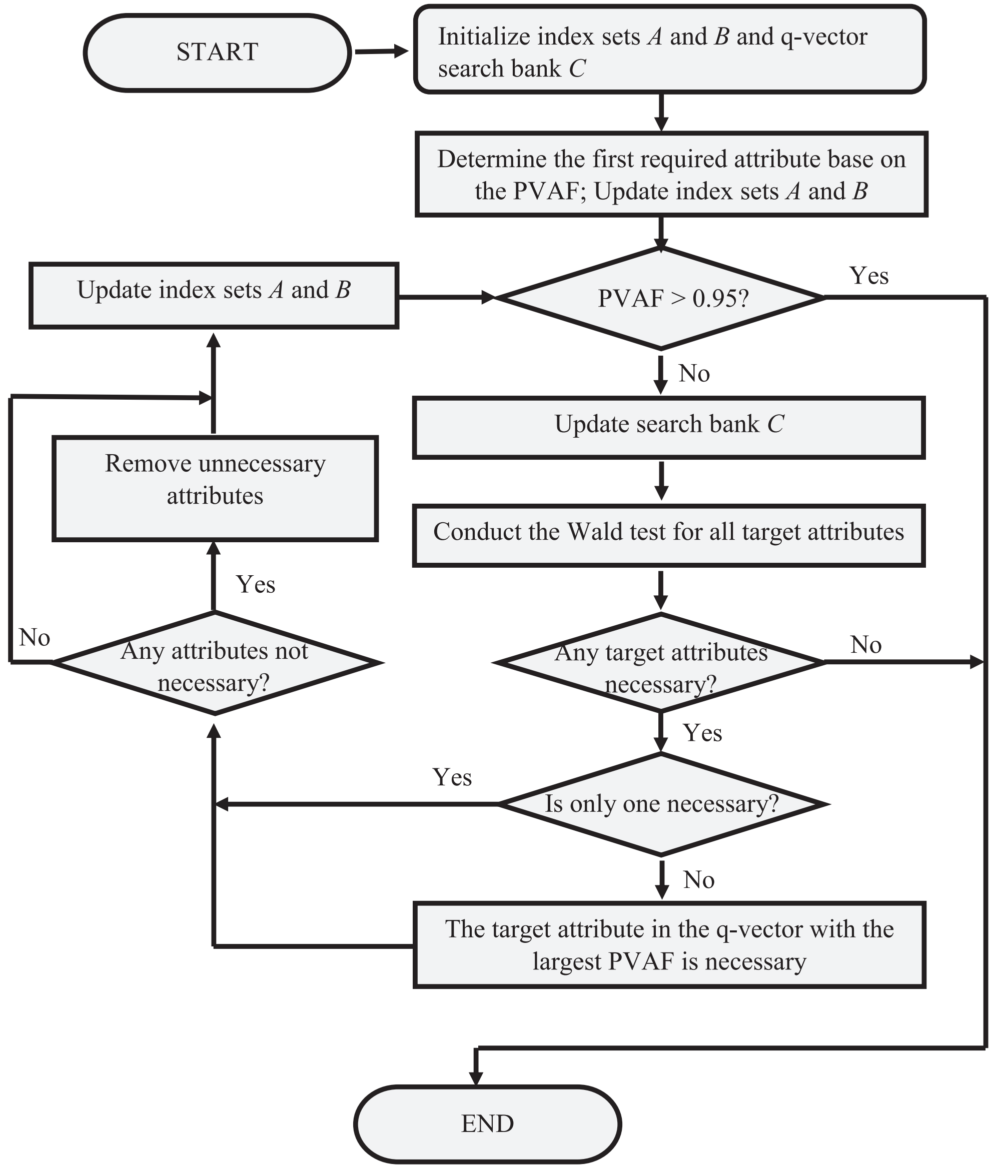

START
Initialize index sets *A* and *B* and q-vector search bank *C*
Determine the first required attribute base on the PVAF; Update index sets *A* and *B*
PVAF > 0.95?
Yes
No
Update index sets *A* and *B*
Update search bank *C*
Remove unnecessary attributes
Conduct the Wald test for all target attributes
Yes
No
Any attributes not necessary?
Any target attributes necessary?
No
Yes
Yes
Is only one necessary?
No
The target attribute in the q-vector with the largest PVAF is necessary
END

Figure 1. Flowchart of the stepwise Q-matrix validation.

category. This process is said to be implemented in a stepwise manner in that the necessity of the attributes is evaluated iteratively, similar to the stepwise procedure for model selection in linear regression (Efroymson, 1960). It should be noted that at the beginning of Step 2, the PVAF of the current *q*-vector is calculated and compared with 0.95. This evaluation is not mandatory, but it might be useful when sample size is large, in which condition, the hypothesis test tends to reject the null hypothesis and result in overspecified *q*-vectors.

In addition, the calculations of the GDI and the Wald statistics involve the estimation of the processing functions based on each competing *q*-vector for the studied category. It is straightforward to recalibrate the data based on each competing q-vector; however, this can be computationally intensive. An alternative solution, which is adopted in this study, is the expectation-maximization (EM) based approximation, similar to de la Torre (2008) and de la Torre and Chiu (2016). The sequential G-DINA model is fitted to the data based on the provisional (i.e., original) Q-matrix, and the posterior probability for individual i (i.e., $P(\boldsymbol{\alpha}_c|\boldsymbol{Y}_i)$) can be calculated using equation (4). Note that if the q-vector of category h of item j changes, the reduced attribute profile $\boldsymbol{\alpha}^*_{ljh}$ also changes. Therefore, let $\tilde{\boldsymbol{q}}_{jh}$ denote a competing q-vector of category h of item j and let $\tilde{\boldsymbol{\alpha}}^*_{ljh}$ be the corresponding reduced attribute profile. Also, let $\tilde{\boldsymbol{\alpha}}^*_{ljh} = \tilde{M}_{jh}(\boldsymbol{\alpha}_c)$ if latent class c is collapsed into latent group l, and denote $\tilde{C}_{ljh} = \{\boldsymbol{\alpha}_c : \tilde{M}_{jh}(\boldsymbol{\alpha}_c) = \tilde{\boldsymbol{\alpha}}^*_{ljh}\}$. The probability of individual i having a reduced attribute profile $\tilde{\boldsymbol{\alpha}}^*_{ljh}$ for category h of item j under the competing q-vector $\tilde{\boldsymbol{q}}_{jh}$ can be approximated by

$$P(\tilde{\boldsymbol{\alpha}}^*_{ljh}|\boldsymbol{Y}_i) \approx \sum_{\boldsymbol{\alpha}_c \in \tilde{C}_{ljh}} P(\boldsymbol{\alpha}_c|\boldsymbol{Y}_i), \tag{14}$$

and the processing functions can be estimated by

$$\hat{s}(\tilde{\boldsymbol{\alpha}}^*_{ljh}) = \frac{\sum_{i=1}^{N} P(\tilde{\boldsymbol{\alpha}}^*_{ljh}|\boldsymbol{Y}_i) I(Y_{ij} \geq h)}{\sum_{i=1}^{N} P(\tilde{\boldsymbol{\alpha}}^*_{ljh}|\boldsymbol{Y}_i) I(Y_{ij} \geq h-1)}. \tag{15}$$

With the estimated processing functions, the covariance matrix was estimated with the outer product of gradient approach using equation (13). It should be noted that although the approximation in equation (14) reduces the computational burden dramatically, its adequacy depends on how well the posterior distribution can be estimated based on the original Q-matrix.

6. Simulation studies

Two simulation studies were conducted to evaluate the performance of the proposed stepwise Q-matrix validation method. Specifically, simulation study 1 explored the performance of the proposed method when the processing functions conform to some reduced CDMs (i.e., the DINA, DINO and A-CDM models), whereas simulation study 2 specified the processing function as a more general form (i.e., the G-DINA model), and compared the proposed stepwise method with the ς^2 method of de la Torre and Chiu (2016). The factors considered in these two studies are summarized in Table 1.

6.1. Simulation study 1

6.1.1. Design

In this study, the number of items and attributes were fixed to $J = 23$ and $K = 5$, respectively. The sample sizes were $N = 1{,}000$, 2,000 and 4,000. Item quality had three levels: $s(\boldsymbol{\alpha}^*_{ljh} = 0) = 1 - \mathrm{s}(\boldsymbol{\alpha}^*_{ljh} = 1) = 0.1,\ 0.2$ or 0.3 for all categories of all items, representing high, moderate and low quality, respectively. When the processing function is the A-CDM, each required attribute contributed equally to the processing function.

Table 1. Summary of factors in simulation studies

Factors	Simulation study 1		Simulation study 2
N		1,000, 2,000, 4,000	
(J, K)		(23, 5)	
Misspecification generation		Random	
Percentage of misspecified entries (m)		0%, 10%, 20%	
Attribute distribution		Uniform, Higher-order	
Processing function	DINA/DINO/A-CDM		G-DINA
$[s_{jh}(\mathbf{0}), s_{jh}(\mathbf{1})]$	[0.1, 0.9]/[0.2, 0.8]/ [0.3, 0.7]		[U(0.1, 0.3), U(0.7, 0.9)]
Q-matrix validation method	Stepwise method		Stepwise method, ς^2 method

Individuals' attribute patterns were generated from two different distributions. In the uniform distribution, all possible attribute patterns are equally likely, whereas in the higher-order distribution (de la Torre & Douglas, 2004), the probability of mastering attribute k for individual i is defined as

$$P(\alpha_k = 1|\theta_i, \delta_k) = \frac{\exp(\theta_i - \delta_k)}{1 + \exp(\theta_i - \delta_k)}. \tag{16}$$

Here, θ_i represents the ability of examinee i and has been drawn from the standard normal distribution, and δ_k is the difficulty of attribute k, which was randomly drawn from one of the five equal intervals from -1.5 to 1.5.

Table 2 gives the Q-matrix for the simulation studies. Out of 23 items, there are five two-category items (i.e., categories 0 and 1), 12 three-category items and six four-category items. Misspecified Q-matrices were constructed by altering 10% or 20% entries (i.e., m = 10%, 20%) in the correct Q-matrix randomly from 0 to 1 or from 1 to 0 with the constraints that each non-zero category measured at least one attribute and that each attribute was required by at least one non-zero category. The processing functions used for data simulation were the DINA model, DINO model, and A-CDM. In each condition, 200 data sets were simulated. The GDINA R package (Ma & de la Torre, 2017) was used for data simulation and model estimation, and the stepwise Q-matrix validation was implemented in the R programming environment (R Core Team, 2017).

An initial recovery rate (IRR) is defined as the proportion of times the first required attributes determined by the stepwise procedure for all non-zero categories are indeed required relative to the total number of non-zero categories. An average IRR across all replications for each condition is used to evaluate the performance of the GDI in selecting the first required attribute. To examine the performance of the Q-matrix validation procedure, the true positive rate and true negative rate were calculated. The true positive rate is the percentage of misspecified entries that were correctly identified, and the true negative rate is the percentage of correct entries that were correctly retained.

Table 2. Q-matrix for simulation studies

Item	Category	α_1	α_2	α_3	α_4	α_5	Item	Category	α_1	α_2	α_3	α_4	α_5
1	1	1	0	0	0	0	13	1	1	0	0	0	0
1	2	0	1	0	0	0	13	2	0	1	0	0	0
2	1	0	0	1	0	0	13	3	0	0	1	0	0
2	2	0	0	0	1	0	14	1	0	0	0	1	0
3	1	0	0	0	0	1	14	2	0	0	1	0	0
3	2	1	0	0	0	0	14	3	0	1	0	0	0
4	1	0	1	0	0	0	15	1	0	0	0	0	1
4	2	0	0	1	0	0	15	2	0	0	1	0	0
5	1	0	0	1	1	0	15	3	1	0	0	1	0
5	2	0	0	0	0	1	16	1	0	0	0	1	0
6	1	1	1	0	0	0	16	2	1	0	0	0	0
6	2	0	0	0	1	0	16	3	0	0	1	0	1
7	1	0	0	1	1	1	17	1	1	0	0	0	0
7	2	0	1	0	0	0	17	2	0	1	0	0	0
8	1	1	1	0	1	0	17	3	0	1	0	1	1
8	2	0	0	0	0	1	18	1	0	1	0	0	0
9	1	0	0	0	0	1	18	2	0	0	0	1	0
9	2	1	0	1	0	0	18	3	1	0	1	0	1
10	1	0	1	0	0	0	19	1	1	0	0	0	0
10	2	1	0	0	0	1	20	1	0	1	0	0	0
11	1	0	0	0	0	1	21	1	0	0	1	0	0
11	2	0	1	1	1	0	22	1	0	0	0	1	0
12	1	0	0	1	0	0	23	1	0	0	0	0	1
12	2	1	0	0	1	1							

6.1.2. Results

Table 3 gives the IRRs when items were of low quality. When item quality was high or moderate, the IRRs were always >99%, which implies that the GDI has excellent performance in selecting the initial required attribute under these conditions, and so the results were omitted. When item quality was low, the IRRs were still very good with a

Table 3. Initial recovery rates for reduced models when item quality was low

Processing function	m	Uniform			Higher-order		
		N = 1,000	N = 2,000	N = 4,000	N = 1,000	N = 2,000	N = 4,000
DINA	0%	0.997	0.999	1.000	0.999	1.000	1.000
	10%	0.991	0.998	0.999	0.989	0.997	0.999
	20%	0.977	0.994	0.996	0.958	0.972	0.984
DINO	0%	0.998	1.000	1.000	0.998	1.000	1.000
	10%	0.993	0.998	0.999	0.992	0.998	0.999
	20%	0.976	0.989	0.995	0.963	0.982	0.987
A-CDM	0%	0.997	1.000	1.000	0.999	1.000	1.000
	10%	0.994	0.999	1.000	0.993	0.998	0.999
	20%	0.980	0.993	0.998	0.953	0.977	0.982

Note. Here, m is the percentage of misspecifications and N is the sample size.

minimum value of 95.3%, which occurred under $N = 1{,}000$, 20% misspecifications, higher-order attribute distribution and A-CDM processing function. From Table 3, the IRR increased as the sample size increased or the percentage of misspecification decreased. Regarding attribute distributions, similar IRRs were observed when there were 10% misspecifications or less, but lower IRRs were observed for the higher-order attribute distribution when there were 20% misspecifications. The IRRs were similar across different processing functions.

Table 4 gives the true positive rates across sample sizes, item qualities, attribute distributions, percentages of misspecifications and processing functions. Item quality influences the true positive rates. The average true positive rate was 98.8% with a minimum value of 97.4% when items were of high quality, and 96.2% with a minimum value of 91% when items were of moderate quality. When item quality was low, however, the average true positive rate dropped to 78.9% with a minimum value of 66.4%. The impact of sample sizes, attribute distributions and percentages of misspecifications was apparent when items were of moderate or low quality. Specifically, the true positive rate increased as the sample size increased, or the percentage of misspecifications decreased. Also, uniformly distributed attributes yielded higher true positive rates than higher-order attributes. When items were of high quality, however, the impact of other factors was not always consistent, which might be caused by a ceiling effect in that the range of true positive rate was only 1.9%.

Table 5 gives the true negative rates for the stepwise Q-matrix validation method across varied conditions. It can be observed that across all conditions, the validation method performed excellently. Even when item quality was low, the average true negative rate was 97.5% with the minimum values of 95.3%. The average true negative rate for high item quality conditions (i.e., 99.4%) is slightly higher than that for moderate item quality conditions (i.e., 99.3%). However, with large sample size and small percentage of misspecification, items of moderate quality could have slightly larger true negative rates than items of high quality. The true negative rates increased as sample size increased. In addition, the uniformly distributed attributes produced higher true negative rates than the higher-order attributes. There was no apparent difference in true negative rate among different percentages of misspecification and different processing functions.

6.2. Simulation study 2

6.2.1. Design

In simulation study 1, the performance of the stepwise Q-matrix validation method was examined when the processing functions were reduced CDMs. Simulation study 2 used the G-DINA model as the processing function, which relaxes the assumptions about the condensation rule for each category. A more realistic condition for item quality was also considered, where $s(\boldsymbol{\alpha}_{ljh}^{*} = \mathbf{0})$ and $s(\boldsymbol{\alpha}_{ljh}^{*} = \mathbf{1})$ were drawn from $U(0.1, 0.3)$ and $U(0.7, 0.9)$, respectively. When $K_{jh}^{*} > 1$, the processing functions for latent classes with $\boldsymbol{\alpha}_{ljh}^{*}$ not equal to $\mathbf{0}$ or $\mathbf{1}$ were drawn from the uniform distribution $U[s(\boldsymbol{\alpha}_{ljh}^{*} = \mathbf{0}), s(\boldsymbol{\alpha}_{ljh}^{*} = \mathbf{1})]$. The processing functions were simulated with the monotonic constraint that mastering an additional attribute would not produce a lower processing function. This study also compared the stepwise method with the ς^2 approach of de la Torre and Chiu (2016). In each condition, 200 data sets were generated. As in the previous study, the IRR was used to evaluate the performance of the GDI in identifying the initial required attribute for the stepwise method, and the true positive and true negative rates were used to compare the performance of the stepwise validation procedure and the ς^2 approach of de la Torre and Chiu (2016).

Table 4. True positive of the stepwise Q-matrix validation method for reduced models

			Uniform			Higher-order		
Processing function	*m*	Item quality	*N* = 1,000	*N* = 2,000	*N* = 4,000	*N* = 1,000	*N* = 2,000	*N* = 4,000
DINA	10%	High	0.993	0.993	0.996	0.991	0.994	0.996
		Moderate	0.969	0.984	0.992	0.952	0.970	0.986
		Low	0.799	0.843	0.897	0.772	0.806	0.846
	20%	High	0.991	0.991	0.997	0.983	0.986	0.990
		Moderate	0.953	0.975	0.984	0.931	0.937	0.962
		Low	0.746	0.792	0.857	0.692	0.704	0.737
DINO	10%	High	0.983	0.993	0.996	0.979	0.989	0.993
		Moderate	0.955	0.978	0.991	0.935	0.963	0.984
		Low	0.802	0.851	0.891	0.777	0.815	0.829
	20%	High	0.986	0.986	0.996	0.978	0.985	0.993
		Moderate	0.947	0.974	0.989	0.920	0.943	0.964
		Low	0.749	0.810	0.857	0.711	0.733	0.751
A-CDM	10%	High	0.987	0.991	0.994	0.985	0.994	0.993
		Moderate	0.947	0.974	0.989	0.936	0.961	0.981
		Low	0.783	0.839	0.895	0.762	0.777	0.826
	20%	High	0.989	0.993	0.992	0.981	0.988	0.990
		Moderate	0.936	0.969	0.984	0.911	0.945	0.969
		Low	0.735	0.792	0.854	0.664	0.704	0.718

Note. Here, *m* is the percentage of misspecifications and *N* is the sample size.

Table 5. True negative of the stepwise Q-matrix validation method for reduced models

Processing		Item	Uniform			Higher-order		
function	*m*	quality	N = 1,000	N = 2,000	N = 4,000	N = 1,000	N = 2,000	N = 4,000
DINA	0%	High	0.995	0.996	0.998	0.993	0.995	0.998
		Moderate	0.996	0.999	1.000	0.993	0.998	1.000
		Low	0.963	0.985	0.997	0.963	0.983	0.993
	10%	High	0.996	0.997	0.999	0.993	0.996	0.998
		Moderate	0.996	0.999	1.000	0.991	0.997	0.999
		Low	0.962	0.983	0.996	0.962	0.981	0.992
	20%	High	0.996	0.997	0.999	0.991	0.993	0.995
		Moderate	0.993	0.997	0.998	0.987	0.992	0.994
		Low	0.959	0.981	0.994	0.959	0.976	0.988
DINO	0%	High	0.989	0.994	0.998	0.985	0.990	0.996
		Moderate	0.989	0.995	0.999	0.984	0.993	0.997
		Low	0.962	0.981	0.992	0.959	0.973	0.987
	10%	High	0.990	0.994	0.998	0.985	0.991	0.996
		Moderate	0.989	0.994	0.999	0.981	0.992	0.997
		Low	0.962	0.979	0.991	0.957	0.974	0.985
	20%	High	0.990	0.994	0.999	0.984	0.990	0.995
		Moderate	0.986	0.993	0.998	0.977	0.988	0.994
		Low	0.955	0.974	0.988	0.954	0.970	0.980
A-CDM	0%	High	0.992	0.996	0.997	0.990	0.994	0.997
		Moderate	0.988	0.998	1.000	0.984	0.995	0.999
		Low	0.956	0.978	0.992	0.954	0.974	0.990
	10%	High	0.994	0.996	0.998	0.991	0.995	0.997
		Moderate	0.988	0.997	1.000	0.983	0.995	0.999
		Low	0.958	0.978	0.992	0.956	0.975	0.989
	20%	High	0.993	0.998	0.997	0.990	0.995	0.997
		Moderate	0.985	0.996	0.998	0.980	0.993	0.997
		Low	0.956	0.976	0.991	0.953	0.973	0.987

Note. Here, *m* is the percentage of misspecifications and *N* is the sample size.

6.2.2. Results

Table 6 gives the IRRs of the stepwise Q-matrix validation method across sample sizes, attribute distributions and percentages of misspecifications. The GDI had excellent power to identify the first required attribute with a minimum IRR of 99.4%. It can also be observed that the IRR increased as the sample size increased and the percentage of misspecification decreased.

Table 6. Initial recovery rate of the stepwise method for the G-DINA processing function

	Uniform			Higher-order		
m	N = 1,000	N = 2,000	N = 4,000	N = 1,000	N = 2,000	N = 4,000
0%	0.999	1.000	1.000	0.999	1.000	1.000
10%	0.998	1.000	1.000	0.998	0.999	1.000
20%	0.994	0.998	0.999	0.996	0.997	1.000

Note. Here, *m* is the percentage of misspecifications and *N* is the sample size.

True positive and true negative rates are given in Tables 7 and 8. The stepwise Q-matrix validation procedure performs well in correcting the misspecifications and retaining the correct q-entries in the Q-matrix. Specifically, across all conditions, the true positive and true negative rates were >0.9 and 0.96, respectively. In addition, both true positive and true negative rates increased as sample sizes increased or the percentage of misspecifications decreased. Compared with the higher-order attribute distribution, the true positive and true negative rates were higher under the uniform distribution.

The proposed stepwise procedure outperformed the ς^2 approach of de la Torre and Chiu (2016) across all conditions. Specifically, the differences in true positive rates ranged between 0.091 and 0.36 and the differences in true negative rates ranged between 0.055 and 0.267. The differences tended to be substantial when sample size was small, but less marked when $N = 4{,}000$.

It can be observed from Table 8 that the ς^2 approach had higher true negative rates as the proportion of misspecifications increased. To partly address this counterintuitive result, we can define an overall accuracy index by finding the proportion of elements that are correctly changed or retained. This can be calculated by $m \times$ true positive rate + $(1 - m) \times$ true negative rate from Tables 7 and 8 directly. With this index, we can see that the overall accuracy of the ς^2 approach decreased when the proportion of misspecifications increased from 10% to 20%.

Table 7. True positive rates of the stepwise and PVAF methods for the G-DINA processing function

		Uniform			Higher-order		
Method	m	$N = 1{,}000$	$N = 2{,}000$	$N = 4{,}000$	$N = 1{,}000$	$N = 2{,}000$	$N = 4{,}000$
Stepwise	10%	0.948	0.965	0.978	0.932	0.953	0.988
	20%	0.930	0.958	0.975	0.901	0.925	0.976
ς^2 approach	10%	0.593	0.778	0.887	0.583	0.750	0.897
	20%	0.570	0.760	0.883	0.544	0.722	0.879

Note. Here, m is the percentage of misspecifications and N is the sample size.

Table 8. True negative rates of the stepwise and PVAF methods for the G-DINA processing function

		Uniform			Higher-order		
Method	m	$N = 1{,}000$	$N = 2{,}000$	$N = 4{,}000$	$N = 1{,}000$	$N = 2{,}000$	$N = 4{,}000$
Stepwise	0%	0.979	0.988	0.993	0.969	0.979	0.989
	10%	0.977	0.987	0.993	0.972	0.983	0.988
	20%	0.973	0.985	0.992	0.964	0.977	0.985
ς^2 approach	0%	0.713	0.853	0.930	0.711	0.845	0.929
	10%	0.724	0.857	0.932	0.726	0.860	0.930
	20%	0.734	0.866	0.937	0.735	0.862	0.924

Note. Here, m is the percentage of misspecifications and N is the sample size.

7. Real data analysis

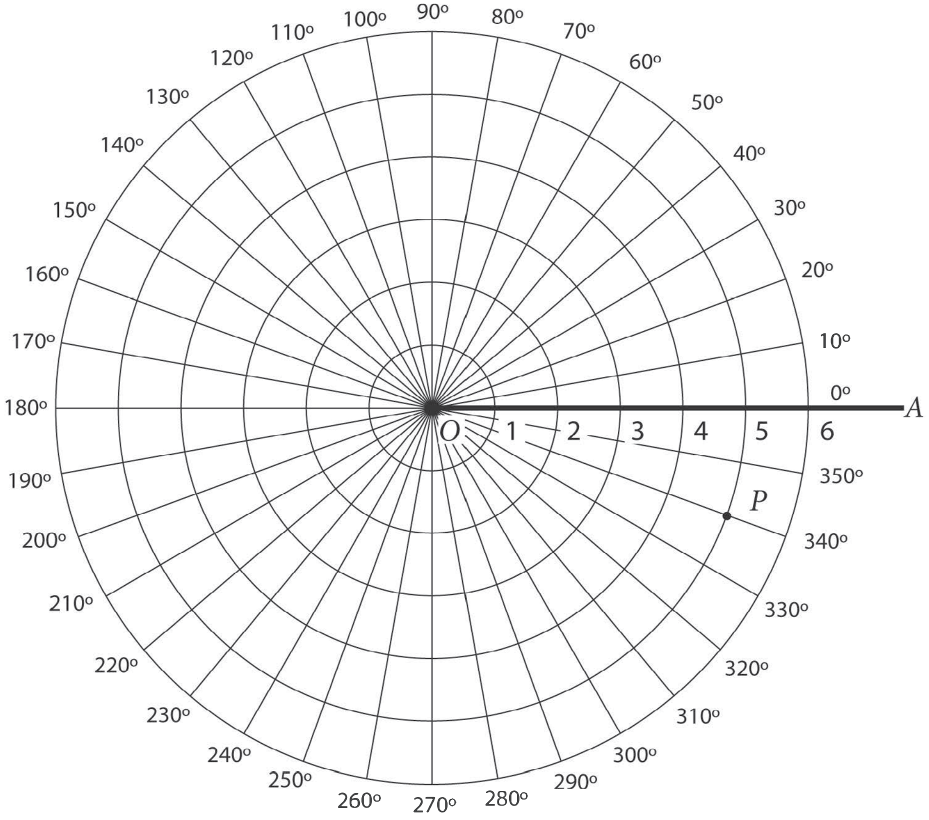
In this paper, we analysed 23 items from Blocks 6 and 7 of the TIMSS 2011 eighth-grade mathematics assessment. The data consist of non-missing responses of 748 students from the United States. Park, Lee, and Johnson (2017) developed the Q-matrix for all released items in the TIMSS 2011 eighth-grade mathematics assessment. The 23 items analysed in this study, including 22 dichotomously scored items and one polytomously scored item, measure seven out of nine attributes identified by Park *et al.* (2017), As shown in Figure 2 the polytomously scored item (i.e., Item 11) involves two tasks where the first requires students to mark two points on a graph and the second asks students to find the measure of angle based on these two points. Although students' responses to these two tasks are recorded separately, they will obtain a score of 1 if the first task is undertaken successfully and a score of 2 if both tasks are carried out successfully (Foy, Arora, & Stanco, 2013, p. 106). According to Park *et al.* (2017), the first task involves α_6 (i.e., location and movement) and α_7 (i.e., data organization, representation and interpretations), and the second task requires α_5 (i.e., lines, angles and shapes) for finding the measure of the angle, as well as α_6 and α_7 for performing the first task successfully. However, for the category level Q-matrix, only the attributes required directly by each task need to be explicitly indicated, which means only α_5 needs to be assumed necessary for the second task. The Q-matrix is given in Table 9.

To conduct the stepwise Q-matrix validation, the sequential G-DINA model was fitted to the data. For all items except Item 11, the sequential G-DINA model is equivalent to the G-DINA model. Monotonic constraints were imposed so that the processing function does not decrease if an additional attribute is mastered. Based on the stepwise procedure, modifications were suggested to six entries of 23 items as shown in Table 9. The provisional *q*-vectors for Item 11 in Figure 2 were verified by the stepwise procedure and no modifications were suggested.

The details of the stepwise Q-matrix validation procedure for the first category of Item 11 are given in Table 10. Because seven attributes are involved, at the beginning, $A = \varnothing$ and $B = \{1, 2, \ldots, 7\}$. In Step 1, there are seven single-attribute *q*-vector candidates in the initial search bank C, namely, $C = \{(1, 0, 0, 0, 0, 0, 0, 0), (0, 1, 0, 0, 0, 0, 0, 0), \ldots, (0, 0, 0, 0, 0, 0, 0, 1)\}$, and their PVAFs are given in Table 10. The *q*-vector measuring α_7 had the highest PVAF (i.e., 0.627), and thus was believed to be the best single-attribute *q*-vector. Because the highest PVAF for single-attribute *q*-vectors was <0.95, Step 2 was implemented. In Step 2, $A = \{7\}$, and the search bank C consists of six candidate *q*-vectors each measuring two attributes including α_7 and another target attribute. The Wald test was used to evaluate whether the target attributes were statistically necessary. The corresponding *p*-values are reported as p[entry] in Table 10. When α_7 was assumed to be required, adding any of the target attributes could result in better model-data fit. Because adding α_6 produced the highest PVAF, the *q*-vector measuring α_6 and α_7 was deemed the best. After the 'entry' step, the Wald test was used to evaluate whether α_7 was statistically necessary when α_6 was added in the 'removal' step. The associated *p*-value is denoted as p [removal] in the table. It turned out that α_7 was still statistically necessary when α_6 was assumed required. Because the PVAF was 0.967, which is >0.95, Step 2 terminated, and the suggested *q*-vector for this item was 0000011 with a PVAF of 0.967.

The model-data fits based on the original and suggested Q-matrices were compared. The sequential G-DINA model based on the suggested Q-matrix had a better relative fit (AIC = 19,148 and BIC = 20,030) than that based on the original Q-matrix (AIC = 19,288 and BIC = 20,179), implying that the suggested Q-matrix is preferred to the original one, though this does not guarantee that the suggested Q-matrix is correct. It is worth emphasizing again that the stepwise Q-matrix validation method should be used with intent to provide ancillary information to aid experts judgements rather than to replace the experts in determining the association between attributes and items.

The diagram shows a system for locating points



In this system, the position of a point *P* is described by its distance from origin, *O*, and the amount of counterclockwise turn from a baseline *OA* to *OP*. Thus, the coordinates of *P* are (5, 340°).

A. Mark the points *B* (3, 30°) and *C* (4, 120°) on the graph above.

B. Draw the angle *BOC*. What is the measure of angle *BOC*?

Angle *BOC* = ___________°

SOURCE: TIMSS 2011 Assessment. Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Figure 2. Item M042300Z in TIMSS 2011 assessment.

Table 9. Q-matrix for the TIMSS 2011 data

Item	TIMSS item ID	α_1	α_2	α_3	α_4	α_5	α_6	α_7
1	M042041	0	1	0	0	0	0	0
2	M042024	0	1	0	0	0	0	0
3	M042016	1	0	0	0	0	0	$\underline{1}$
4	M042002	1	0	0	0	0	0	0
5	M042198A	0	0	1	0	0	0	$\underline{0}$
6	M042198B	0	0	1	0	0	0	0
7	M042198C	0	0	1	0	0	0	0
8	M042077	1	0	0	1	0	0	0
9	M042235	0	0	0	1	0	0	0
10	M042150	0	0	0	0	1	0	0
11	M042300Z	0	0	0	0	0	1	1
11	M042300Z	0	0	0	0	1	0	0
12	M042169A	0	0	0	0	0	0	1
13	M042169B	0	0	0	0	0	0	1
14	M042169C	0	0	0	0	0	0	1
15	M032352	1	0	$\underline{1}$	0	0	0	1
16	M032725	0	1	0	0	0	$\underline{0}$	0
17	M032738	0	0	0	1	0	0	0
18	M032295	0	0	0	1	0	0	0
19	M032331	0	0	0	0	1	1	0
20	M032679	0	0	0	0	1	1	0
21	M032047	1	0	0	$\underline{1}$	0	0	0
22	M032398	0	0	0	0	1	0	0
23	M032424	0	$\underline{0}$	0	1	0	0	0

Notes. Modifications were suggested to underlined entries. α_1, whole numbers and integers; α_2, fractions, decimals and proportions; α_3, patterns; α_4, expressions, equations and functions; α_5, lines, angles and shapes; α_6, location and movement; α_7, data organization, representation and interpretations.

8. Discussion

The importance of a correctly specified Q-matrix has been recently recognized by many researchers, but research on the Q-matrix validation mainly centres on dichotomous responses. The stepwise Q-matrix validation procedure developed in this study can be used to validate the association between attributes and problem-solving steps of polytomously scored items empirically based on a recently developed CDM (i.e., the sequential G-DINA model). The stepwise Q-matrix validation method incorporates a formal hypothesis test with an effect size measure. Specifically, the procedure intends to identify all statistically required attributes for each non-zero category based on the Wald test, which takes the item parameter estimation errors into account. The GDI or PVAF from de la Torre and Chiu (2016) is used as an effect size measure to select the initial required attribute and to exclude the attributes that are identified as statistically necessary but without substantial contributions. Note that the processing functions can have various forms (e.g., conjunctive, disjunctive, additive) and the stepwise Q-matrix validation procedure does not assume that the forms of the processing functions are known. In addition, the stepwise procedure can also be applied to dichotomous response

Table 10. An illustration of the stepwise Q-matrix validation algorithm

Step	Candidate *q*-vectors	PVAF	*p*[entry]	*p*[removal]	Decision
Step 1	$A = \varnothing, B = \{1, \ldots, 7\}$				
	(1000000)	0.401			
	(0100000)	0.430			
	(0010000)	0.305			
	(0001000)	0.401			
	(0000100)	0.314			
	(0000010)	0.611			
	(0000001)	0.627			✓
Step 2					
Entry	$A = \{7\}, B = \{1, 2, 3, 4, 5, 6\}$				
	(1000001)	0.710	<.001		
	(0100001)	0.676	<.001		
	(0010001)	0.692	<.001		
	(0001001)	0.659	<.001		
	(0000101)	0.697	<.001		
	(0000011)	0.967	<.001		✓
Removal	$A = \{6, 7\}, B = \{1, 2, 3, 4, 5\}$				
	(0000011)	0.967		<.001†	✓

Note. †Denotes that this hypothesis test assesses whether α_7 is statistically necessary when α_6 is present.

data directly without the assumption about the specific forms of the CDMs involved, as long as they are special cases of the G-DINA model.

The GDI was used to identify the first required attribute, and under most conditions, it performed excellently. We also conducted a small simulation study to evaluate whether the stepwise validation method could be further improved if the first required attribute was always selected correctly. It turns out that under the condition $N = 1{,}000$, 20% misspecifications, low item quality, higher-order attribute distribution and *A*-CDM processing function, where the GDI had the worst performance in the first step, the stepwise Q-matrix validation can be improved by only 3.3% in terms of the true positive rate, and by 0.3% in terms of the true negative rate. Under other conditions, where the GDI had better performance, the improvements were even smaller.

The stepwise Q-matrix validation procedure requires that the provisional Q-matrix be largely correct. Simulation studies showed that the procedure performed well even when 20% elements in the Q-matrix were randomly misspecified. Nevertheless, its performance might be further improved after some modifications. First, the stepwise Q-matrix validation can be carried out iteratively. After obtaining the suggested Q-matrix, it can be used as the provisional Q-matrix for another round of validation. In doing so, the stepwise Q-matrix validation procedure needs to be repeated multiple times until no elements are altered in the suggested Q-matrix. Second, the posterior distribution of each individual can be updated after revising the *q*-vectors of each item. However, this modification could be more time-consuming than the current implementation. Third, the Wald test is a vital component of the stepwise validation procedure. Researchers have shown that the performance of the Wald test for model comparison and differential item functioning detection could be further improved if a better estimated variance-covariance matrix is adopted (Liu, Andersson, Xin, Zhang, & Wang, 2019; Ma & de la Torre, 2019; Ma *et al.*, 2017). Further research can investigate the performance of the stepwise Q-matrix validation procedure using the Wald test with different covariance matrix estimators.

In addition, the stepwise Q-matrix validation method assumes that the number of attributes measured by the assessment is known. The impact of missing one or more required attributes in the provisional Q-matrix is worth investigating. Second, although the sequential G-DINA model can be used for both ordinal and nominal responses, the stepwise validation method is only suitable for ordinal response data. It is important to consider exploring how to extend the proposed methods for nominal response data. Third, this study fixed the Q-matrix in simulation studies, but the structure of the Q-matrix can have an impact on the performance of the Q-matrix validation procedure. Future research might vary the characteristics of the Q-matrix such as the number of items, attributes and categories in order to examine their impact on the validation accuracy. Further, the sequential G-DINA model assumes that attributes are binary, but it should be straightforward to extend the model to accommodate polytomous attributes as in Chen and de la Torre (2013). Future research might investigate how the stepwise Q-matrix validation procedure can be extended accordingly. Lastly, as a method that can be used for dichotomous response data, it would be interesting to compare it with other Q-matrix validation methods.

Acknowledgements

This article is based on the first author's doctoral dissertation completed at Rutgers, The State University of New Jersey, under the guidance of the second author.

References

Agresti, A. (2013). *Categorical data analysis*. New York, NY: Wiley.

Chen, J. (2017). A residual-based approach to validate *Q*-matrix specifications. *Applied Psychological Measurement*, *41*, 277–293. https://doi.org/10.1177/0146621616686021

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

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, *37*, 419–437. https://doi.org/10.1177/0146621613479818

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based on diagnostic classification models. *Journal of the American Statistical Association*, *110*, 850–866. https://doi.org/10.1080/01621459.2014.934827

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*, 598–618. https://doi.org/10.1177/0146621613488436

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, *36*, 447–468. https://doi.org/10.1177/0146621612449069

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. https://doi.org/10.1111/j.1745-3984.2008.00069.x

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

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273. https://doi.org/10.1007/s11336-015-9467-8 de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353. https://doi.org/10.1007/bf02295640

de la Torre, J., van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, *51*, 281–296. https://doi.org/10.1080/07481756.2017.1327286

Efroymson, A. (1960). Multiple regression analysis. In A. Ralston & H. S. Wilf (Eds.), *Mathematical methods for digital computers* (pp. 191–203). New York, NY: Wiley.

Foy, P., Arora, A., & Stanco, G. M. (2013). *TIMSS 2011 user guide for the international database*. TIMSS & PIRLS International Study Center.

Gu, Y., Liu, J., Xu, G., & Ying, Z. (2018). Hypothesis testing of the Q-matrix. *Psychometrika*, *83*, 515–537. https://doi.org/10.1007/s11336-018-9629-6

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

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*, 98–125. https://doi.org/10.1111/jedm.12036

Liu, Y., Andersson, B., Xin, T., Zhang, H., & Wang, L. (2019). Improved Wald statistics for item-level model comparison in diagnostic classification models. *Applied Psychological Measurement*, *43*, 402–414. https://doi.org/10.1177/0146621618798664

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, *36*, 548–564. https://doi.org/10.1177/0146621612456591

Liu, J., Xu, G., & Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli*, *19*, 1790–1817. https://doi.org/10.3150/12-bej430

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. https://doi.org/10.1111/bmsp.12070

Ma, W., & de la Torre, J. (2017). *GDINA: The generalized DINA model framework*. [Computer software version 1.4.2]. Retrieved from https://CRAN.R-project.org/package=GDINA

Ma, W., & de la Torre, J. (2019). Category-level model selection for the sequential G-DINA model. *Journal of Educational and Behavioral Statistics*, *44*, 45–77. https://doi.org/10.3102/1076998618792484

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*, 200–217. https://doi.org/10.1177/0146621615621717

Ma, W., Terzi, R., Lee, S., & de la Torre, J. (2017). *Multiple group cognitive diagnosis models and their applications in detecting differential item functioning*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Antonio, TX.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. https://doi.org/10.1007/BF02296272

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*, 91–100. https://doi.org/10.1177/014662169501900110

Park, J. Y., Lee, Y.-S., & Johnson, M. S. (2017). An efficient standard error estimator of the DINA model parameters when analysing clustered data. *International Journal of Quantitative Research in Education*, *4*, 159–190. https://doi.org/10.1504/IJQRE.2017.086507

R Core Team (2017). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*, 78–96. https://doi.org/10.1177/0013164407301545

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). Berlin, Germany: Springer. https://doi.org/10.1007/978-1-4757-2691-6

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*, *41*, 614–631. https://doi.org/10.1177/0146621617707510

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, *19*, 506–532. https://doi.org/10.1177/1094428116630065

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

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

Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, *26*, 237–255. https://doi.org/10.1007/s13394-013-0090-7

Tutz, G. (1997). Sequential models for ordered responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 139–152). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-1-4757-2691-6

Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A step model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–138). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-1-4757-2691-6

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307. https://doi.org/10.1348/000711007x193957

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482. https://doi.org/10.2307/1990256

Wang, W., Song, L., Ding, S., Meng, Y., Cao, C., & Jie, Y. (2018). An EM-based method for Q-matrix validation. *Applied Psychological Measurement*, *42*, 446–459. https://doi.org/10.1177/0146621617752991

Received 25 September 2017; revised version received 26 November 2018