Article

Applied Psychological Measurement
2018, Vol. 42(6) 446–459
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621617752991
journals.sagepub.com/home/apm
SAGE

An EM-Based Method for Q-Matrix Validation

Wenyi Wang[1], Lihong Song[1], Shuliang Ding[1], Yaru Meng[2], Canxi Cao[3], and Yongjing Jie[3]

Abstract

With the purpose to assist the subject matter experts in specifying their Q-matrices, the authors used expectation–maximization (EM)–based algorithm to investigate three alternative Q-matrix validation methods, namely, the maximum likelihood estimation (MLE), the marginal maximum likelihood estimation (MMLE), and the intersection and difference (ID) method. Their efficiency was compared, respectively, with that of the sequential EM-based δ method and its extension (ς^2), the γ method, and the nonparametric method in terms of correct recovery rate, true negative rate, and true positive rate under the deterministic-inputs, noisy "and" gate (DINA) model and the reduced reparameterized unified model (rRUM). Simulation results showed that for the rRUM, the MLE performed better for low-quality tests, whereas the MMLE worked better for high-quality tests. For the DINA model, the ID method tended to produce better quality Q-matrix estimates than other methods for large sample sizes (i.e., 500 or 1,000). In addition, the Q-matrix was more precisely estimated under the discrete uniform distribution than under the multivariate normal threshold model for all the above methods. On average, the ς^2 and ID method with higher true negative rates are better for correcting misspecified Q-entries, whereas the MLE with higher true positive rates is better for retaining the correct Q-entries. Experiment results on real data set confirmed the effectiveness of the MLE.

Keywords

cognitive diagnosis, Q-matrix, EM algorithm, DINA model, reduced RUM, fraction-subtraction data

In educational assessment, cognitive diagnostic assessment (CDA) that combines psychometrics and cognitive science has received increased attention (Leighton & Gierl, 2007; Rupp, Templin, & Henson, 2010; K. K. Tatsuoka, 2009). This approach potentially provides useful diagnostic information regarding students' strengths and weaknesses, and can facilitate individualized learning (Chang, 2015; Chang & Wang, 2016). Cognitive diagnostic models (CDMs) often

[1] Jiangxi Normal University, Jiangxi, China
[2] Xi'an Jiaotong University, Shaanxi, China
[3] University of Illinois at Urbana–Champaign, IL, USA

Corresponding Author:
Wenyi Wang, School of Computer and Information Engineering, Jiangxi Normal University, 99 Ziyang Avenue, Nanchang, Jiangxi 330022, China.
Email: wenyiwang@jxnu.edu.cn

utilize a Q-matrix (Embretson, 1984; K. K. Tatsuoka, 1990, 1995, 2009), whose entries are 1 or 0, $q_{jk} = 1$ meaning that attribute k is involved in correctly answering item j, otherwise, $q_{jk} = 0$.

Correct specification of the Q-matrix is a fundamental step to guarantee the test validity for CDA (McGlohen & Chang, 2008; Im & Corter, 2011). Its procedure is usually an iterative process (Buck et al., 1998; Jang, 2009): (a) The provisional Q-matrix is primarily exploratory based on a current related theory, subject matter experts' judgment, and item analysis and (b) the modified Q-matrix is primarily confirmatory which is based on statistical methods. The above two steps represent the qualitative and quantitative methods, respectively, and either of them alone is not enough to guarantee the correctness of a Q-matrix.

In practice, expert judgment may introduce some uncertain elements into the provisional Q-matrix (DeCarlo, 2012), making it difficult to specify correctly in CDA (DeCarlo, 2011; Jang, 2009). Previous studies have shown that even a small amount of Q-matrix misspecification could degrade the precision of estimated item parameters, resulting in the decrease in the classification accuracy of CDMs (Baker, 1993; Rupp & Templin, 2008; Im & Corter, 2011). To improve the quality of a Q-matrix, researchers have proposed several quantitative methods for Q-matrix validation, such as the sequential expectation–maximization (EM)–based δ method or δ method (de la Torre, 2008) and its extension ς^2 method (de la Torre & Chiu, 2016; Huo & de la Torre, 2013), the γ method (Tu, Cai, & Dai, 2012), the Bayesian approach (DeCarlo, 2012), the data-driven approach (Liu, Xu, & Ying, 2012, 2013), the nonparametric Q-matrix refinement method (Chiu, 2013), and the stepwise reduction algorithm (Hartz, 2002).

The primary advantage of those methods is that they can incorporate expert's Q-matrix and item response data into Q-matrix validation. However, its disadvantages are very obvious. Five of them are as follows: (a) The δ, ς^2, and γ methods rely on particular cutoff values. In fact, different values should be assigned to items with different number of attributes or item quality (de la Torre & Chiu, 2016; Huo & de la Torre, 2013); (b) the performance of the γ method is not very satisfactory when the number of attributes required in correctly solving an item is above three, which is illustrated in the simulation study below; (c) the Bayesian approach requires that the uncertain entries in the Q-matrix should be identified in advance. It could also be used in a more exploratory manner; however, the robustness of this method remains to be explored (DeCarlo, 2012); (d) the data-driven approach (Liu et al., 2012, 2013) is not easy to compute when the number of items and/or the number of attributes is large (Chiu, 2013), though it could be used without experts' Q-matrix; and (e) the nonparametric method is preferred when the underlying model is unknown, but sometimes it would be less efficient than the parametric method if the underlying model fits the data. In spite of their respective limitations, all the above methods are being widely used.

The traditional methods like the maximum likelihood estimation (MLE) and the marginal maximum likelihood estimation (MMLE) were proposed to estimate the q-vector of an item because it is very similar to estimating attribute pattern of an examinee. Several online calibration approaches, including the MLE-based or MMLE-based method (Y. Chen, Liu, & Ying, 2015; P. Chen & Wang, 2016; P. Chen, Xin, Wang, & Chang, 2010, 2012; Wainer & Mislevy, 1990; W. Y. Wang, Ding, & You, 2011), have been proposed to estimate item parameters or the q-vector in computerized adaptive testing (CAT) or cognitive diagnostic computerized adaptive testing (CD-CAT). One important distinction between Q-matrix validation and online Q-matrix calibration is that the former is often applied to refine a provisional Q-matrix, whereas the latter is usually used to estimate the q-vectors of raw items. Whereas the fact is that little is known about how to extend the traditional online calibration methods to Q-matrix validation, nor are there enough related studies comparing the current Q-matrix validation methods. With this gap in mind, the authors in this article explore an EM-based approach to assist subject matter experts in specifying their Q-matrices. The authors propose three alternative Q-matrix validation methods based on their previous study (W. Y. Wang, Ding, & Song, 2013), none of which need to set cutoff values or give possible misspecified Q-entries.

An EM-Based Approach

The authors intend to make a comparison among the existing methods and the newly developed ones. As most of the existing methods are based on the deterministic-inputs, noisy ''and'' gate (DINA) model (Junker & Sijtsma, 2001) and the reduced reparameterized unified model (rRUM; Hartz, 2002), this article also chose these two models. Let X_{ij} be the response of examinee i to item j, $i = 1, \ldots, N$, $j = 1, \ldots, J$. Let $\boldsymbol{\alpha}_\mathbf{i}$ be examinee i attribute pattern. Let $\boldsymbol{q}_j$ be the q-vector of item j in the Q-matrix. The item response function for the DINA model is as follows:

$$P_j(\boldsymbol{\alpha}_\mathbf{i}) = P\left(X_{ij} = 1|\boldsymbol{\alpha}_\mathbf{i}\right) = g_j^{1-\eta_{ij}}\left(1 - s_j\right)^{\eta_{ij}}, \tag{1}$$

where $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ is an ideal latent response, and s_j and g_j are slipping and guessing parameters of item j.

The item response function for the rRUM is as follows:

$$P_j(\boldsymbol{\alpha}_\mathbf{i}) = P\left(X_{ij} = 1|\boldsymbol{\alpha}_\mathbf{i}\right) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*(1-\alpha_{ik})q_{jk}}, \tag{2}$$

where the baseline parameter π_j^* is the probability of correct response to item j given that an examinee has mastered all the required attributes for the item, and the probability of correct response to item j is proportional to the penalty parameters r_{jk}^* when an examinee has not mastered attribute k.

The three methods proposed for Q-matrix validation in this study are the MLE, the MMLE, and an intersection and difference (ID) estimation. The ID method is based on two-set operations (intersection and difference) in terms of set theory. Before estimating the uncertain q-vector for an item, a parameter space or a reduced matrix $\boldsymbol{Q}_r$, which represents a set of potential q-vectors, must be specified. If K attributes are independent, the rows of $\boldsymbol{Q}_r$ consist of all combinations of attributes and $\boldsymbol{Q}_r$ contains $2^K - 1$ rows. For estimating examinees' attribute patterns, let $\boldsymbol{Q}_s$ be the universal set of attribute patterns (K. K. Tatsuoka, 1995), which consists of $\boldsymbol{Q}_r$ with an added zero vector.

Given a provisional Q-matrix and response data, the MLE, MMLE, and ID methods can be friendly implemented in the EM algorithm (de la Torre, 2009; Feng, Habing, & Huebner, 2014) by considering q-vectors as item-specific parameters to be estimated. At cycle t of the EM algorithm, let $\hat{\boldsymbol{\beta}}_j^{(t)} = (\hat{s}_j^{(t)}, \hat{g}_j^{(t)})$ and $\hat{\boldsymbol{\beta}}_j^{(t)} = (\hat{\pi}_j^{*(t)}, \hat{\boldsymbol{r}}_j^{*(t)})$ be the item parameter estimates of the DINA model and rRUM, respectively. $\hat{\boldsymbol{\alpha}}_i^{(t)}$ denotes the examinee i's attribute pattern estimate. $\hat{\boldsymbol{q}}_j^{(t)}$ represents the q-vector estimate. The initial value of the q-vector for item j, denoted by $\boldsymbol{q}_j^{(0)}$, is the jth row of the provisional Q-matrix. Four steps are involved in these methods:

Step 1. Obtain a provisional Q-matrix from subject matter experts.

Step 2. Run one EM cycle to estimate the item parameters, the examinees' attribute patterns, and their distributions.

Step 3. Estimate the q-vector for each item via the MLE, MMLE, or ID method, and then update the Q-matrix. The three methods differ in Step 3. Next, the authors will give a detailed description of the MLE, MMLE, and ID methods, followed by the comparison in between.
Step 4. Repeat the second and third step until the convergence criterion is satisfied (when the discrepancy of the relative log-marginalized likelihood between the previous and current estimate is smaller than 0.001, the convergence is reached).

MLE Method

At cycle t of EM algorithm, item parameter $\hat{\boldsymbol{\beta}}_j^{(t)}$ and examinee attribute pattern $\hat{\boldsymbol{\alpha}}_i^{(t)}$ are estimated. Assuming local independence exists, to obtain vector $\boldsymbol{q}_j$, the authors replace $\hat{\boldsymbol{\alpha}}_i^{(t)}$ and $\hat{\boldsymbol{\beta}}_j^{(t)}$ in the following equation with their newly estimated values:

$$\hat{\boldsymbol{q}}_j^{(t)} = \arg\max_{q_h \in Q_r} L\left(\boldsymbol{X}_j | \hat{\boldsymbol{\alpha}}_i^{(t)}, \hat{\boldsymbol{\beta}}_j^{(t)}, \boldsymbol{q}_h\right), \tag{3}$$

where $L(\boldsymbol{X}_j|\hat{\boldsymbol{\alpha}}_i^{(t)}, \hat{\boldsymbol{\beta}}_j^{(t)}, \boldsymbol{q}_h) = \prod_{i=1}^{N} P_{jh}(\hat{\boldsymbol{\alpha}}_i^{(t)})^{X_{ij}}(1 - P_{jh}(\hat{\boldsymbol{\alpha}}_i^{(t)}))^{1-X_{ij}}$ and $P_{jh}(\hat{\boldsymbol{\alpha}}_i^{(t)})$ is the item response function of the DINA model or the rRUM. For item j, let $r_{cj}^{(t)}$ and $w_{cj}^{(t)}$ correspond to the correct and incorrect frequency of all examinees with an attribute pattern $\boldsymbol{\alpha}_c$, respectively. Taking the natural logarithm of Formula 3 yields the following:

$$\hat{\boldsymbol{q}}_j^{(t)} = \arg\max_{q_h \in Q_r} \sum_{\alpha_c \in Q_s} \left(r_{cj}^{(t)} \ln P_{jh}(\boldsymbol{\alpha}_c) + w_{cj}^{(t)} \ln\left(1 - P_{jh}(\boldsymbol{\alpha}_c)\right)\right). \tag{4}$$

MMLE Method

In traditional item response theory, poorly calibrated items may result in an inaccurate estimation of the latent trait. For the purpose of a precise item parameter calibration, the MMLE method should take account of measurement errors derived from the latent trait estimates (Wainer & Mislevy, 1990), which can be effectively addressed, whereas the MLE method ignores these errors. The former has been widely used for the partial Bayesian models, and it only places a prior distribution on the examinees' population parameters (DiBello, Roussos, & Stout, 2007). In the MMLE, posterior distribution $\boldsymbol{\pi}(\boldsymbol{\alpha}_c|\boldsymbol{X}_i, \hat{\boldsymbol{\beta}}_j^{(t)}, \hat{\boldsymbol{q}}_j^{(t-1)})$ is used to estimate the q-vector. Assuming local independence exists, the marginal likelihood function can be written as follows:

$$\hat{\boldsymbol{q}}_j^{(t)} = \arg\max_{\boldsymbol{q}_h \in \boldsymbol{Q}_r} \prod_{i=1}^{N} L\left(X_{ij} | \hat{\boldsymbol{\beta}}_j^{(t)}, \boldsymbol{q}_h\right), \tag{5}$$

where $\mathrm{L}(X_{ij}|\hat{\boldsymbol{\beta}}_j^{(t)}, \boldsymbol{q}_h) = \sum_{\boldsymbol{\alpha}_c \in \boldsymbol{Q}_s} (L(X_{ij}|\boldsymbol{\alpha}_c, \boldsymbol{\beta}_j^{(t)}, \boldsymbol{q}_h)\boldsymbol{\pi}(\boldsymbol{\alpha}_c|\boldsymbol{X}_i, \boldsymbol{\beta}_j^{(t)}, \hat{\boldsymbol{q}}_j^{(t-1)}))$. Note that the number or value of each item parameter varies with different CDM Q-matrices. For the DINA model, only values of item parameters change. While under the rRUM, the following issues should be considered: First, the number of item parameters will vary with different q-vectors; second, the rRUM with q-vector as an additional parameter is nonidentifiable. A solution is needed to deal with these issues. Because the MLE and MMLE method need to search all the possible q-vectors and require item parameters to estimate the q-vector of item j, the initial value of $\hat{r}_{jk}^{*(t-1)}$

for $\hat{\boldsymbol{q}}_{jk}^{(t-1)} = 0$ is generated uniformly between 0.1 and 0.6 before estimating the $\hat{\boldsymbol{q}}_j^{(t)}$ vector at each cycle of EM algorithm. Given $\hat{r}_{j1}^{*(t-1)}, \hat{r}_{j2}^{*(t-1)}, \ldots, \hat{r}_{jK}^{*(t-1)}$ for all attributes, the $\hat{\boldsymbol{q}}_j^{(t)}$ can be obtained by calculating the (marginal) likelihood function. An identifiability restriction is imposed by putting the upper bound of r_{jk}^* equal to 0.6 in order that the rRUM model with a q-vector parameter can be identifiable. It is easy to show that item response function for $\hat{r}_{jk}^* = 1$ and $q_{jk} = 1$ is equivalent to $q_{jk} = 0$, because in these two cases, the term $r_{jk}^{*(1-\alpha_{ik})q_{jk}}$ in the rRUM satisfied $1^{(1-\alpha_{ik})\times 1} = r_{jk}^{*(1-\alpha_{ik})\times 0} = 1$ for each α_{ik}. For example, for a hypothetical item with $\hat{\boldsymbol{q}}_j = (1, 0), \hat{\pi}_j^* = 0.8$, and $\hat{r}_{j1}^* = 0.6$, the probabilities of correct response for four attribute patterns, such as $(1, 1)$, $(1, 0)$, $(0, 1)$, and $(0, 0)$, are 0.8, 0.8, 0.48, and 0.48, respectively. If $\hat{\boldsymbol{q}}_j = (1, 1), \hat{\pi}_j^* = 0.8$, $\hat{r}_{j1}^* = 0.6$, and $\hat{r}_{j2}^* = 1$, the probabilities do not change. The upper bound is set equal to 0.6 partially because if r_{jk}^* is close to 0.9, the corresponding attribute should be eliminated for item j (Hartz, 2002).

ID Method

Based on the definition of noncompensatory/conjunctive (DiBello et al., 2007), both the DINA model and rRUM are categorized into special cases of this model, as you can see in original Tables 4 and 5 in DiBello et al. (2007). For the noncompensatory/conjunctive model (DiBello et al., 2007), it is often reasonable to assume that if most examinees with an attribute pattern solved item j correctly, the set of attributes required by item j is a subset of the mastery attribute set. The implementation of the algorithm for the ID method is based on the assumptions mentioned above and the lattice theory (Rosen, 2011). The theoretical foundation and two artificial examples for the ID method are described in Online Appendix A.

At cycle t of the EM algorithm, let $n_{cj}^{(t)} = r_{cj}^{(t)} + w_{cj}^{(t)}$, $p_{cj}^{(t)} = r_{cj}^{(t)} / n_{cj}^{(t)}$, where $r_{cj}^{(t)}$ and $w_{cj}^{(t)}$ are defined as described above. If $p_{cj}^{(t)} > 1 - p_{cj}^{(t)}$, it implies that $\boldsymbol{q}_j$ belongs to $\boldsymbol{L}_{\boldsymbol{\alpha}_\mathrm{c}}$, where $\boldsymbol{L}_{\boldsymbol{\alpha}_\mathrm{c}} = \{\boldsymbol{q}_t | \forall \boldsymbol{q}_t \in \boldsymbol{Q}_\mathrm{r} \text{ and } \boldsymbol{q}_t \leq \boldsymbol{\alpha}_c\}$ represents the lower bound of $\boldsymbol{\alpha}_c$ in terms of the lattice theory. That is, $\boldsymbol{L}_{\boldsymbol{\alpha}_\mathrm{c}}$ is a set of all possible q-vectors of item j which can be responded correctly by an examinee with $\boldsymbol{\alpha}_c$ without slipping. On the contrary, if $p_{cj}^{(t)} < 1 - p_{cj}^{(t)}$, it means that $\boldsymbol{q}_j$ does not belong to $\boldsymbol{L}_{\boldsymbol{\alpha}_\mathrm{c}}$. Four steps of the ID method are as follows:

Step 1. Let the candidate set $\boldsymbol{C}_j$ of $\boldsymbol{q}_j$ be $\boldsymbol{Q}_{\boldsymbol{r}}$ for item j.
Step 2. Sort $\mathrm{Ratio}_{\mathrm{cj}} = \max(p_{cj}^{(t)}/(1 - p_{cj}^{(t)}), \left(1 - p_{cj}^{(t)}\right)/p_{cj}^{(t)})(\boldsymbol{\alpha}_\mathrm{c} \in \boldsymbol{Q}_s)$ in descending order for item j.
Step 3. Select the attribute pattern with large $\mathrm{Ratio}_{\mathrm{cj}}$ sequentially to adjust the candidate set $\boldsymbol{C}_j$. If $p_{cj}^{(t)} > 1 - p_{cj}^{(t)}$, then let the candidate set $\boldsymbol{C}_j$ be the intersection of $\boldsymbol{C}_j$ and $\boldsymbol{L}_{\boldsymbol{\alpha}_\mathrm{c}}$(i.e., $\boldsymbol{C}_j = \boldsymbol{C}_j \bigcap \boldsymbol{L}_{\boldsymbol{\alpha}_\mathrm{c}}$); otherwise, let the candidate set $\boldsymbol{C}_j$ be the difference of $\boldsymbol{C}_j$ and $\boldsymbol{L}_{\alpha_c}$ (i.e., $\boldsymbol{C}_j = \boldsymbol{C}_j - \boldsymbol{L}_{\alpha_c}$), meaning that $\boldsymbol{C}_j$ is the intersection of $\boldsymbol{C}_j$ and the up-set of element $\boldsymbol{\alpha}_\mathrm{c}$ (C. Tatsuoka, 2002).
Step 4. Repeat Step 3 until $\boldsymbol{C}_j$ contains only one element or the iteration number of Step 3 is equal to the number of elements in the universal set of attribute patterns. If $\boldsymbol{C}_j$ contains more than one element, then let $\left\{\boldsymbol{q}_j\right\}$ be equal to a zero vector; otherwise, let $\left\{\boldsymbol{q}_j\right\}$ be equal to $\boldsymbol{C}_j$. The zero vector means that it is difficult to choose a best estimate by the algorithm. In practice, although $\boldsymbol{C}_j$ contains more than one element, these results at a coarser grain level can also provide information that helps experts specify the q-vector.

Similarities and Differences Between These Three Methods

All the three methods can successfully estimate the q-vector based on examinees' attribute patterns. For the MLE method, the whole estimation process can be regarded as a joint maximum likelihood estimation (JMLE) method, while for the MMLE, it sometimes could be thought as empirical Bayes (Casella, 1985; Ivezic, Connolly, VanderPlas, & Gray, 2014). It is important to note that the inherent drawback of JMLE is that the estimators of the item parameters are not statistical consistent (de la Torre, 2009). However, JMLE can be applied to estimate item parameters and attribute patterns very effectively under various CDMs, including the DINA model, the rRUM, the deterministic input noisy or model, and the noisy input, deterministic and model (Y. Chen et al., 2015; Zheng, Chiu, & Douglas, 2015). The ID method can be regarded as a nonparametric method because it requires the estimation of examinees' attribute patterns only, whereas the MLE and MMLE methods require a parametric model for calculating the likelihood function.

Simulation Study

Simulation Design

To investigate whether these methods can work under certain conditions, simulated data were generated using five attributes. In the simulation study, the correct Q-matrix was fixed as the reduced Q-matrix with 31 items including all the possible nonzero q-vectors to examine the robustness of the estimation methods. This Q-matrix, with an identity or a reachability (***R***) matrix, was called a complete Q-matrix (Chiu, Douglas, & Li, 2009) or a necessary and sufficient Q-matrix (Ding, Yang, & Wang, 2010). Because the complete Q-matrix can distinguish all the ideal item response patterns, the correct classification rate can thus be improved.

Four important factors were included in the design of the simulation study under the DINA model or the rRUM: (a) the source of the examinees' attribute patterns (discrete uniform distribution and multivariate normal threshold model), (b) the number of examinees (N = 300, 500, and 1,000), (c) the quality of items (items with s, g~U(0.05, 0.25) or π^*~U(0.8, 0.98) and r^*~U(0.1, 0.6) were labeled high quality; items with s, g~U(0.05, 0.4) or π^*~U(0.75, 0.95) and r^*~U(0.2, 0.95) were labeled low quality), and (d) the percentage of misspecified Q-entries (0%, 10%, 20%, 30%, and 40%). Item quality in this study was defined as the average of the discriminating powers of items in a test (Cui, Gierl, & Chang, 2012) or item parameters (Ma, Iaconangelo, & de la Torre, 2016). In practice, item quality would be defined in terms of both discriminating power and coverage of the content specifications (Xing & Hambleton, 2004). In general, for the DINA model, a high-quality or "good" item will have small slipping and guessing parameters (Rupp et al., 2010), which means that the item discrimination powers are large (Cui et al., 2012). For the rRUM, a high-quality or "good" item will have a high π_j^* and low r_{jk}^* parameters (Rupp et al., 2010). A total of 60 conditions were simulated (2 correlations ×3 sample sizes ×2 item parameters ×5 misspecification). Two hundred replication data sets were simulated for each condition.

Simulation Data

For the discrete uniform distribution, attribute patterns were generated to take each of the 2^5 possible patterns with equal probability for each sample size. Attribute patterns were generated from the multivariate normal threshold model with all the means equal to 0, all the variances and covariances in the variance–covariance matrix equal to 1.00 and 0.50, respectively, following the process used in Chiu et al. (2009). Moreover, the correlation coefficient (ρ) between any pair of attributes is equal to 0.50. Item parameters were randomly generated across replications. The correct Q-matrix was used to generate the item responses based on the DINA model and the rRUM.

Random errors were added to the correct Q-matrix (i.e., error-free) by randomly changing a specified percentage of the elements. The percentage of the elements needed for the change was consistent with the error rates (from 0 to 0.4 with Step 0.1), so the number of elements changed in the correct Q-matrix varies from 0 to 62 equal to the error rates × the number of items (31) × the number of attributes (5). A computer program was designed to achieve this by first selecting an item and an attribute at random, and then reversing the current value of that cell (0 to 1 or 1 to 0) in the Q-matrix (Baker, 1993). The constraints imposed on the generation of the error Q-matrix were that each attribute was at least measured by one item, and each item measured at least one attribute. The provisional Q-matrices for each error rate thus resulted.

Methods and Evaluation Criteria

A computer program based on the EM algorithm (de la Torre, 2009; Feng et al., 2014) was written in MATLAB 2008. For each data set, the performance of three new methods under the DINA model was compared with the γ method, the δ method, and Chiu's nonparametric method. In the pilot study, seven cutoff values ($\varepsilon = 0, 0.01, 0.05, 0.10, 0.20, 0.25$, and 0.30) in the δ method were used to select the candidate q-vectors. For similar simulation conditions as described above, the results indicated that the cutoff value ε between 0.10 and 0.20 could be regarded as a reasonable value (see Table B1 in Online Appendix B). Therefore, only one cutoff value ($\varepsilon = 0.20$) was used in the following simulation study. In the rRUM, the new methods were only compared with Chiu's nonparametric method and the ς^2 method because the γ method and the δ method cannot be implemented. The nonparametric method relies on the ideal response pattern which is computed following the method proposed by Chiu and Douglas (2013). For the ς^2 method, based on earlier work by Huo and de la Torre (2013), two cutoff values of 0.005 ($N = 300$) and 0.0025 ($N = 500$ and 1,000) were applied.

The results reported in this study focused on the Q-matrix estimate, because it was directly related to the performance of each method. The correct recovery rate (CRR) is equal to the ratio of the number of correct Q-entries in the estimated Q-matrix to the total number of Q-entries (Chiu, 2013). For each condition, the mean and standard deviation of the CRR values of the 200 replications were reported for each method. In addition, the authors are interested in whether the differences of the largest mean values of CRR were statistically significant from the others. It should be noted that if their means were almost the same or statistically insignificantly different, there would be several promising candidates. In this case, it was advised to perform a paired t test with the null hypothesis H_0 that the mean of the differences between the largest and other values of CRRs is equal to zero, against the alternative hypothesis H_1 that the null is false. The null hypothesis was tested at the 5% level of significance. The results of test hypotheses were given in Tables B2 and B3 in Online Appendix B. The results of the paired t test were very similar to that of the Wilcoxon signed rank test. To obtain insight into the performance of these methods in two different aspects, the true positive and true negative rates of Q-entries were presented, which were also used for evaluating the performance of the ς^2 method (de la Torre & Chiu, 2016). The true positive rate indicates the proportion of correctly specified Q-entries that was retained. The true negative rate indicates the proportion of misspecified Q-entries correctly estimated.

Results

The EM-based algorithm in the majority conditions had achieved convergence when the criterion was pegged at 0.001. The means of the CRR values of 200 replications for each method are shown in Tables B2 and B3, in which the largest CRRs were highlighted in boldface and nonsignificant CRRs associated with the largest rates in boldface italics. For all conditions, the distribution of the standard deviation of the CRR had a mean of 0.04 and a standard deviation of 0.03 (minimum = 0, maximum = 0.12). Detailed results of standard deviations are available by contacting the first author.

The impact of the source for attribute patterns. Tables B2 and B3 show that the quality of the provisional Q-matrix was improved. The Q-matrix was more precisely estimated under the discrete uniform distribution than under the realistic multivariate normal threshold model. This finding was aligned with the results by Chiu (2013). One reason for this result is that some attribute patterns contained too few examinees under multivariate normal threshold model to identify some misspecified q-vectors, noticing that if $\rho = 0.5$ was positive, then an individual with a specific attribute was more likely to have mastered the second attribute; the other reason is that the prior distribution only matched the discrete uniform distribution.

The impact of sample size. For three sample sizes, Tables B2 and B3 show the accuracy of Q-matrix estimates. In comparison, they showed a clear improvement from the sample size of 500 to 1,000. For the larger sample size (1,000), the mean of the CRR in many cases was above 0.9 for the MMLE and MLE methods within a low (0.1) or moderate (0.2) degree of Q-matrix misspecification. However, increasing sample size hardly improved the accuracy of Q-matrix estimates when the degree of the Q-matrix misspecification was high.

The impact of item parameter and Q-matrix misspecification. The smaller slipping and guessing parameters or penalty parameters corresponded to the better performance of Q-matrix estimates. This is because those smaller values contribute to higher correct classification rates for attribute patterns. The Q-matrix was better recovered from a slight or moderate misspecification than a serious one. Table B2 illustrates that CRRs decreased dramatically when a larger degree of Q-matrix misspecification (e.g., 0.30 or 0.40) was involved.

The impact of the items with different numbers of required attributes. To demonstrate the possible effects of the number of attributes required for an item on Q-matrix estimation, Figure B1 in Online Appendix B shows the mean of CRRs regarding different numbers of attributes. For all methods except the ID method, the accuracy decreased as the number of required attributes increased. It is expected that the performance of the γ method under the DINA model was unacceptable when the number of attributes measured by an item was above three. For other methods under the DINA model, there were relatively small differences across the different numbers of attributes. Similar results were obtained in the rRUM.

The impact of the estimation method. Results of the six methods compared under the DINA model are shown in Table B2. According to the hypothesized test results (see column 12 in Table B2), the MMLE method yielded more significantly accurate Q-matrix estimates than other methods when the data were from the multivariate normal threshold model. The average CRRs across 30 conditions for the γ, MLE, ID, MMLE, δ, and nonparametric methods were 0.7557, 0.8225, 0.7978, 0.8312, 0.7272, and 0.8245, respectively. However, when the data followed the discrete uniform distribution, the corresponding average CRRs were, respectively, 0.7900, 0.8772, 0.8976, 0.8925, 0.8678, and 0.8895, for the six methods. These results indicated that, the ID method, on average, produced better quality Q-matrix estimates than other methods, particularly when the sample size was large (i.e., 500 or 1,000). The table also shows that the nonparametric method resulted in better Q-matrix estimates only when 10% or 20% of the entries in a provisional Q-matrix were randomly changed.

For the results under the rRUM, see Table B3. It shows that (a) the MMLE method outperformed other methods when the data were generated from the multivariate normal threshold model and the quality of items was high (π^*~U(0.8, 0.98) and r^*~U(0.1, 0.6)). The MLE method was only slightly inferior to the MMLE method (i.e., 0.8607 vs. 0.8653 for the average CRRs across 15 conditions); (b) the MLE method introduced more precise Q-matrix estimates than other methods when the quality of items was low (π^*~U(0.75, 0.95) and r^*~U(0.2, 0.95)), regardless of attribute pattern distributions. Take the multivariate normal threshold model as an example, the average CRRs across 15 conditions for the MLE, ID, MMLE, nonparametric, and ς^2 methods were 0.8129, 0.6863, 0.8038, 0.7720, and 0.7829; (c) the ς^2 method performed best with an average CRR of 0.9257 under conditions of large sample sizes (i.e., 500 or 1,000), high quality of items, and discrete uniform distribution; and (d) the ID and nonparametric methods did not perform well under the rRUM, consistent with the theorem results (S. Wang & Douglas, 2015), because the simulated penalty parameters of some items did not satisfy the condition (a.3) specified by S. Wang and Douglas (2015). Tables B4 and B5 in Online Appendix B provide the true positive and true negative rates of Q-entries. An interesting result was found that the ς^2 and ID methods have higher true negative rates, and the MLE method has higher true positive rates.

Real Data and Analysis

The performance of the above Q-matrix validation methods was examined through real data analysis. These methods are applied to the fraction-subtraction data set (K. K. Tatsuoka, 1990; C. Tatsuoka, 2002), which consists of 536 examinees. The Q-matrix, which consists of 15 items, is the same as the one used by de la Torre (2008, see original Table 7) and DeCarlo (2012, see original Table 7). The labels of the attributes are (a) performing a basic fraction-subtraction operation, (b) simplifying/reducing, (c) separating whole numbers from fractions, (d) borrowing one from a whole number to a fraction, and (e) converting whole numbers to fractions. Based on existing results (de la Torre, 2008; Huo & de la Torre, 2013), a small cutoff value (ε or ϵ) of 0.005 was used for the δ and ς^2 methods, respectively.

The DINA model and the rRUM were used to analyze the data. For the DINA model, the initial values of item parameters were randomly drawn from U(0.05, 0.4). For the rRUM, the initial values of item parameters were randomly drawn from π^*~U(0.8, 0.98) and r^*~U(0.1, 0.9). Due to the relatively small sample size, different initial item parameters led to slight different estimates of the item parameters, Q-matrix, and fit indices (–2 log likelihood [–2LL], Akaike information criterion [AIC], and Bayesian information criterion [BIC]). Thus, each of the methods was run for 200 times to compute the average of the fit index (Kang & Cohen, 2007) and the Q-matrix. Table B6 in Online Appendix B shows the average of the fit indices. The –2LL, AIC, and BIC were smaller for the Q-matrix obtained from the MLE method or the ς^2 method, indicating better relative fit than that of the other Q-matrices. Figure B2 in Online Appendix B shows the distribution of –2LL for all replications. Judging by the figures, one can easily see that the MLE produced the lowest –2LL in most replications.

Table B7 in Online Appendix B shows the modified Q-matrix from the MLE method. Attributes α_2 and α_4 should be included in Item 2, and attributes α_1 and α_3 should be dropped from Item 2. This result might be useful to subject experts. For example, one can only simplify $2\frac{3}{2}$ to $3\frac{1}{2}$, and get a correct answer with high probability by knowing that the minuend and subtrahend are equal. Similarly, one can borrow one from the minuend, and then get an answer of zero. In other words, an examinee mastering attributes α_2 and/or α_4 can still have a very high probability of correct response. For Item 5, the proposed q-vector is the same as the original q-vector. A strategy to solve the problem in Item 5 could be $3\frac{7}{8}-2\xrightarrow{\alpha_3}3-2+\frac{7}{8}\xrightarrow{\alpha_1}1\frac{7}{8}$. The results of Items 10 and 14 are consistent with the results reported by DeCarlo (2012). Although the MLE method provided some interesting results for the Q-matrix validation as shown above, statistical methods could result in false suggestions because of noise in data. For example, the results of the MLE method indicated that attribute α_5 is necessary for Items 12 and 15, which actually is not the case. A better strategy is to discuss with domain experts, as suggested by de la Torre (2008).

Conclusion and Discussion

In conclusion, this study introduced three methods for validating the Q-matrix given the provisional Q-matrix and response data. Simulation results showed that these methods exhibit varying degrees of effectiveness in terms of CRR under different conditions. When determining which method should be used, it is important to note that (a) the MLE method worked better for a test with low-quality items under the rRUM; (b) the MMLE method performed better for a test with high-quality items under the rRUM; (c) the ς^2 and ID methods were better for correcting misspecified Q-entries, whereas the MLE method was better for retaining the correct Q-entries; (d) the Q-matrix was more precisely estimated for all methods under discrete uniform distribution than under multivariate normal threshold model; and (d) the ID, MMLE, and nonparametric methods performed well in different conditions under the DINA model.

The contributions of this study are that (a) the proposed validation methods do not need to set cutoff values and specify uncertain entries. Instead, it utilizes the expert's judgment from the provisional Q-matrix; (b) the proposed validation methods can be easily implemented in the EM algorithm in the DINA model and the rRUM; and (c) the MLE method is an efficient approach for Q-matrix validation in both simulation and real data analyses, and its computation time is comparatively short. On a laptop computer with two 2.1-GHz processors and 2 GB of memory in the MATLAB 2008 software environment, the MLE, ID, and MMLE via the EM algorithm took an average of less than 20 s, 20 s, and 1 min, respectively, to run each data set with the sample size of 1,000. The MLE and MMLE methods should search through all possible q-vectors. When the number of attributes is high, an parallel implementation of the EM algorithm may be considered. An improved parallel EM algorithm has been proposed by von Davier (2017) for estimating generalized latent variable models. Moreover, MapReduce in cloud computing can increase the efficiency of all methods. For example, distributed computing can be used to estimate posterior distributions of attribute patterns separately in subsamples, and then estimate q-vectors of exclusive item sets separately by using a large number of computers (nodes).

The idea of the MLE and MMLE in this study is related to the previous studies in CD-CAT (Y. Chen et al., 2015; P. Chen & Xin, 2011; P. Chen et al., 2012). The MLE and MMLE are quite similar to three online calibration methods, namely, Cognitive Diagnostic-Method A (CD-Method A), Cognitive Diagnostic–Multiple EM Cycles (CD-MEM) proposed by P. Chen and Xin (2011), and the joint estimation algorithm (JEA) proposed by Y. Chen et al. (2015). The CD-Method A and CD-MEM firstly were used to estimate the q-vectors of new items. The CD-Method A and CD-MEM were then used to estimate the item parameters of new items by P. Chen et al. (2012). Based on these two studies, the JEA was developed to estimate both the q-vectors and item parameters of new items. This study further extends these methods to Q-matrix validation.

Some future research directions are also pointed out. First, it is necessary to consider how to determine the number of attributes for either Q-matrix validation method or the data-driven approach (Liu et al., 2012, 2013). It is important to recognize that the number of attributes was fixed to five in this study. Future research might investigate how to eliminate or add attributes by considering not only some fit statistics (J. Chen, de la Torre, & Zhang, 2013) but also the validity of classification results (Cui et al., 2012; W. Y. Wang, Song, Chen, Meng, & Ding, 2015). As the authors mentioned in the simulation study, they only used the cutoff value from the previous study by Huo and de la Torre (2013) for the ς^2 method. It should be noted that de la Torre and Chiu (2016) proposed a cutoff value based on the proportion of variance accounted for (PVAF) by a particular q-vector relative to the maximum ς^2. From the results of these two papers, the cutoff values specified in both Huo and de la Torre (2013) and de la Torre and Chiu (2016) performed very well. Thus, it would be interesting to compare the performance of the two cutoff values under the ς^2 method.

Second, it is worthwhile to explore the impact of attribute hierarchy on Q-matrix specification, because only the independent structure was considered in the simulation study. If an attribute hierarchy is well-defined, on one hand, the reduced Q-matrix (Leighton, Gierl, & Hunka, 2004; K. K. Tatsuoka, 1995) could be taken as a parameter space, that is, all possible q-vectors for an item in the estimation are restricted by the rows of $\boldsymbol{Q}_r$. The $\boldsymbol{Q}_r$ matrix can be obtained by imposing the constraints of the attribute hierarchy (Leighton et al., 2004; K. K. Tatsuoka, 1995) or determined by using the augment algorithm (Ding, Luo, Cai, Lin, & Wang, 2008). On the other hand, the difference in performance between the proposed and the existing methods is advised for investigation in a further study under different designs of the Q-matrix.

Finally, one limitation of this study is that the upper bound of r_{jk}^* is set at 0.6 for solving the identifiability issue. Another limitation is that the true CDMs were assumed to be given in the simulation study, while this is never the case in real applications (Lei & Li, 2016). However, the performance of the proposed methods relies on CDM and its classification accuracy of attribute patterns. It is well known that when the Q-matrix is correctly specified, misspecification of CDM would affect classification accuracy (Lei & Li, 2016; Ma et al., 2016). For example, the classification accuracy of attribute patterns will decrease when a general CDM is used, which is different from the true model, particularly when the sample size is small and items are of low quality (Ma et al., 2016). With CDMs uncertain, the saturated model plays an important role in detecting Q-matrix misspecifications (J. Chen et al., 2013). In the future study, investigation of these methods in some general CDMs including the generalized DINA model (de la Torre, 2011), log-linear CDM (Henson, Templin, & Willse, 2009), or the general diagnostic model (von Davier, 2005) could be more useful because in that case, identifying the model is not necessary.

Acknowledgment

The authors thank Dr. Hua-Hua Chang and two anonymous reviewers for their valuable comments on earlier versions of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was partially supported by the National Natural Science Foundation of China (Grants 31500909, 31360237, and 31160203), the Key Project of National Education Science "Twelfth Five Year Plan" of Ministry of Education of China (Grant DHA150285), the National Social

Science Foundation of China (Grant 16BYY096), the Humanities and Social Sciences Research Foundation of Ministry of Education of China (Grant 12YJA740057), the National Natural Science Foundation of Jiangxi (Grant 20161BAB212044), the Social Science Foundation of Jiangxi (Grant 17JY10), and the Education Science Foundation of Jiangxi (Grant 13YB032).

Supplemental Material

Supplementary material is available for this article online.

References

Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement*, *17*, 201-210.

Buck, G., VanEssen, T., Tatsuoka, K. K., Kostin, I., Lutz, D., & Phelps, M. (1998). *Development, selection and validation of a set of cognitive and linguistic attributes for the SAT I verbal sentence completion section (RR-98-23)*. Princeton, NJ: Educational Testing Services.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistical Association*, *39*, 83-87.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*, 1-20.

Chang, H.-H., & Wang, W. Y. (2016). ‘‘Internet plus’’ measurement and evaluation: A new way for adaptive learning. *Journal of Jiangxi Normal University (Natural Science)*, *40*, 441-455.

Chen, J., de la, Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*, 123-140.

Chen, P., & Wang, C. (2016). A new online calibration method for multidimensional computerized adaptive testing. *Psychometrika*, *81*, 674-701.

Chen, P., & Xin, T. (2011, April). *Item replenishing in cognitive diagnostic computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2010). *A comparative study on on-line calibration in cognitive diagnostic computerized adaptive testing*. Paper presented at the 75th Meeting of the Psychometric Society, Athens, GA.

Chen, P., Xin, T., Wang, C., & Chang, H.-H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, *77*, 201-222.

Chen, Y., Liu, J., & Ying, Z. (2015). Online item calibration for Q-matrix in CD-CAT. *Applied Psychological Measurement*, *39*, 5-15.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*, 598-618.

Chiu, C.-Y., & Douglas, J. A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, *30*, 225-250.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633-665.

Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*, 19-38.

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

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, *36*, 447-468.

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.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115-130.

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

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253-273.

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

Ding, S.-L., Luo, F., Cai, Y., Lin, H.-J., & Wang, X.-B. (2008). Complement to Tatsuoka's Q-matrix theory. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 417-424). Tokyo, Japan: Universal Academy Press.

Ding, S.-L., Yang, S. Q., & Wang, W. Y. (2010). The importance of reachability matrix in constructing cognitively diagnostic testing. *Journal of Jiangxi Normal University (Natural Science)*, *34*, 490-495.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika*, *49*, 175-186.

Feng, Y., Habing, B. T., & Huebner, A. (2014). Parameter estimation of the reduced RUM using the EM algorithm. *Applied Psychological Measurement*, *38*, 137-150.

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

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

Huo, Y., & de la Torre, J. (2013, April). *Data-driven Q-matrix specification for subsequent test forms*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Im, S., & Corter, J. E. (2011). Statistical consequences of attribute misspecification in the rule space method. *Educational and Psychological Measurement*, *71*, 712-731.

Ivezic, Z., Connolly, A. J., VanderPlas, J. T., & Gray, A. (2014). *Statistics, data mining, and machine learning in astronomy*. Princeton, NJ: Princeton University Press.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing*, *26*, 31-73.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258-272.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*, 331-358.

Lei, P.-W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Applied Psychological Measurement*, *40*, 405-417.

Leighton, J. P., & 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., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205-237.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, *36*, 548-564.

Liu, J., Xu, G., & Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli*, *19*, 1790-1817.

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*, 200-217.

McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*, 808-821.

Rosen, K. H. (2011). *Discrete mathematics and its applications* (7th ed.). New York, NY: McGraw-Hill.

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.

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

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C*, *51*, 337-350.

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

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessments* (pp. 327-359). Hillsdale, NJ: Lawrence Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York, NY: Taylor & Francis.

Tu, D.-B., Cai, Y., & Dai, H.-Q. (2012). A new method of Q-matrix validation based on DINA model. *Acta Psychologica Sinica*, *44*, 558-568.

von Davier, M. (2005). *A general diagnostic model applied to language testing data (RR-05-16)*. Princeton, NJ: Educational Testing Service.

von Davier, M. (2017). New results on an improved parallel EM algorithm for estimating generalized latent variable models. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative Psychology. IMPS 2016. Springer Proceedings in Mathematics & Statistics* (*Vol. 196*, pp. 1-8). Cham, Switzerland: Springer.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 65-102). Hillsdale, NJ: Lawrence Erlbaum.

Wang, S., & Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika*, *80*, 85-100.

Wang, W. Y., Ding, S. L., & Song, L. H. (2013, April). *New Q-matrix validation methods and their sensitivity under the DINA model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. San Francisco, CA.

Wang, W. Y., Ding, S. L., & You, X. F. (2011). On-line item attribute identification in cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica*, *43*, 964-976.

Wang, W. Y., Song, L. H., Chen, P., Meng, Y. R., & Ding, S. L. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, *52*, 457-476.

Xing, D., & Hambleton, R. K. (2004). Test design, item quality, and item bank size on the psychometric properties of computer-based credentialing examinations. *Educational and Psychological Measurement*, *64*, 5-21.

Zheng, Y., Chiu, C. -Y., & Douglas, J. A. (2015). *NPCD: Nonparametric methods for cognitive diagnosis* (R Package Version 1.0-9). Retrieved from https://cran.r-project.org/web/packages/NPCD/NPCD.pdf