A Semi-supervised Learning Method for Q-Matrix Specification Under the DINA and DINO Model With Independent Structure

Wenyi Wang[1], Lihong Song[2]*, Shuliang Ding[1], Teng Wang[1], Peng Gao[1] and Jian Xiong[1]

[1] School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China, [2] Elementary Education College, Jiangxi Normal University, Nanchang, China

Cognitive diagnosis assessment (CDA) can be regarded as a kind of formative assessments because it is intended to promote assessment for learning and modify instruction and learning in classrooms by providing the formative diagnostic information about students' cognitive strengths and weaknesses. CDA has two phases, like a statistical pattern recognition. The first phase is feature generation, followed by classification stage. A Q-matrix, which describes the relationship between items and latent skills, corresponds to the feature generation phase in statistical pattern recognition. Feature generation is of paramount importance in any pattern recognition task. In practice, the Q-matrix is difficult to specify correctly in cognitive diagnosis and misspecification of the Q-matrix can seriously affect the accuracy of the classification of examinees. Based on the fact that any columns of a reduced Q-matrix can be expressed by the columns of a reachability R matrix under the logical OR operation, a semi-supervised learning approach and an optimal design for examinee sampling were proposed for Q-matrix specification under the conjunctive and disjunctive model with independent structure. This method only required subject matter experts specifying a R matrix corresponding to a small part of test items for the independent structure in which the R matrix is an identity matrix. Simulation and real data analysis showed that the new method with the optimal design is promising in terms of correct recovery rates of q-entries.

Keywords: cognitive diagnostic assessment, Q-matrix, the augment algorithm, the reachability matrix, the conjunctive model, the disjunctive model

INTRODUCTION

In educational assessment, cognitive diagnostic assessment (CDA) that combines psychometrics and cognitive science has received increased attention recently (Leighton and Gierl, 2007; Tatsuoka, 2009; Rupp et al., 2010). This approach potentially provides useful diagnostic information regarding students' strengths and weaknesses, and can facilitate individualized learning (Chang, 2015). Cognitive diagnostic models (CDMs) often utilize a Q-matrix (Embretson, 1984; Tatsuoka, 1990, 1995, 2009). Tatsuoka (2009) pointed out that "Tatsuoka (1990) organized the underlying cognitive processing skills and knowledge that are required in answering test items correctly in a Q-matrix, in which the rows represent attributes and the columns represent items." The entries of a Q-matrix

are 1 or 0, denoted by q_{kj}. If attribute k is involved in correctly answering item j, then $q_{kj} = 1$, and $q_{kj} = 0$ otherwise. The definition of Q-matrix in Tatsuoka (1990) is used in our study. Recently, one common representation of a Q-matrix is that in which the rows represent items and the columns represent attributes (Ma and de la Torre, 2020; Zhan et al., 2020). It should be noted that the representation of the Q-matrix that they used in the study differs from the traditional one.

Cognitive diagnostic assessment has two phases, like statistical pattern recognition and classification methodology. The first phase is feature generation, and then classification stage follows. The specification of Q-matrix corresponds to the feature extractor phase in statistical pattern recognition and classification problems. Feature generation is of paramount importance in any pattern recognition task. So, the Q-matrix plays a very important role in establishing the relation between latent attribute patterns and ideal/latent response patterns.

In practice, the Q-matrix is difficult to specify correctly in cognitive diagnostic assessment (Jang, 2009; DeCarlo, 2011) and misspecification of the Q-matrix can seriously affect the accuracy of both item parameter estimates and the classification of examinees (de la Torre, 2008; Rupp and Templin, 2008). Researchers have proposed several quantitative methods for deriving or refining Q-matrix. These methods can be classified into two categories (Xu and Desmarais, 2018): (a) the unsupervised method, including but not limited to the q-matrix method (Barnes, 2003, 2011), the non-negative matrix factorization technique (Desmarais, 2011; Desmarais et al., 2012; Desmarais and Naceur, 2013) or alternate least-square factorization method (Desmarais et al., 2014; Xu and Desmarais, 2016), the data-driven approach (Liu et al., 2012, 2013), and the exploratory factor analysis method (Barnes, 2003; Close, 2012; Wang et al., 2018b, 2020), and (b) the supervised method, including the sequential EM-based δ method (de la Torre, 2008) and its extension ς^2 method (de la Torre and Chiu, 2016), the Bayesian approach (DeCarlo, 2012), the non-parametric Q-matrix refinement method (Chiu, 2013), the stepwise reduction algorithm (Hartz, 2002), the EM-based methods (Wang et al., 2018a), the residual-based or item fit statistic approach (Chen, 2017; Kang et al., 2018) and so on.

The unsupervised method is deriving a Q-matrix only from test data or item responses. The unsupervised method is very useful because there are many existing tests without specifying the Q-matrix but with test response data. However, it would be difficult to identify the number of latent skills and be slightly more difficult to understand results from real data. A study of Beheshti et al. (2012) found that the number of latent skills estimated from real data is not well-aligned with the assessment of experts.

The supervised method can incorporate the information of experts' Q-matrix and test response data to refine or validate the provisional Q-matrix. If the provisional Q-matrix is unknown for an existing test, the supervised methods cannot be used. Furthermore, this method often needs a high-quality provisional Q-matrix for a whole test. If the provisional Q-matrix is specified by subject matter experts but contains a large amount of misspecification, it will be difficult for the recovery of a high-quality Q-matrix through the supervised method,

because the performance of the supervised method relies on the precision of classification of attribute patterns resulting from the provisional Q-matrix (de la Torre, 2008; Rupp and Templin, 2008).

Specifying a Q-matrix for a whole test by experts can be a time-consuming and fatigue process. The purpose of this study is to propose a semi-supervised method for Q-matrix specification in order to check whether only some of items needs to be identified by experts. The semi-supervised method falls between unsupervised and supervised methods.

MODEL AND METHOD

Model

Let K be the number of attributes. Let X_{ij} be a binary random variable to denote the response of examinee i to item j, $i = 1, 2, \ldots, N, j = 1, 2, \ldots, J$. Let $\boldsymbol{\alpha}_i$ be a column vector to denote an attribute mastery pattern or a knowledge state from the universal set of knowledge states. Moreover, Q-matrix that specifies the item-attribute relationship is a $K \times J$ matrix, in which entry $q_{kj} = 1$ if attribute k is required for answering item j correctly; otherwise, $q_{kj} = 0$.

The item response function for the deterministic inputs, noisy "and" gate (DINA) model (Haertel, 1989; Junker and Sijtsma, 2001; Chiu and Douglas, 2013) is as follows:

$$P_j(\boldsymbol{\alpha}_i) = P(X_{ij} = 1 | \boldsymbol{\alpha}_i) = g_j^{1-\eta_{ij}}(1-s_j)^{\eta_{ij}}, \quad (1)$$

where a deterministic latent response $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ki}^{q_{kj}}$ indicates whether or not examinee i possesses all of the attributes required by item j. A value of $\eta_{ij} = 1$ means that examinee i has mastered all of the attributes required by item j, and $\eta_{ij} = 0$ otherwise. The slip parameter s_j refers to the probability of an incorrect response to the item j when $\eta_{ij} = 1$, and the guessing parameter g_j refers to the probability of a correct response to item j when $\eta_{ij} = 0$. Let $\mathbf{B} = (\eta_{ij})$ be a deterministic latent response matrix for the DINA model.

The item response function for the deterministic inputs, noisy "or" gate (DINO) model (Templin and Henson, 2006; Chiu and Douglas, 2013) is as follows:

$$P_j(\boldsymbol{\alpha}_i) = P(X_{ij} = 1 | \boldsymbol{\alpha}_i) = (1-s_j)^{w_{ij}} g_j^{1-w_{ij}}, \quad (2)$$

where $w_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ki})^{q_{kj}}$ is a deterministic latent response. As in the DINA model, s_j and g_j are the slip and guessing parameters of item j. The DINA and DINO model are conjunctive and disjunctive models (Maris, 1999), respectively. Let $\mathbf{W} = (w_{ij})$ be a deterministic latent response matrix for the DINO model.

A Semi-supervised Learning Approach for the Conjunctive Model

In the rule space method (Tatsuoka, 2009) or the attribute hierarchy method (Leighton et al., 2004), the adjacency matrix denoted by $\mathbf{A}$ represents the direct relationship among attributes. We denote the entry in row k_1 and column k_2 of $\mathbf{A}$ by $a_{k_1 k_2}$. If a direct prerequisite relation exists from attribute k_1 to attribute k_2, then $a_{k_1 k_2} = 1$, and $a_{k_1 k_2} = 0$ otherwise. Let $\mathbf{R}$ denote

a reachability matrix of order (K, K) to specify the direct and indirect relationships among attributes. The $\mathbf{R}$ matrix is given by $\mathbf{R} = (\mathbf{A} + \mathbf{I})^K$ with respect to Boolean operations, where $\mathbf{I}$ is an identity matrix. The reduced Q matrix denoted by $\mathbf{Q_r}$ is obtained by removing the items (columns) that do not satisfy the specified relationships from the incidence Q matrix. The columns of $\mathbf{Q_r}$ and the zero vector forms the student matrix denoted by $\mathbf{Q_s}$ in which the columns forms the universal set of attribute patterns. If K attributes are independent, $\mathbf{A}$ is a zero matrix, $\mathbf{R}$ with K columns is an identity matrix, $\mathbf{Q_r}$ with $2^K - 1$ columns does not include the zero vector, and $\mathbf{Q_s}$ with 2^K columns contains all possible combinations of attribute patterns.

We assume that the cognitive requirement for the multiple skills within an item is conjunctive (Maris, 1999), that is, answering an item correctly requires mastery of all the skills required by that item. For the conjunctive model, Example 1 will show the relationship of latent responses on items with q-vectors corresponding to $\mathbf{R}$ and $\mathbf{Q_r}$.

Example 1 for an independent structure. Let $K = 2$, $\mathbf{R} = \begin{bmatrix} \mathbf{r_1} & \mathbf{r_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\mathbf{Q_r} = \begin{bmatrix} \mathbf{q_1} & \mathbf{q_2} & \mathbf{q_3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$, and $\mathbf{Q_s} = \begin{bmatrix} \boldsymbol{\alpha_1} & \boldsymbol{\alpha_2} & \boldsymbol{\alpha_3} & \boldsymbol{\alpha_4} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$.

Given $\mathbf{Q_s}$ and a test Q-matrix of $\mathbf{Q_r}$, a latent response matrix $\mathbf{B} = \begin{bmatrix} \boldsymbol{\eta_1} & \boldsymbol{\eta_2} & \boldsymbol{\eta_3} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$ can be calculated, in which the entry in row i and column j is the deterministic latent response of η_{ij}. If 0 corresponds to F (false) and 1 corresponds to T (true), the logical conjunction and disjunction operators, $\vee$ and $\wedge$, can be applied to two binary vectors of equal length, by taking the bitwise AND or OR of each pair of bits at corresponding positions. It can be observed that $\boldsymbol{\eta_3} = \boldsymbol{\eta_1} \wedge \boldsymbol{\eta_2}$, where $\boldsymbol{\eta_3} = \boldsymbol{\eta_1} \wedge \boldsymbol{\eta_2}$ is the conjunction of $\boldsymbol{\eta_1}$ and $\boldsymbol{\eta_2}$. This is because the relationship $\mathbf{q_3} = \mathbf{q_1} \vee \mathbf{q_2}$ is true, where $\mathbf{q_1} \vee \mathbf{q_2}$ is the disjunction of $\mathbf{q_1}$ and $\mathbf{q_2}$.

Example 1 illustrates the following fact. For the conjunctive model, consider two latent response matrices denoted by $\mathbf{B_1}$ and $\mathbf{B_2}$ from two tests corresponding two Q-matrices $\mathbf{Q_r}$ and $\mathbf{R}$, where denoted as a reachability matrix. It means that $\mathbf{B_1}$ and $\mathbf{B_2}$ can be generated, respectively from the reduced Q-matrix and the reachability matrix based on the universal set of attribute patterns. From the example above, then any columns of the $\mathbf{B_1}$ can be expressed by the columns of the $\mathbf{B_2}$ under the logical AND operation. This is because the augmented algorithm proposed by Ding et al. (2008, 2009) in the generalized Q-matrix theory (Ding et al., 2015) provided the useful fact that any columns of the reduced Q-matrix can be expressed by the columns of the reachability matrix under the logical OR operation. The argument in Example 1 can be adapted to prove the following theorem.

Theorem 1. For the conjunctive model, if K attributes are independent, then $\mathbf{q_j} = \vee_{l \in S_j} \mathbf{r_l}$ if and only if $\eta_{ij} = \wedge_{l \in S_j} \eta_{il}$, where $\boldsymbol{\alpha}_i$ is any column of $\mathbf{Q_s}$ and S_j is a subset of $\{1, 2, \dots, K\}$.

Proof: If $\mathbf{q_j} = \vee_{l \in S_j} \mathbf{r_l}$, we need to consider two cases, when $\eta_{ij} = 1$ and $\eta_{ij} = 0$. If $\eta_{ij} = 1$ for $\boldsymbol{\alpha}_i$ as a column of $\mathbf{Q_s}$, we know that $\alpha_{ki} = 1$ for all attributes k with $q_{kj} = 1$ by the

definition of the deterministic latent response. That is, examinee i has mastered all the skills required by item j. Since $\mathbf{q_j} = \vee_{l \in S_j} \mathbf{r_l}$, then by the definition of conjunction, we can conclude that $\alpha_{ki} = 1$ for all attributes k with $r_{kl} = 1$ for all $l \in S_j$. We now use the definition of the deterministic latent response to conclude that $\eta_{il} = 1$ for all $l \in S_j$, that is, $\wedge_{l \in S_j} \eta_{il} = 1$. This shows that $\eta_{ij} = \wedge_{l \in S_j} \eta_{il}$ when $\eta_{ij} = 1$. If $\eta_{ij} = 0$ for $\boldsymbol{\alpha}_i$ as a column of $\mathbf{Q_s}$, we know that $\alpha_{ki} = 0$ for at least one of attributes with $q_{kj} = 1$ by the definition of the deterministic latent response. That is, examinee i has not mastered all the skills required by item j. Since $q_{kj} = 1$ and $\mathbf{q_j} = \vee_{l \in S_j} \mathbf{r_l}$, there is an item l in S_j such that $r_{kl} = 1$. This means that item l measured attribute k. Since $\alpha_{ki} = 0$, then by the definition of the deterministic latent response, it follows that $\eta_{il} = 0$ for at least one of items in S_j, that is, $\wedge_{l \in S_j} \eta_{il} = 0$. This show that $\eta_{ij} = \wedge_{l \in S_j} \eta_{il}$ when $\eta_{ij} = 0$. Next, we try to prove the converse. First suppose that there exists an attribute $k \in \{1, 2, \dots, K\}$ such that $\vee_{l \in S_j} r_{kl} = 1$ and $q_{kj} = 0$. Since $\vee_{l \in S_j} r_{kl} = 1$, we know that there exists an item $l \in S_j$ with $r_{kl} = 1$. Due to the arbitrariness of $\boldsymbol{\alpha}_i$, let $\boldsymbol{\alpha}_i = \mathbf{1} - \mathbf{e}_k$, where $\mathbf{1} = (1 \ 1 \ \dots \ 1)^T$ and $\mathbf{e}_k$ is the vector with a 1 in the kth entry and 0's elsewhere. This is a contradiction, because we know that $\eta_{ij} = 1$, while $\wedge_{l \in S_j} \eta_{il} = 0$. Similarly, we assume that there exists an attribute $k \in \{1, 2, \dots, K\}$ such that $\vee_{l \in S_j} r_{kl} = 0$ and $q_{kj} = 1$. One can still take $\boldsymbol{\alpha}_i = \mathbf{1} - \mathbf{e}_k$. This is also a contradiction, because we know that $\eta_{ij} = 0$, while $\wedge_{l \in S_j} \eta_{il} = 1$. The proof is complete.

The important fact about Theorem 1 is that if a latent response matrix is calculated from a Q-matrix, the relationship between the columns in the Q-matrix can be constructed from the relationship between the corresponding columns in the latent response matrix. It should be noted that an observed item response is a function of an underlying latent response and slip and guessing parameters. In other words, the noise introduced in the process is due to slip and guessing parameters.

Next, we will introduce a semi-supervised learning method for Q-matrix specification for the conjunctive model by using the result of Theorem 1 and considering the noise in item responses. Without loss of generality, we begin by arbitrarily assigning q-vector $\mathbf{q_j}$ to item j. Given a test Q-matrix, written as $\mathbf{Q}_t = [\mathbf{R}_{K \times K} \ \mathbf{q_j}] = [\mathbf{r_1} \ \mathbf{r_2} \ \dots \ \mathbf{r_K} \ \mathbf{q_j}]$, where $\mathbf{R}$ is a reachability matrix specified by subject matter experts and the remaining $\mathbf{q_j}$ is unknown. Let $\mathbf{U} = [\mathbf{X}_{N \times K} \ \mathbf{Y}_{N \times 1}]$ be an item response matrix on $\mathbf{Q}_t$, where N is the sample size. The estimate of $\mathbf{q_j}$ can be written as

$$\hat{\mathbf{q}}_j = \vee_{r_k \in \hat{S}_j} \mathbf{r}_k, \tag{3}$$

where logical OR is applied to the corresponding entries of the columns in the following set of $\hat{S}_j$

$$\hat{S}_j = \underset{S \in \mathrm{P}(\{r_1, r_2, \dots, r_K\}) - \varnothing}{\arg\min} (Y_j - \wedge_{r_k \in S} X_k)^T (Y_j - \wedge_{r_k \in S} X_k), \tag{4}$$

where $\mathrm{P}(\{r_1, r_2, \dots, r_K\})$ is the power set of the set $\{\mathbf{r_1}, \mathbf{r_2}, \dots, \mathbf{r_K}\}$. The exhaustive method with time complexity $O(2^K)$ provided a simple way to find a global solution of $\hat{S}_j$.

A Semi-supervised Learning Approach for the Disjunctive Model

For the disjunctive model, the deterministic latent response on an item is correct if and only if an examinee has mastered at least one of the skills required by the item. This is illustrated in Example 2. Similar to what we did in Example 1, Example 2 will show the relationship of latent responses on items with q-vectors corresponding to $\mathbf{R}$ and $\mathbf{Q_r}$.

Example 2 for an independent structure. Let $K = 2$, $\mathbf{R} = \begin{bmatrix} \mathbf{r_1} & \mathbf{r_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\mathbf{Q_s} = \begin{bmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \boldsymbol{\alpha}_3 & \boldsymbol{\alpha}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$, and

$\mathbf{Q_r} = \begin{bmatrix} \mathbf{q_1} & \mathbf{q_2} & \mathbf{q_3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$. From $\mathbf{Q_s}$ and $\mathbf{Q_r}$, a latent response

matrix $\mathbf{W}_1 = \begin{bmatrix} \mathbf{w_1} & \mathbf{w_2} & \mathbf{w_3} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ can be calculated, in

which the entry in row i and column j is the deterministic latent response of w_{ij}. It can be observed that $\mathbf{w_3} = \mathbf{w_1} \vee \mathbf{w_2}$. This is because the relationship $\mathbf{q_3} = \mathbf{q_1} \vee \mathbf{q_2}$ is true.

Consider a latent response matrix, denoted by $\mathbf{W}_2 = \begin{bmatrix} \mathbf{w_1} & \mathbf{w_2} \end{bmatrix}$, corresponding to the $\mathbf{R}$ matrix. The fact illustrated in Example 2 is that any columns of the $\mathbf{W}_1$ can be expressed by the columns of the $\mathbf{W}_2$ under the logical OR operation for the disjunctive model. This is also because the augmented algorithm proposed by Ding et al. (2008, 2009) in the generalized Q-matrix theory (Ding et al., 2015) provided the useful fact that any columns of the reduced Q-matrix can be expressed by the columns of the reachability matrix under the logical OR operation. The following theorem gives the precise statement.

Theorem 2. For the disjunctive model, if K attributes are independent, then $\mathbf{q_j} = \vee_{l \in S_j} \mathbf{r_l}$ if and only if $w_{ij} = \vee_{l \in S_j} w_{il}$, where $\boldsymbol{\alpha}_i$ is any column of $\mathbf{Q_s}$ and S_j is a subset of $\{1, 2, \ldots, K\}$.

Proof: If $\mathbf{q_j} = \vee_{l \in S_j} \mathbf{r_l}$, we need to consider two cases, when $w_{ij} = 1$ and $w_{ij} = 0$. If $w_{ij} = 1$ for $\boldsymbol{\alpha}_i$ as a column of $\mathbf{Q_s}$, we know that $\alpha_{ki} = 1$ for at least one of attributes k with $q_{kj} = 1$ by the definition of the deterministic latent response. That is, examinee i has mastered at least one of the attributes required by item j. Without loss of generality, we assume $\alpha_{ki} = 1$ and $q_{kj} = 1$. Since $\mathbf{q_j} = \vee_{l \in S_j} \mathbf{r_l}$, then by the definition of disjunction, we can conclude that $r_{kl} = 1$ is true for at least one of $l \in S_j$. From the definition of the deterministic latent response, it follows that there is at least one item $l \in S_j$ such that $w_{il} = 1$, that is, $\vee_{l \in S_j} w_{il} = 1$. This show that $w_{ij} = \vee_{l \in S_j} w_{il}$ when $w_{ij} = 1$. If $w_{ij} = 0$ for $\boldsymbol{\alpha}_i$ as a column of $\mathbf{Q_s}$, we know that $w_{ki} = 0$ for all of attributes with $q_{kj} = 1$ by the definition of the deterministic latent response. That is, examinee i has not mastered any skills required by item j. Since $\mathbf{q_j} = \vee_{l \in S_j} \mathbf{r_l}$, examinee i has not mastered any skills required by any item $l \in S_j$. If we suppose that examinee i has mastered at least one of attributes required by an item $l \in S_j$, then $w_{ij} = 1$, which is a contradiction. It means that item l measured attribute k. It follows that $w_{il} = 0$ for all of items in S_j, that is, $\vee_{l \in S_j} w_{il} = 0$, directly from the definition of the deterministic latent response. This show that $w_{ij} = \vee_{l \in S_j} w_{il}$ when $w_{ij} = 0$. Next, we use a proof by contradiction to prove the converse. First assume that there exists an attribute $k \in \{1, 2, \ldots, K\}$ such that $\vee_{l \in S_j} r_{kl} = 1$ and

$q_{kj} = 0$. Since $\vee_{l \in S_j} r_{kl} = 1$, we know that there exists an item $l \in S_j$ with $r_{kl} = 1$. Due to the arbitrariness of $\boldsymbol{\alpha}_i$, let $\boldsymbol{\alpha}_i = \mathbf{e}_k$, where $\mathbf{e}_k$ is the vector with a 1 in the kth entry and 0's elsewhere. Then, we have $w_{il} = 1$ and $w_{ij} = 0$. Since $w_{ij} = \vee_{l \in S_j} w_{il}$, we know that $w_{ij} = 1$ and arrive at a contradiction. Similarly, we assume that there exists an attribute $k \in \{1, 2, \ldots, K\}$ such that $\vee_{l \in S_j} r_{kl} = 0$ and $q_{kj} = 1$. One can still take $\boldsymbol{\alpha}_i = \mathbf{e}_k$. This is also a contradiction, because we know that $w_{ij} = 1$, while $\wedge_{l \in S_j} w_{il} = 0$. The proof is complete.

The important fact about Theorem 2 is that one can derive the relationship between the columns of a Q-matrix from the relationship between the columns of corresponding latent response matrix. For considering the noise introduced in item responses due to slipping and guessing, we will introduce a semi-supervised learning method for Q-matrix specification for the disjunctive model by using the result of Theorem 2. Without loss of generality, we begin by arbitrarily assigning a q-vector to $\mathbf{q_j}$. Given a test Q-matrix, written as $\mathbf{Q}_t = [\mathbf{R}_{K \times K} \ \mathbf{q_j}] = [\mathbf{r_1} \ \mathbf{r_2} \ \ldots \ \mathbf{r_K} \ \mathbf{q_j}]$, where $\mathbf{R}$ is a reachability matrix specified by subject matter experts and the remaining $\mathbf{q_j}$ is unknown. Let $\mathbf{U} = [\mathbf{X}_{N \times K} \ \mathbf{Y}_{N \times 1}]$ be an item response matrix on $\mathbf{Q}_t$. The estimate of $\mathbf{q_j}$ can be written as

$$\hat{\mathbf{q}}_j = \vee_{r_k \in \hat{S}_j} \mathbf{r_k}, \tag{5}$$

where logical OR is applied to the corresponding entries of the columns in the following set of $\hat{S}_j$

$$\hat{S}_j = \underset{S \in \text{P}(\{r_1, r_2, \ldots, r_K\}) - \varnothing}{\arg \min} (Y_j - \vee_{r_k \in S} X_k)^T (Y_j - \vee_{r_k \in S} X_k), \tag{6}$$

where $\text{P}(\{r_1, r_2, \ldots, r_K\})$ is the power set of the set $\{\mathbf{r_1}, \mathbf{r_2}, \ldots, \mathbf{r_K}\}$. The exhaustive method with time complexity $O(2^K)$ provided a simple way to find a global solution of $\hat{S}_j$.

A SIMULATION STUDY

Study Design

A simulation study was conducted to investigate the performance of the new method under five factors, such as sample size, item parameters for items corresponding to a reachability matrix, item parameters for new or raw items with unknown q-vectors, two cognitive diagnostic models (the DINA and DINO model), and two designs. Five attributes were considered in the simulation study. Matlab 2015a and R-3.6.1 were used for estimating unknown Q-matrix and analyzing real data below.

In the simulation study, a test Q-matrix $\mathbf{Q_t} = [\mathbf{R} \ \mathbf{Q_r}]$ consists of an identity or a reachability matrix and a reduced Q-matrix, where the reduced Q-matrix with 31 items includes all non-zero possible q-vectors. The number of examinees has 10 levels, such as $N = 30, 60, \ldots,$ and 300. Item parameters for $\mathbf{R}$ and $\mathbf{Q_r}$ have 10 levels, such as $0, 0.05, \ldots,$ and 0.45. In general, for the DINA or DINO model, a high quality or "good" item will have small slip and guessing parameters (Rupp et al., 2010), which means that the noise are small.

Random and optimal designs were considered in the simulation study. For the random design, attribute patterns for

examinees were generated by taking each of the 2^5 possible patterns with equal probability for each sample size. From the proof of Theorem 1 above, we know that the following set of attribute patterns for examinees plays a very important role in discriminating latent response vectors of different q-vectors under the DINA model

$$S_{DINA} = \{\mathbf{1} - \mathbf{e}_1 \mathbf{1} - \mathbf{e}_2, \ldots, \mathbf{1} - \mathbf{e}_K\}\{\begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \ldots, \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 0 \end{bmatrix}\} \quad (7)$$

where $\mathbf{e}_k$ is the vector with a 1 in the kth entry and 0's otherwise. From the proof of Theorem 2 above, another set of attribute patterns for examinees plays a very important role in discriminating latent response vectors of different q-vectors under the DINO model as follows

$$S_{DINO} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K\}\{\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}\}, \quad (8)$$

where $\mathbf{e}_k$ is the vector with a 1 in the kth entry and 0's otherwise. For the optimal design, attribute patterns for examinees under the DINA or DINA model were randomly drawn with replacement from the set of S_{DINA} or S_{DINO}, respectively. Optimal designs for two models are possible to meet the needs of learners at different stages of skills and knowledge acquisition. For example, the attribute patterns in S_{DINO} containing only one skill. This condition is really improbable for summary assessments in real situations, but is expected to be common for novice learners with respect to the new content to be learned in formative assessments or classroom assessments.

Data Simulation

Simulated data were generated using five attributes. Based on the simulated Q-matrix, item parameters, and attribute patterns, item responses are generated in the following way

$$X_{ij} = \begin{cases} 1, & if \ u \leq P_j(\boldsymbol{\alpha}_i), \\ 0, & otherwise, \end{cases} \quad (9)$$

where u is a random value from a Uniform (0, 1) distribution and $P_j(\boldsymbol{\alpha_i})$ is the item response function of the DINA or DINO model. A total of 4,000 conditions were simulated (10 sample sizes × 10 item parameters × 10 item parameters × 2 models × 2 designs). Thirty replication data sets were simulated for each condition.

Evaluation Criterion

The performance of the new method is evaluated in terms of the correct recovery rate (CRR) of q-entries. The correct recovery rate equals the ratio of the number of correct q-entries in the estimated Q-matrix to the total number of q-entries (Chiu, 2013)

$$CRR = \frac{1}{KM} \sum_{k=1}^{K} \sum_{j=1}^{M} \mathbf{I}(\hat{q}_{kj} = q_{kj}), \quad (10)$$

TABLE 1 | Mean and standard deviation (in brackets) of correct recovery rate of q-entries for two models and two designs.

Sample size	The DINA model		The DINO model	
	Random design	Optimal design	Random design	Optimal design
30	0.651 (0.126)	0.720 (0.147)	0.653 (0.126)	0.721 (0.146)
60	0.699 (0.145)	0.769 (0.153)	0.700 (0.144)	0.770 (0.151)
90	0.725 (0.149)	0.796 (0.152)	0.726 (0.149)	0.796 (0.151)
120	0.742 (0.151)	0.815 (0.149)	0.743 (0.151)	0.815 (0.148)
150	0.756 (0.153)	0.827 (0.146)	0.754 (0.153)	0.829 (0.145)
180	0.764 (0.153)	0.839 (0.143)	0.765 (0.153)	0.838 (0.144)
210	0.772 (0.153)	0.847 (0.141)	0.772 (0.151)	0.846 (0.141)
240	0.779 (0.152)	0.854 (0.138)	0.777 (0.152)	0.854 (0.138)
270	0.784 (0.151)	0.858 (0.137)	0.783 (0.152)	0.858 (0.137)
300	0.789 (0.151)	0.863 (0.135)	0.789 (0.152)	0.864 (0.135)
Mean	0.746 (0.154)	0.819 (0.151)	0.746 (0.154)	0.819 (0.150)

where $M = 31$ is the number of columns of the unknown Q-matrix $\mathbf{Q_r}$, q_{kj} is an (k, j)th entry of the simulated $\mathbf{Q_r}$, and $\hat{q}_{kj}$ is an (k, j) entry of the $\hat{\mathbf{Q}}_{\mathbf{r}}$ estimated from the new method. The mean and standard deviation of the CRR values of the 30 replications were reported for each condition.

Results

Table 1 lists descriptive statistics of correct recovery rate of q-entries for two models and two designs across other conditions. It is clear that the mean of correct recovery rates of q-entries tends to increase as sample size increases, but sample size has slightly affected the standard deviations of correct recovery rates. It should be noted that the mean of correct recovery rates of the optimal design is larger than that of the random design. The semi-supervised learning method for q-matrix specification performed similarly under two cognitive diagnostic models. In addition, since there are 32 possible attribute patterns, no all attribute patterns can be observed in the first sample size condition ($N = 30$). This might lead to lower rate of correct recovery observed for this condition.

Table 2 shows the correct recovery rates of q-entries from the new method with sample size of 300 for the DINA model under the random design. From correct recovery rates of q-entries, when item parameters for items with known (i.e., the reachability matrix) and unknown q-vectors are ≤0.2, most of the average of correct recovery rates of q-entries for the semi-supervised method are larger than or equal to 0.9. From trends of marginal means of last rows and columns in **Table 2**, item parameters of the reachability matrix have a relatively larger impact on the performance of the semi-supervised method than item parameters with unknown q-vectors.

Table 3 presents the correct recovery rates of q-entries from the new method with sample size of 300 for the DINA model under the optimal design. From correct recovery rates of q-entries, when item parameters for items with known and unknown q-vectors are ≤0.25, the average of correct recovery

TABLE 2 | The correct recovery rates of q-entries with sample size of 300 for the DINA model and random design.

Item parameters for the R	Item parameters for the new items										
	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	M
0.00	**1.000**	**1.000**	**1.000**	**0.999**	**0.998**	**0.993**	**0.974**	**0.929**	0.837	0.665	0.939
0.05	**1.000**	**1.000**	**0.997**	**0.995**	**0.991**	**0.976**	**0.946**	**0.908**	0.791	0.642	0.925
0.10	**0.994**	**0.992**	**0.984**	**0.984**	**0.967**	**0.954**	**0.927**	0.857	0.767	0.632	0.906
0.15	**0.974**	**0.968**	**0.958**	**0.954**	**0.932**	**0.915**	0.866	0.807	0.724	0.608	0.870
0.20	**0.927**	**0.922**	**0.910**	**0.901**	0.881	0.860	0.826	0.776	0.692	0.591	0.829
0.25	0.866	0.846	0.850	0.839	0.825	0.802	0.776	0.733	0.652	0.567	0.775
0.30	0.791	0.801	0.782	0.793	0.760	0.735	0.727	0.670	0.624	0.563	0.725
0.35	0.728	0.718	0.720	0.709	0.709	0.698	0.683	0.637	0.604	0.546	0.675
0.40	0.673	0.686	0.681	0.680	0.668	0.643	0.620	0.608	0.589	0.527	0.638
0.45	0.647	0.634	0.623	0.620	0.615	0.612	0.604	0.575	0.575	0.537	0.604
M	0.860	0.857	0.851	0.847	0.835	0.819	0.795	0.750	0.686	0.588	0.789

The bold values are larger than 0.9.

TABLE 3 | The correct recovery rates of q-entries with sample size of 300 for the DINA model and optimal design.

Item parameters for the R	Item parameters for the new items										
	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	M
0.00	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**	**0.991**	**0.939**	0.780	0.971
0.05	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**	**0.997**	**0.981**	**0.917**	0.754	0.965
0.10	**1.000**	**1.000**	**1.000**	**0.999**	**0.998**	**0.995**	**0.981**	**0.955**	0.871	0.714	0.951
0.15	**1.000**	**1.000**	**0.998**	**0.992**	**0.988**	**0.981**	**0.959**	**0.915**	0.841	0.688	0.936
0.20	**0.986**	**0.986**	**0.982**	**0.973**	**0.962**	**0.952**	**0.918**	0.875	0.785	0.661	0.908
0.25	**0.959**	**0.952**	**0.947**	**0.933**	**0.926**	**0.909**	0.879	0.829	0.754	0.624	0.871
0.30	**0.930**	**0.914**	**0.912**	**0.909**	0.883	0.865	0.832	0.786	0.712	0.610	0.835
0.35	0.886	0.888	0.880	0.865	0.847	0.813	0.780	0.724	0.662	0.573	0.792
0.40	0.847	0.834	0.816	0.804	0.774	0.753	0.721	0.679	0.629	0.562	0.742
0.45	0.749	0.738	0.717	0.720	0.688	0.687	0.656	0.619	0.581	0.556	0.671
M	0.936	0.931	0.925	0.920	0.907	0.895	0.872	0.835	0.769	0.652	0.864

The bold values are larger than 0.9.

rates of q-entries for the semi-supervised method are larger than or equal to 0.9. However, item parameters for known q-vectors have slightly larger impact on the performance of the semi-supervised method than for unknown q-vectors, because the row means decreased more quickly than the column means. We need to compare the **Tables 2**, **3** to see which designs are promising. The number of correct recovery rates above 0.9 in **Table 3** were found to be larger than that of **Table 2**. **Tables 4**, **5** show the correct recovery rates of q-entries from the new method with sample size of 300 for the DINO model under the random and optimal design. It can be observed that results for the DINO model are the same as those for the DINA model described above.

REAL DATA ANALYSIS

The purpose of the real data analysis is to examinee whether the proposed method is promising for a non-independent structure under the conjunctive model based on an intuitive fact from the following example.

Example 3 for an unstructured hierarchy under the conjunctive model. Let $K = 3$, $\mathbf{R} = [\mathbf{r_1}\ \mathbf{r_2}\ \mathbf{r_3}] = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $\mathbf{Q_r} = [\mathbf{q_1}\ \mathbf{q_2}\ \mathbf{q_3}\ \mathbf{q_4}] = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$, and

$\mathbf{Q_s} = [\boldsymbol{\alpha_0}\ \boldsymbol{\alpha_1}\ \boldsymbol{\alpha_2}\ \boldsymbol{\alpha_3}\ \boldsymbol{\alpha_4}] = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$. From the ideal

response matrix $\mathbf{B} = [\boldsymbol{\eta_1}\ \boldsymbol{\eta_2}\ \boldsymbol{\eta_3}\ \boldsymbol{\eta_4}] = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$, it can be

observed that $\boldsymbol{\eta_4} = \boldsymbol{\eta_2} \wedge \boldsymbol{\eta_3}$ or $\boldsymbol{\eta_4} = \boldsymbol{\eta_1} \wedge \boldsymbol{\eta_2} \wedge \boldsymbol{\eta_3}$. This is because the relationship $\mathbf{q_4} = \mathbf{q_2} \vee \mathbf{q_3}$ or $\mathbf{q_4} = \mathbf{q_1} \vee \mathbf{q_2} \vee \mathbf{q_3}$ is true.

A common data set pertaining to fraction-subtraction data contains 20 items and 536 examines (de la Torre and Douglas, 2004). In our real data analysis, we focused on the analysis of a subset of test items where the expert Q-matrix comes

TABLE 4 | The correct recovery rates of q-entries with sample size of 300 for the DINO model and random design.

Item parameters for the R	Item parameters for the new items										
	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	M
0.00	**1.000**	**1.000**	**1.000**	**0.999**	**0.998**	**0.990**	**0.976**	**0.929**	0.840	0.685	0.942
0.05	**1.000**	**0.999**	**0.998**	**0.996**	**0.991**	**0.980**	**0.946**	0.889	0.796	0.650	0.925
0.10	**0.996**	**0.991**	**0.992**	**0.985**	**0.968**	**0.945**	**0.918**	0.848	0.767	0.639	0.905
0.15	**0.971**	**0.969**	**0.963**	**0.947**	**0.931**	**0.912**	0.864	0.806	0.724	0.604	0.869
0.20	**0.927**	**0.921**	**0.915**	**0.905**	0.885	0.854	0.819	0.763	0.687	0.578	0.825
0.25	0.855	0.854	0.865	0.842	0.831	0.797	0.773	0.733	0.665	0.572	0.779
0.30	0.787	0.795	0.789	0.774	0.763	0.743	0.722	0.677	0.629	0.571	0.725
0.35	0.734	0.721	0.717	0.720	0.720	0.686	0.662	0.634	0.597	0.550	0.674
0.40	0.677	0.678	0.689	0.675	0.660	0.654	0.623	0.624	0.573	0.536	0.639
0.45	0.632	0.630	0.628	0.633	0.611	0.608	0.600	0.586	0.571	0.525	0.603
M	0.858	0.856	0.856	0.848	0.836	0.817	0.790	0.749	0.685	0.591	0.789

The bold values are larger than 0.9.

TABLE 5 | The correct recovery rates of q-entries with sample size of 300 for the DINO model and optimal design.

Item parameters for the R	Item parameters for the new items										
	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	M
0.00	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**	**0.992**	**0.939**	0.780	0.971
0.05	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.995**	**0.973**	**0.916**	0.744	0.963
0.10	**1.000**	**1.000**	**1.000**	**1.000**	**0.998**	**0.994**	**0.983**	**0.954**	0.873	0.715	0.952
0.15	**0.998**	**0.999**	**0.996**	**0.995**	**0.985**	**0.978**	**0.957**	**0.922**	0.836	0.695	0.936
0.20	**0.991**	**0.986**	**0.982**	**0.970**	**0.961**	**0.946**	**0.916**	0.873	0.795	0.666	0.909
0.25	**0.957**	**0.951**	**0.951**	**0.935**	**0.925**	**0.909**	0.885	0.829	0.752	0.628	0.872
0.30	**0.922**	**0.910**	**0.913**	**0.902**	0.894	0.866	0.840	0.772	0.705	0.595	0.832
0.35	0.887	0.882	0.870	0.862	0.842	0.807	0.787	0.731	0.678	0.571	0.792
0.40	0.838	0.830	0.812	0.802	0.788	0.749	0.715	0.675	0.620	0.560	0.739
0.45	0.737	0.733	0.733	0.703	0.690	0.663	0.645	0.619	0.595	0.544	0.666
M	0.933	0.929	0.926	0.917	0.908	0.891	0.872	0.834	0.771	0.650	0.863

The bold values are larger than 0.9.

from **Table 7** both in de la Torre (2008) or DeCarlo (2012). The labels given to the five skills are (A1) performing basic fraction-subtraction operation, (A2) simplifying/reducing, (A3) separating whole numbers from fractions, (A4) borrowing one from whole number to fraction, and (A5) converting whole numbers to fractions.

We assumed the corresponding Q-matrix of items 3, 8, 9, 12, and 10 known since these item parameters are relatively small and the q-vectors of other items are combinations of q-vectors for these five items. Then, the semi-supervised method was applied to estimate q-vectors for the other 10 items. Results in **Table 6** show that the agreement rate of q-entries between the estimate and expert Q-matrix on the 10 items is 84%. The estimated q-entries suggest that items 4, 7, 13, 14, and 15 do not require attribute 2 (simplifying/reducing). Item 4 (similar to item 14) do not required attribute A2, which is consistent with results from DeCarlo (2012). Items 7, 13, and 15 can be answered correctly by using attributes required by item 12.

The estimated q-vector of item 1 has largest discrepancy with the expert q-vector. The reason might be that solving item 1 correctly needs to find a common denominator and then performs basic fraction-subtraction operation. The guessing and slip parameter of item 1 are 0.0001 and 0.2769 under the expert q-vector, respectively. The guessing and slip parameter of item 1 are 0.3408 and 0.0716 under the estimated q-vector, respectively. Since item 1 requires an extra attribute (i.e., find a common denominator), the slip parameter for the expert q-vector is relatively large, while the estimated q-vector contains some unnecessary attributes, the guessing parameter is relatively large. In the estimated Q-matrix, attribute A4 has been added to item 11. The guessing probability of item 11 increased sensibly (from 0.10 to 0.48). It indicated that attribute A4 is not necessary for item 11 because this item is different from items 7, 12, and so on.

The generalized DINA model (GDINA; de la Torre, 2011), the DINA model, the linear logistic model (LLM; Fischer, 1995), and

TABLE 6 | The expert and estimated Q-matrix and item parameters estimates of the DINA model for the fractional subtraction data.

No.	Items	The expert Q-matrix and item parameters estimates of the DINA model							The estimated Q-matrix and item parameters estimates of the DINA model						
		A1	A2	A3	A4	A5	$\hat{g}$	$\hat{s}$	A1	A2	A3	A4	A5	$\hat{g}$	$\hat{s}$
1	$\frac{3}{4} - \frac{3}{8}$	1	0	0	0	0	0.00	0.28	1	0	**1**	**1**	**1**	0.34	0.07
2	$3\frac{1}{2} - 2\frac{3}{2}$	1	1	1	1	0	0.21	0.12	1	1	1	1	0	0.21	0.11
3	$\frac{6}{7} - \frac{4}{7}$	1	0	0	0	0	0.14	0.04	1	0	0	0	0	0.10	0.05
4	$3 - 2\frac{1}{5}$	1	1	1	1	1	0.12	0.13	1	**0**	1	1	1	0.12	0.18
5	$3\frac{7}{8} - 2$	1	0	1	0	0	0.34	0.25	1	0	1	0	0	0.35	0.25
6	$4\frac{4}{12} - 2\frac{7}{12}$	1	1	1	1	0	0.03	0.23	1	1	1	1	0	0.03	0.23
7	$4\frac{1}{3} - 2\frac{4}{3}$	1	1	1	1	0	0.07	0.08	1	**0**	1	1	0	0.07	0.08
8	$\frac{11}{8} - \frac{1}{8}$	1	1	0	0	0	0.16	0.05	1	1	0	0	0	0.09	0.04
9	$3\frac{4}{5} - 3\frac{2}{5}$	1	0	1	0	0	0.08	0.06	1	0	1	0	0	0.04	0.04
10	$2 - \frac{1}{3}$	1	0	1	1	1	0.17	0.07	1	0	1	1	1	0.15	0.09
11	$4\frac{5}{7} - 1\frac{4}{7}$	1	0	1	0	0	0.10	0.10	1	0	1	**1**	0	0.48	0.07
12	$7\frac{3}{5} - \frac{4}{5}$	1	0	1	1	0	0.03	0.13	1	0	1	1	0	0.05	0.14
13	$4\frac{1}{10} - 2\frac{8}{10}$	1	1	1	1	0	0.13	0.16	1	**0**	1	1	0	0.13	0.16
14	$4 - 1\frac{4}{3}$	1	1	1	1	1	0.02	0.20	1	**0**	1	1	1	0.01	0.24
15	$4\frac{1}{3} - 1\frac{5}{3}$	1	1	1	1	0	0.01	0.18	1	**0**	1	1	0	0.01	0.19

The bold values are the changes.

the reduced reparametrized unified model (R-RUM; Hartz, 2002) were applied to fit the fraction-subtraction data with the expert or estimated Q-matrix. Under the DINA model, the means of the estimates of the guessing and slip parameter for the expert Q-matrix are 0.1080 and 0.1381, respectively, while for the revised Q-matrix, they are 0.1440 and 0.1295, respectively. It means that the estimates of the slip parameter become lower, but the guessing parameters tend to be larger. **Table 7** presents fit results for the fraction subtraction data using the expert and estimated q-matrix. The LLM with the estimated Q-matrix is the best-fitting CDM and the R-RUM with the estimated Q-matrix is slightly worse, whereas the estimated Q-matrix performed worse than the expert Q-matrix only in the DINA model.

CONCLUSION AND DISCUSSION

The supervised methods rely on a provisional Q-matrix for a whole test, the estimates of examinees' attribute patterns and their accuracy. It is not suitable for the case of a provisional Q-matrix with a large amount of misspecification. The purpose of this study is to propose the semi-supervised method under independent structure based on item responses and a reachability R matrix corresponding to a small part of test item specified by subject matter experts. The new method doesn't need to estimate examinees' attribute patterns. The main conclusion of this study is that the new method will play a very important role in assist

subject matter experts for Q-matrix specification because it is hard to correctly specify a Q-matrix with a large number of test items by subject matter experts. It may be useful for cognitive diagnostic assessment to facilitate teaching and learning.

The generalized Q-matrix theory has been shown that each column in the reduced Q-matrix can be expressed as a logical disjunction of some of columns of the reachability matrix. With the aid of this theory, this study takes a look inside a latent response matrix and reveals an interesting and useful relationship hidden in its columns. If a latent response matrix is calculated from a Q-matrix under the conjunctive model, a column in the latent response matrix is the conjunction of some other columns in this matrix if and only if the corresponding column of the Q-matrix can be written as the disjunction of their corresponding columns. While for the disjunctive model, the columns of the latent response matrix have exactly the same disjunction relationships as the columns of the Q-matrix. Because any conjunction or disjunction relationship among the columns of a latent response matrix would imply a disjunction relationship among the columns of a Q-matrix, then we are expected that the relationship between the columns in the Q-matrix can be constructed from the relationship between the corresponding columns in an observed response matrix, resulting from the latent response matrix by adding the noise or random errors. Another reason for this expectation is that each entry in the observed response matrix is modeled as a noisy observation of the corresponding entry in the latent

TABLE 7 | Fit results for the fraction subtraction data using the expert and estimated Q-matrix.

Q-matrix	CDM	−LL2	AIC	BIC
Expert Q-matrix	GDINA	6,695	7,133	8,071
Estimated Q-matrix	GDINA	6,548	6,910	7,686
Expert Q-matrix	DINA	6,912	7,034	7,295
Estimated Q-matrix	DINA	7,030	7,152	7,413
Expert Q-matrix	LLM	6,595	6,781	7,179
Estimated Q-matrix	LLM	6,523	6,707	7,102
Expert Q-matrix	R-RUM	6,696	6,882	7,280
Estimated Q-matrix	R-RUM	6,543	6,727	7,122

−2LL, −2 log likelihood; AIC, Akaike's information criterion; BIC, Bayesian information criterion (Chen et al., 2013).

response matrix through slip and guessing parameters (Junker and Sijtsma, 2001) and the discrepancies between the latent and observed response matrices are considered as random errors (Tatsuoka, 1987).

From the key theoretical results above, the semi-supervised method and an optimal design were then proposed for Q-matrix specification based on test response data and a reachability matrix specified by subject matter experts, and the simulation study was conducted to investigate the performance of the new method and the optimal design for examinee sampling in terms of the CRR of q-entries. From the CRR of q-entries, it is clear found that: (a) for the random design, when item parameters for items with known and unknown q-vectors are ≤0.20, the average of CRRs of q-entries for the semi-supervised method is larger than or equal to 0.9, (b) for the optimal design, when item parameters for items with known and unknown q-vectors are ≤0.25, the average of CRRs of q-entries for the semi-supervised method is larger than or equal to 0.9, and (c) item parameters of the reachability matrix have a larger impact on the performance of the semi-supervised method than item parameters with unknown q-vectors.

Finally, based on the results obtained in this study, some problems worthy of study in the future are put forward. First, how to effectively use the most of data or information on some other items for which experts have also specified q-vectors, because as the increase of the number of item specified q-vectors, the time complexity (more specifically, exponential time) of the exhaustive method grows much faster? If the number of items is increased to double or triple the number of attributes corresponding to the reachability matrix, one should investigate whether choosing a small part of items with high quality will reduce the noise of the responses and improve the estimation of q entries of unknown items. Second, in the simulation study, we know exactly how many attributes all items include. However, in the real situation, some items with unknown Q-matrix may mix additional attributes not specified in the reachability matrix because we haven't reviewed all items. Thus, we should explore a novel or revised method for identifying the possibility of extra attribute(s). Third, if the Q-matrix obtained from the semi-supervised method is taken as an initial matrix or a provisional Q-matrix of the existing supervised methods, is it possible to further improve the recovery of Q-matrix? From the results of the study, it can be seen that item parameters or random errors of item responses have an impact on the recovery of Q-matrix. If there is a method to reduce noise in item responses, the recovery of Q-matrix may be further improved. We only considered the small set of items with known q-vectors and fixed item parameters. Additional work is needed to further examine the impact of not only error patterns for known q-vectors but different item parameters for test items. Fourth, the current study focused on the DINA and DINO model only. In the future, the proposed method should be applied to general families of cognitive diagnostic models such as the generalized DINA model (de la Torre, 2011), the log-linear cognitive diagnostic model (Henson et al., 2009), the general diagnostic model (von Davier, 2008), testlet cognitive diagnosis model (Zhan et al., 2018), or polytomous cognitive diagnosis models (Chen and de la Torre, 2018; Ma, 2019). Lastly, since only the independent attribute structure in the simulation study and hierarchy structures for the conjunctive model in real data analysis were considered, the proposed method for other attribute hierarchies with different cognitive assumptions is worth studying.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.rdocumentation.org/packages/CDM/versions/7.4-19/topics/fraction.subtraction.data.

AUTHOR'S NOTE

Based on the fact that any columns of a reduced Q-matrix can be expressed by the columns of a reachability R matrix under the logical OR operation, a semi-supervised learning approach and an optimal design for examinee sampling were proposed for Q-matrix specification under the conjunctive and disjunctive model. This method only required subject matter experts specifying a R matrix corresponding to a small part of test items. Simulation and real data analysis showed that the new method with the optimal design is promising in terms of correct recovery rates of q-entries.

AUTHOR CONTRIBUTIONS

WW, LS, and TW conducted a design of the study, data analysis, paper writing, and revision. SD revised the paper. PG and JX give some descriptions of data analysis. All authors contributed to the article and approved the submitted version.

FUNDING

REFERENCES

Barnes, T. (2011). "Novel derivation and application of skill matrices: the q-matrix method," in *Handbook of Educational Data Mining,* eds C. Romero, Sebastian Ventura, M. Pechenizkiy, and R. S. J. D. Baker (Boca Raton, FL: CRC Press), 159–172.

Barnes, T. M. (2003). *The q-matrix method of fault-tolerant teaching in knowledge assessment and data mining* (Unpublished Doctoral dissertation). Raleigh, NC: North Carolina State University.

Beheshti, B., Desmarais, M., and Naceur, R. (2012). "Methods to find the number of latent skills," in *Proceedings of the 5th International Conference on Educational Data Mining* (Chania).

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika* 80, 1–20. doi: 10.1007/S11336-014-9401-5

Chen, J. (2017). A residual-based approach to validate q-matrix specifications. *Appl. Psych. Meas.* 41, 277–293. doi: 10.1177/0146621616686021

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

Chen, J.-S., de la Torre, J., and Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *J. Educ. Meas.* 50, 123–140. doi: 10.1111/j.1745-3984.2012.00185.x

Chiu, C.-Y. (2013). Statistical refinement of the q-matrix in cognitive diagnosis. *Appl. Psych. Meas.* 37, 598–618. doi: 10.1177/0146621613488436

Chiu, C.-Y., and Douglas, J. A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *J. Classif.* 30, 225–250. doi: 10.1007/s00357-013-9132-9

Close, C. N. (2012). *An exploratory technique for finding the q-matrix for the DINA model in cognitive diagnostic assessment: Combining theory with data* (Unpublished Doctoral dissertation). Minneapolis, MN: University of Minnesota, Educational Psychology.

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

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., and Chiu, C.-Y. (2016). A general method of empirical q-matrix validation. *Psychometrika* 81, 253–273. doi: 10.1007/s11336-015-9467-8

de la Torre, J., and Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* 69, 333–353. doi: 10.1007/BF02295640

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

DeCarlo, L. T. (2012). Recognizing uncertainty in the q-matrix via a bayesian extension of the DINA model. *Appl. Psych. Meas.* 36, 447–468. doi: 10.1177/0146621612449069

Desmarais, M. (2011). "Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization," in *The 4th International Conference on Educational Data Mining (EDM 2011),* eds M. Pechenizkiy, C. Conati, S. Ventura, C. Romero, and J. Stamper (Eindhoven).

Desmarais, M. C., Beheshti, B., and Naceur, R. (2012). "Item to skills mapping: deriving a conjunctive q-matrix from data" in *Intelligent Tutoring Systems Lecture Notes in Computer. Science,* Vol. 7315, eds S. A. Cerri, W. J. Clancey, G. Papadourakis, and K. Panourgia (Berlin; Heidelberg: Springer), 454–463.

Desmarais, M. C., Beheshti, B., and Xu, P. (2014). "The refinement of a q-matrix assessing methods to validate tasks to skills mapping," in *Proceedings of the 7th International Conference on Educational Data Mining*, eds J. Stamper, Z. Pardos, M. Mavrikis, and B. M. McLaren (London).

Desmarais, M. C., and Naceur, R. (2013). "A matrix factorization method for mapping items to skills and for enhancing expert-based q-matrices," in *Artificial Intelligence in Education (AIED 2013, LNAI 7926),* eds H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik (Heidelberg: Springer), p. 441–450.

Ding, S.-L., Luo, F., Yan, C., Lin, H.-J., and Wang, X.-B. (2008). "Complement to tatsuoka's q matrix theory," in *New Trends in Psychometrics,* eds K. Shigemasu, A. Okada, T. Imaizumi, and T. Hoshino (Tokyo: Universal Academy Press), 417–424.

Ding, S.-L., Zhu, Y.-F., Lin, H.-J., and Cai, Y. (2009). Modification of tatsuoka's q matrix theory. *Acta. Psy. Sinica* 41, 175–181. doi: 10.3724/SP.J.1041.2009.00175

Ding, S. L., Luo, F., Wang, W. Y., and Xiong, J. (2015). "Dichotomous and polytomous q matrix theory," in *Quantitative Psychology Research: The 80th Annual Meeting of the Psychometric Society,* eds L. A. V. D. Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, and M. Wiberg (Beijing: Springer).

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika* 49, 175–186. doi: 10.1007/BF02294171

Fischer, G. H. (1995). "The linear logistic test model," in *Rasch Models: Foundations, Recent Developments, and Applications,* eds G. H. Fischer and I. W. Molenaar (New York, NY: Springer-Verlag), 131–155.

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

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

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* 74, 191–210. doi: 10.1007/S11336-008-9089-5

Jang, E. E. (2009). Cognitive diagnostic assessment of l2 reading comprehension ability: validity arguments for fusion model application to languedge assessment. *Lang. Test.* 26, 31–73. doi: 10.1177/0265532208097336

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psych. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Kang, C., Yang, Y., and Zeng, P. (2018). Q-matrix refinement based on item fit statistic rmsea. *Appl. Psych. Meas.* 43, 527–542. doi: 10.1177/0146621618813104

Leighton, J. P., and Gierl, M. J. (eds.). (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications.* New York, NY: Cambridge University Press.

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

Liu, J., Xu, G., and Ying, Z. (2012). Data-driven learning of q-matrix. *Appl. Psych. Meas.* 36, 548–564. doi: 10.1177/0146621612461894

Liu, J., Xu, G., and Ying, Z. (2013). Theory of self-learning q-matrix. *Bernoulli* 19, 1790–1817. doi: 10.3150/12-BEJ430

Ma, W. (2019). A diagnostic tree model for polytomous responses with multiple strategies. *Br. J. Math. Stat. Psychol.* 72, 61–82. doi: 10.1111/bmsp.12137

Ma, W., and de la Torre, J. (2020). An empirical Q-matrix validation method for the sequential generalized DINA model. *Br. J. Math. Stat. Psychol.* 73, 142–163. doi: 10.1111/bmsp.12156

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

Rupp, A. A., and Templin, J. (2008). The effects of q-matrix misspecification on parameter estimates and classification accuracy in the dina model. *Educ. Psychol. Meas.* 68, 78–96. doi: 10.1177/0013164407301545

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

Tatsuoka, K. K. (1987). Bug distribution and statistical pattern classification. *Psychometrika* 52, 193–206. doi: 10.1007/BF02294234

Tatsuoka, K. K. (1990). "Toward an integration of item-response theory and cognitive error diagnosis," in *Diagnostic Monitoring of Skill and Knowledge Acquisition,* eds N. Frederiksen, R. L. Glaser, A. M. Lesgold, and M. G. Safto (Hillsdale, NJ: Erlbaum), p. 453–488.

Tatsuoka, K. K. (1995). "Architecture of knowledge structures and cognitive diagnosis: a statistical pattern classification approach," in *Cognitively Diagnostic Assessments,* eds P. D. Nichols, S. F. Chipman, and R. L. Brennan (Hillsdale: Erlbaum), p. 327–359.

Tatsuoka, K. K. (2009). *Cognitive Assessment: An Introduction to the Rule Space Method.* New York, NY: Taylor and Francis Group.

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

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

Wang, W., Gao, P., Song, L., and Wang, T. (2020). The improved exploratory method of q-matrix specification with noise preprocessing. *J. Jiangxi Normal Univ. (Nat. Sci.)* 44, 136–141.

Wang, W., Song, L., and Ding, S. (2018b). "An exploratory discrete factor loading method for q-matrix specification in cognitive diagnostic models," in *Quantitative Psychology. IMPS 2017. Springer Proceedings in Mathematics and Statistics, Vol. 233*, eds M. Wiberg, S. Culpepper, R. Janssen, J. González, and D. Molenaar (Cham: Springer).

Wang, W. Y., Song, L. H., Ding, S. L., Meng, Y. R., Cao, C. X., and Jie, Y. J. (2018a). An EM-based method for Q-matrix validation. *Appl. Psych. Meas.* 42, 46–459. doi: 10.1177/0146621617752991

Xu, P., and Desmarais, M. C. (2016). "Boosted decision tree for q-matrix refinement," in *Proceedings of the 9th International Conference on Educational Data Mining*, eds T. Barnes, M. Chi, and M. Feng (Raleigh, NC).

Xu, P., and Desmarais, M. C. (2018). "An empirical research on identifiability and q-matrix design for DINA model," in *11th International Conference on Educational Data Mining (EDM)*, eds K. E. Boyer and M. Yudelson (Raleigh, NC).

Zhan, P., Liao, M., and Bian, Y. (2018). Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Front. Psychol.* 9:607. doi: 10.3389/fpsyg.2018. 00607

Zhan, P., Ma, W., Jiao, H., and Ding, S. (2020). A sequential higher order latent structural model for hierarchical attributes in cognitive diagnostic assessments. *Appl. Psych. Meas.* 44, 65–83. doi: 10.1177/0146621619832935