Improving Robustness in Q-Matrix Validation Using an Iterative and Dynamic Procedure

Article *in* Applied Psychological Measurement · March 2020

DOI: 10.1177/0146621620909904

4 authors:

Pablo Nájera
Universidad Pontificia Comillas

Miguel A. Sorrel
Universidad Autónoma de Madrid

Jimmy de la Torre
The University of Hong Kong

Francisco J Abad
Universidad Autónoma de Madrid

This article should be cited as:

Nájera, P., Sorrel, M.A., de la Torre, J., & Abad, F. J. (in press). Improving robustness in Q-matrix validation using an iterative and dynamic procedure. *Applied Psychological Measurement*. doi:10.1177/0146621620909904

Improving Robustness in Q-Matrix Validation using an Iterative and Dynamic Procedure

Pablo Nájera[a], Miguel A. Sorrel[a], Jimmy de la Torre[b], & Francisco José Abad[a]

[a]: Autonomous University of Madrid; [b]: The University of Hong Kong

Author Note

Pablo Nájera, Miguel A. Sorrel, and Francisco José Abad, Department of Social Psychology and Methodology, Autonomous University of Madrid, Spain. Jimmy de la Torre, Faculty of Education, The University of Hong Kong, Hong Kong.

This research was partially supported by Grant PSI2017-85022-P (Ministerio de Ciencia, Innovación y Universidades, Spain).

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

Correspondence concerning this article should be addressed to Miguel A. Sorrel, Department of Social Psychology and Methodology, Autonomous University of Madrid, Ciudad Universitaria de Cantoblanco, Madrid 28049, Spain, e-mail: miguel.sorrel@uam.es.

Improving robustness in Q-matrix validation using an iterative and dynamic procedure

Abstract

In the context of cognitive diagnosis models, a Q-matrix reflects the correspondence between attributes and items. The Q-matrix construction process is typically subjective in nature, which may lead to misspecifications. All this can negatively affect the attribute classification accuracy. In response, several methods of empirical Q-matrix validation have been developed. The general discrimination index (GDI) method has some relevant advantages, such as the possibility of being applied to several CDMs. However, the estimation of the GDI relies on the estimation of the latent groups sizes and success probabilities, which is made with the original (possibly misspecified) Q-matrix. This can be a problem, especially in those situations in which there is a great uncertainty about the Q-matrix specification. To address this, the present study investigates the iterative application of the GDI method where only one item is modified at each step of the iterative procedure, and the required cutoff is updated considering the new parameter estimates. A simulation study was conducted to test the performance of the new procedure. Results showed that the performance of the GDI method improved when the application was iterative at the item level and an appropriate cutoff point was used. This was most noticeable when the original Q-matrix misspecification rate was high, where the proposed procedure performed better 96.5% of the times. The results are illustrated using Tatsuoka's fraction-subtraction dataset.

Key words: CDM, G-DINA, Q-matrix, validation, GDI.

Improving robustness in Q-matrix validation using an iterative and dynamic procedure

In the context of cognitive diagnosis assessment, cognitive diagnosis models (CDMs) are latent class multidimensional statistical models that classify examinees as masters or non-masters of different skills. Those skills are often referred to as *attributes*. Several CDMs have been developed in the last years, which can be categorized as either reduced or general models. The reduced models are the most specific ones; they provide low generalization but high parsimony. The *deterministic input noise* and *gate* (DINA; Haertel, 1984; Junker & Sijtsima, 2001), the *deterministic input noise* or *gate* (DINO; Templin & Henson, 2006), and the *noisy input, deterministic output* and *gate* (NIDA; Maris, 1999; Junker & Sijtsima, 2001) are some of the most widely known reduced models. Reduced models are usually preferred because of the less number of parameter estimates and ease of interpretation. However, they make strong assumptions about the data and model fit is therefore compromised. Reduced models are nested in the general models, which allow for greater flexibility, but with more demanding requirements (e.g., larger sample sizes). The *general diagnosis model* (GDM; von Davier, 2005) and the *generalized DINA model* (G-DINA; de la Torre, 2011) are two examples of general models. These models are preferred when there is not enough evidence to assume a specific response process underlying the item responses.

The estimation of a CDM typically requires two inputs: the item responses of the examinees and a Q-matrix (Tatsuoka, 1983). The Q-matrix is a J (number of items) × K (number of attributes) matrix that reflects which attributes are measured by each item. Thus, each item will have a q-vector ($\mathbf{q}_j$), in which each q-entry (q_{jk}) will adopt a value of 1 or 0 denoting if attribute k is relevant for correctly answering item j or not, respectively.

The original Q-matrix construction process should have a theoretical foundation, and thus it is usually performed after a literature review, by analyzing examinees' reports, or by domain experts. These processes are subjective in nature and can lead to some misspecifications in the Q-matrix. These Q-matrix misspecifications negatively affect the estimation of the model parameters and the accuracy of the attribute profile classification (Gao, Miller & Liu, 2017; Rupp & Templin, 2008). For this reason, in the last years, several empirically-based methods of Q-matrix validation have been developed with the aim of detecting and correcting misspecified entries in a Q-matrix.

The present paper will focus on the *general discrimination index* (GDI) method, also known as the general method of Q-matrix validation, developed for the G-DINA framework by de la Torre and Chiu (2016). The structure of the paper will be the following. First, the G-DINA model will be briefly introduced, followed by a description of the GDI method and its advantages and limitations. Second, an item-level iterative procedure for the GDI method is proposed and described. Third, the performance of the iterative procedure is compared to that of the GDI method by means of Monte Carlo simulation. Fourth, a real data illustration is conducted. Finally, a discussion of the results is provided, as well as future research insights and comments on the advantages and limitations of the proposed procedure.

Review of the G-DINA model

The G-DINA model (de la Torre, 2011) is a general, saturated CDM that subsumes most of the reduced models (e.g., DINA, DINO, *A*-CDM). In its original formulation, the probability of success can be decomposed into the sum of the effects due to the presence of specific attributes and their interactions:

$$P(\boldsymbol{\alpha}_{lj}^{*}) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \; \dots + \delta_{12\dots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}\,, \quad (1)$$

where $\boldsymbol{\alpha}_{lj}^{*}$ is the reduced attribute vector whose elements are relevant for solving the item j; δ_{j0} is the intercept of item j; δ_{jk} is the main effect due to α_k; $\delta_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$; and $\delta_{12\dots K_j^*}$ is the interaction effect due to $\alpha_1, \dots, \alpha_{K_j^*}$, where K_j^* is the number of attributes specified for item j.

The GDI method of empirical Q-matrix validation

The GDI method of empirical Q-matrix validation (de la Torre & Chiu, 2016) is a generalization of the δ-method (de la Torre, 2008) that was developed for the DINA model. The GDI method has been shown to perform well under both reduced and general CDMs at detecting and modifying misspecifications in the Q-matrix. Apart from its great flexibility and generalization, this method is included in the `GDINA` package (Ma & de la Torre, 2018) of the R software (R Core Team, 2018) with a low computational cost. This makes it one of the most accessible and easily applicable methods.

This validation method relies on the general discrimination index (GDI; usually represented as ς_j^2), which is the variance of the probabilities of success of the different latent groups that are possible for an item weighted by the posterior distribution of those groups:

$$\varsigma_j^2 = \sum_{l=1}^{2^{K_j^*}} \omega(\boldsymbol{\alpha}_{lj}^*)\left[P(\boldsymbol{\alpha}_{lj}^*) - \bar{P}(\boldsymbol{\alpha}_{lj}^*)\right]^2 \qquad (2)$$

where $2^{K_j^*}$ is the number of possible latent groups for item *j*, $\omega(\boldsymbol{\alpha}_{lj}^*)$ is the posterior probability of examinees in group $\boldsymbol{\alpha}_{lj}^*$, $P(\boldsymbol{\alpha}_{lj}^*)$ is the probability of success for examinees in this group, and $\bar{P}(\boldsymbol{\alpha}_{lj}^*)$ is the weighted mean probability of success across all the $2^{K_j^*}$ possible latent groups for item *j*.

The method is based on the rationale that the correctly specified q-vector will lead to the highest possible item discrimination value; that is, the correct q-vector for an item will be the one that maximizes ς_j^2. When comparing nested q-vectors, the specification of more attributes in the q-vector will lead to a higher ς_j^2, and thus a criterion needs to be included so that the suggested q-vector for all items is not the one containing all the attributes ($\varsigma_{\mathbf{q}_j^{1:K}}^2$). De la Torre and Chiu (2016) defined the *proportion of variance accounted for* (PVAF), which is computed as $\text{PVAF}_{jc} = \varsigma_{\boldsymbol{q}_j^c}^2 / \varsigma_{\mathbf{q}_j^{1:K}}^2$, where *c* reflects each of the $2^{K^*} - 1$ possible q-vectors (note that the zero q-vector, with no attributes specified, is not plausible). The inclusion of spurious attributes is prevented by determining a cutoff point (ϵ, also referred as *EPS* for *epsilon*), so the *suggested* q-vector would be the simplest one (i.e., the one with less attributes specified) among those that fulfill PVAF > *EPS*.

Despite the good performance of the validation method, the original study was not without limitations, as de la Torre and Chiu (2016) noted. For instance, the authors did not justify the election criterion for the value of the *EPS*, which was set to 0.95. This aspect of the method was examined by Nájera, Sorrel, and Abad (2019), who found that the GDI method showed a good performance under a wide set of conditions, given that an optimal *EPS* for each specific condition was used. Specifically, they provided a predictive formula for the optimal *EPS* as a function of the average item quality (*IQ*), the sample size (*N*), and the number of items (*J*):

$$EPS = \text{inv.logit}(-0.405 + 2.867 \cdot IQ + 4.840 \cdot 10^{-4} \cdot N - 3.316 \cdot 10^{-3} \cdot J)\,, \quad (3)$$

where *inv.logit*(·) represents the inverse function of the logit function, computed as $\exp(x)/(1+\exp(x))$. *IQ* is computed as the average item quality ($IQ = \frac{1}{J}\sum_{j=1}^{J} IQ_j$), where IQ_j is the difference in the probability of success between the latent group that possesses all the relevant attributes specified in item *j*, $P_j(\mathbf{1})$, and the one with none of them, $P_j(\mathbf{0})$.

There is another aspect of the GDI method that deserves specific attention. When computing ς_j^2, the method assumes that the Q-matrix is correctly specified: ς_j^2 relies on the estimation of the latent group sizes and their success probabilities, which are estimated using the provisional (misspecified) Q-matrix. As the authors point out, “it would be difficult, if not impossible, for the same experts to correctly specify all the entries of the Q-matrix, particularly when the test is long. Consequently, (b) [this assumption] is expected to always be violated” (de la Torre & Chiu, 2016, p. 258). The authors state that the violation of the assumption "does not automatically invalidate the viability of the proposed method. […] the proposed method appears to be robust when the misspecifications in the Q-matrix is controlled at a reasonable rate, which justifies the usefulness of the method in practice" (de la Torre & Chiu, 2016, p. 258). According to the favorable results found by them with 5% of misspecifications, and with 10% of misspecifications by Nájera et al. (2019), the method seems indeed to be robust when the misspecification rate is low.

However, relying on the experts to make few mistakes while specifying the Q-matrix is another assumption that may not always be realistic or, at least, will remain uncertain. It is reasonable to think that different knowledge domains may vary in terms of Q-matrix specification difficulty. For instance, the Q-matrix of a scholastic exam of mathematical operations seems easier to specify (e.g., "*8 + 3 × 2*", would be easily detected as measuring, for example, "sum" and "multiplication", but not "subtraction" or "division") than the Q-matrix of a reading comprehension test, a clinical diagnostic test, or a test assessing students' competencies (e.g., Sorrel et al. [2016] reported lower inter-rater reliability for more abstract attributes like *"Study attitudes*" compared to attributes easier to objectivize like "*Helping others*"). In fact, the Q-matrix of the popular fraction subtraction data set (Tatsuoka, 1990), which does not belong to a particularly ambiguous knowledge domain, is still controversial (Kang, Yang, & Zeng, 2019). Thus, the degree of uncertainty involved in the process could reasonably be higher than what has been assumed, especially when the response processes of the knowledge domain are somehow subjectively defined. Some authors have taken this point under consideration, and have used in their simulation studies misspecification rates up to 40% (e.g., Wang et al., 2018). In light of the above, it is expected that the GDI method performance will be compromised if the misspecification rate is reasonably high, since the noise entered by the large number of misspecified q-entries can disrupt the calculation of ς_j^2.

Iterative Q-matrix validation methods

One way of mitigating the pernicious effects that the violation of the true Q-matrix assumption may provoke is to apply the validation method with an iterative procedure. Some validation methods follow this rationale. The *iterative modified sequential search algorithm* (IMSSA; Terzi & de la Torre, 2018a) and the *iterative general discrimination index* method (iGDI; Terzi, 2017; Terzi & de la Torre, 2018b) are two validation methods in which all proposed q-vector modifications are introduced in the Q-matrix in each iteration. In this sense, they can be referred to as *test-level* iterative methods. On the other hand, the *Q-matrix refinement method* (QRM; Chiu, 2013) and the data-driven approach proposed by Liu, Xu, and Ying (2012) update the Q-matrix after each q-vector modification; that is, they modify only one item in each iteration. Thus, they can be referred to as an *item-level* iterative method. Even though *test-level* iterative methods can improve the performance of non-iterative methods, it may be more precise to apply the iterative procedure at the item level. At the test-level iteration, the first step will introduce several modifications based on the original and presumably misspecified Q-matrix, and thus the probability of introducing wrong modifications will be high. At the item-level, only the first item will be modified based on the information of the original Q-matrix, while the rest of the items will be modified based on progressively better specified Q-matrices. In the context of the GDI method, this will result in a better recovery of ς_j^2 and a more precisely predicted *EPS* as the iterations take place.

In light of the above, an optimal method should take into consideration the following desired characteristics: first, it should be conducted iteratively; second, the iterations should be applied at the item level; third, if a cutoff point is required, it should be selected by empirical means and updated within each iteration; fourth, it should be applicable to both reduced and general models. Based on this, it is expected that an item-level iterative procedure based on the GDI method, applied with an optimal *EPS* that gets updated after each iteration, will lead to promising results. The steps of the iterative procedure algorithm evaluated in this paper are the following:

Step 1: Estimate the CDM according to the item responses and the provisional Q-matrix (**Q**).

Step 2: Select the *EPS* value.

Step 3: Compute all items' ς_j^2 (and PVAF) for each possible q-vector specification and define, for each item, the set of *appropriate q-vector(s)*, which fulfill(s) PVAF > *EPS*.

Step 4: Select, for each item, the simplest element(s) among all the *appropriate q-vectors*.

4.1: If there is only one element, then it is defined as the *suggested q-vector*.

4.2: If there are more than one element, the one with the highest PVAF is defined as the *suggested q-vector*.

Step 5: Define, for each item, PVAF_j^0 as the PVAF of the *provisional q-vector* specified in **Q**, and PVAF_j^* as the PVAF of the *suggested q-vector*.

Step 6: Calculate all items' $\Delta\mathrm{PVAF}_j$, defined as $\Delta\mathrm{PVAF}_j = \left|\mathrm{PVAF}_j^* - \mathrm{PVAF}_j^0\right|$.

Step 7: Define the *hit item* as the item with the highest $\Delta\mathrm{PVAF}_j$.

Step 8: Update **Q** by changing the *provisional q-vector* by the *suggested q-vector* of the *hit item*.

Step 9: Iterate over Steps 1 to 8 until $\sum_{j=1}^{J} \Delta\mathrm{PVAF}_j = 0$.

Step 2 and Steps 6 and 7 are of special relevance for the iterative procedure. Step 2 dictates which q-vectors are going to become *appropriate q-vectors* in Step 3 and, consequently, which q-vector is going to become the *suggested q-vector* in Step 4. If the *EPS* value is improperly chosen, the *suggested q-vectors* will be more likely to be incorrect. Thus, each iteration will probably increase the distance between the provisional Q-matrix and the true Q-matrix in a sort of "snowball" effect (i.e., errors will lead to more errors), and the ς_j^2 will be worse specified. Hence, it is very important that the *EPS* election criterion is not arbitrary. The predictive formula provided by Nájera et al. (2019; see Equation 3) showed a good performance under a wide range of conditions. Furthermore, it can be easily implemented in the iterative procedure and entails an additional benefit: as the prediction formula considers the average item quality (IQ), which is computed after the model is estimated, the *EPS* in Step 2 can be updated after each iteration. Step 7 is also very important, because the election of the *hit item* can be neither be at random. Especially in the first iterations, in which the Q-matrix will presumably still have several misspecifications, the ς_j^2 is going to be calculated with some error. Steps 6 and 7 are used to select, for each iteration, the q-vector that is more likely to be misspecified. These steps should optimize the performance of the iterative procedure by increasing the probabilities of properly modifying a q-vector in each iteration. The iterations would stop when all the *provisional q-vectors* and *suggested q-vectors* are equal.

Simulation study

A simulation study was conducted to test if the proposed iterative procedure for the GDI method provides better results than the standard (non-iterative) procedure. Two hypotheses were stated: a) the iterative procedure will show a better performance than the standard procedure, especially when the misspecification rate is high, b) this will be true as long as the *EPS* value is properly chosen, based on the predictive formula. The performance of the iterative procedure based on an inappropriate *EPS* value is expected to be worse than that of the standard procedure, due to the "snowball" effect previously described.

Method

Design. The examinees' responses were simulated under the G-DINA model. The number of attributes was fixed at $K = 5$, and the underlying distribution of examinees' attribute patterns was uniform. The number of examinees was fixed at $N = 1000$, the average item quality at $IQ = 0.6$, and the number of items at $J = 30$. Those values are considered to be medium levels of each factor in applied contexts (Nájera et al., 2019). Table 1 shows the Q-matrix used to simulate the examinees' responses ($\mathbf{Q}_{\text{true}}$). The Q-matrix was used in the paper of de la Torre and Chiu (2016). It contains the same number of one-, two- and three-attribute items, and each attribute is measured by the same number of items. Its structure satisfies the required conditions to be a complete (Köhn & Chiu, 2017, 2018) and identifiable (Gu & Xu, in press a, in press b) Q-matrix. Three variables were studied: the proportion of misspecified q-entries or misspecification rate (*MR* = 0.1, 0.2, 0.3, 0.4), the application procedure for the GDI method (iterative, standard), and the *EPS* value (predicted *EPS*, 0.95). Thus, a total of 16 conditions resulted after combining the different factor levels (4 misspecification rates × 2 GDI application procedures × 2 *EPS* values).

Table 1

Q-Matrix for the Simulated Data

Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0	16	0	1	0	1	0
2	0	1	0	0	0	17	0	1	0	0	1
3	0	0	1	0	0	18	0	0	1	1	0
4	0	0	0	1	0	19	0	0	1	0	1
5	0	0	0	0	1	20	0	0	0	1	1
6	1	0	0	0	0	21	1	1	1	0	0
7	0	1	0	0	0	22	1	1	0	1	0
8	0	0	1	0	0	23	1	1	0	0	1
9	0	0	0	1	0	24	1	0	1	1	0
10	0	0	0	0	1	25	1	0	1	0	1
11	1	1	0	0	0	26	1	0	0	1	1
12	1	0	1	0	0	27	0	1	1	1	0
13	1	0	0	1	0	28	0	1	1	0	1
14	1	0	0	0	1	29	0	1	0	1	1
15	0	1	1	0	0	30	0	0	1	1	1

Data generation. The probabilities of success of the latent groups with all the relevant attributes, $P_j(\mathbf{1})$, and the probabilities of success of the latent groups with none of them, $P_j(\mathbf{0})$, were manipulated to generate the item's quality (IQ_j). Specifically, $P_j(\mathbf{1}) = U(0.7, 0.9)$ and $P_j(\mathbf{0}) = U(0.1, 0.3)$, which results in average values of $\bar{P}(\mathbf{1}) \cong 0.8$ and $\bar{P}(\mathbf{0}) \cong 0.2$, giving an average item quality of $IQ = \bar{P}(\mathbf{1}) - \bar{P}(\mathbf{0}) \cong 0.6$. For the other latent groups (those with some of the relevant attributes), the probabilities of success were simulated so that they increased as the number of mastered attributes grew (i.e., monotonicity constraint). Thus, a latent group that masters more attributes than other will always have higher probabilities of success.

Misspecifications in the Q-matrix were introduced randomly with two constraints: first, all items measured at least one attribute, and second, the first five items were not modified. This latter constraint ensured the completeness of the Q-matrix, by assuring that each attribute had, at least, one single-attribute item measuring it (Köhn & Chiu, 2017, 2018).

A total of 200 data sets were generated for each of the conditions. For each data set, the IQ_j were generated according to the aforementioned uniform distribution, and a different misspecified Q-matrix ($\mathbf{Q}_{\text{miss}}$) was produced. All simulations and CDM analyses were performed in R software, using the `GDINA` package.

Dependent variables. Two different types of dependent variables were used to assess the performance of the validation method. First, the Q-matrix recovery rate (QRR) was used to measure the quality of the Q-matrix specification recovery. It reflects the number of q-entries that the method correctly specifies divided by the total number of q-entries ($J \times K$). Second, the proportion of correctly classified attributes (PCA) and the proportion of correctly classified vectors (PCV) were used to reflect the accuracy of attribute profile classification (Ma & de la Torre, 2018). The PCA measures the proportion of entries (i.e., attributes) correctly classified in the $N \times K$ matrix of attribute profile classification, while the PCV reflects the proportion of examinees' attribute profiles that are completely correctly classified (i.e., correctly classified rows in the $N \times K$ matrix of attribute classifications). Please note that the PCV is a stricter measure than the PCA, and will usually obtain lower values. These accuracy measures are of high relevance, since they provide information about the impact of the Q-matrix specification quality in the final output of a CDM.

When applying a Q-matrix validation method, the suggested Q-matrix might show some attributes positions (i.e., columns) interchanged. The possibility of having interchanged attributes increases as the misspecification rate is higher. Thus, for each replica, the suggested Q-matrix was compared with $\mathbf{Q}_{\text{true}}$ by checking the similarity between both matrices' columns. Specifically, the mean absolute difference between the columns was conducted, and the suggested Q-matrix's attribute columns were presented in the order that minimized the difference with the corresponding $\mathbf{Q}_{\text{true}}$ attribute columns. This process is akin to a domain expert labelling the factors when interpreting a factor analysis, where the order of the factors is arbitrary. In the present case, the domain expert will evaluate whether the attributes are correctly labelled.

Results

Before describing the main results, a brief comment about the iterative process (when using the predicted *EPS*) is provided. No convergence problems were registered during the simulation study. Table 2 shows the average number of iterations and number of items modified (with one or more modifications in their q-vector) for each misspecification rate condition. As expected, both measures increased as the misspecification rate did. It is important to note that the number of iterations is usually higher than the number of items modified, given that one item can be modified several times during the iteration procedure. One item can be more properly modified at a later moment of the procedure, when the rest of the Q-matrix is better specified. On the other hand, information about the average *IQ* and *EPS* is given in Table 3. As expected, the initial *IQ* (i.e., the one estimated with the misspecified Q-matrix) rapidly decreased as the misspecification rate increased. However, after the iterative procedure was completed, the final *IQ* was adequately recovered, even for the most unfavorable condition (i.e., $MR = 0.4$). This had an impact on the predicted *EPS*, which also showed an increase from the original misspecified Q-matrix to the final validated Q-matrix.

In the following results, the performance of the standard and iterative procedures, as well as their interaction with the predicted *EPS* and the *EPS* of 0.95, will be described. Tables

4, 5, and 6 show the results for the different dependent variables and conditions of the simulation study in conjunction with the results obtained with the true Q-matrix and the misspecified Q-matrices, which serve as upper and lower baselines, respectively. The type of misspecification error (under- or over-specification) is disaggregated in Table 4. Plots for the distribution of the dependent variables across the 200 replicates per misspecification rate condition are provided in the Online Appendix. The different tables presented here include the median of the 200 replicates due to the existence of asymmetry in the results distributions. Results regarding the QRR, the PCA, and the PCV were consistent and showed similar patterns. Thus, unless otherwise indicated, results for the three measures are described together.

Table 2
Average Number of Iterations and of Modified Items

	Number of iterations				Number of items modified*			
MR	Mean	SD	Min	Max	Mean	SD	Min	Max
0.1	16.9	2.4	10	24	14.6	2.1	9	20
0.2	23.2	2.8	17	31	19.4	1.9	14	23
0.3	29.4	5.0	20	53	22.4	1.8	18	27
0.4	35.3	5.3	26	62	24.2	1.6	19	28

Note. * = with one or more modifications in their q-vector. MR = misspecification rate. This information refers to the iterative procedure in conjunction with the predicted *EPS*.

Table 3
*Average Item Quality (*IQ*) and Used* EPS

	IQ		*EPS*	
MR	Initial	Final	Initial	Final
0.1	0.545	0.574	0.824	0.836
0.2	0.481	0.567	0.795	0.833
0.3	0.421	0.549	0.765	0.825
0.4	0.369	0.531	0.738	0.817

Note. MR = misspecification rate. Initial *IQ* and *EPS* values are obtained with the original misspecified Q-matrix. Final *IQ* and *EPS* values are obtained with the validated Q-matrix after the iterative procedure (using the predicted *EPS*) is completed. Items were simulated with an *IQ* of 0.60.

As can be seen from Tables 4 to 6, the iterative implementation used in conjunction with the predicted *EPS* always led to the best results. The Q-matrix recovery was very close to one when the initial misspecification rate was low (QRR = 0.940), and was still high even when the initial misspecification rate was high (QRR = 0.893). This procedure achieved the highest QRR among the four presented procedures in most of the replicates, especially as the misspecification rate increased. Thus, the iterative-predicted *EPS* implementation obtained the highest QRR 62% of the times (*MR* = 0.1), 85.5% (*MR* = 0.2), 93.5% (*MR* = 0.3), and 96.5% (*MR* = 0.4). It is important to note that, in those replicas in which it did not obtained the highest QRR, it still obtained a QRR close to the highest, with a maximum loss of 0.07 through all misspecification rates. On the other hand, it obtained a QRR up to 0.32 higher than the next best procedure, which reflects the better overall Q-matrix recovery shown in Table 4. According to the GDI method rationale, a higher *EPS* tends to suggest more complex q-vectors (i.e., with more attributes specified), and vice versa; thus, in Table 4 it can be seen that the *EPS* of 0.95 produced more over-specification errors, while the predicted *EPS* produced more under-specifications. The accuracy measures obtained with the iterative-predicted *EPS* procedure were generally close to the upper limit regardless of the misspecification rate. This was especially true for PCA. The misspecification rate affected more severely the rest of the procedures. For example, the range of the median PCA values reported in Table 5 for the standard and iterative implementations used in conjunction with the predicted *EPS* were 0.085 and 0.012, respectively.

Table 4

Medians for the Q-Matrix Recovery Rate (QRR) Results

			Predicted *EPS*		*EPS* = 0.95	
MR	$\mathbf{Q}_{true}$	$\mathbf{Q}_{miss}$	std	ite	std	ite
0.1	1	0.900 (6, 9)	**0.940** (8, 1)	**0.940** (8, 0)	0.887 (1, 16)	0.833 (1, 24)
0.2	1	0.800 (13, 17)	0.907 (11, 3)	**0.933** (9, 1)	0.827 (2, 24.5)	0.780 (1, 32.5)
0.3	1	0.700 (19, 26)	0.817 (17, 11)	**0.913** (11, 2)	0.720 (6, 36)	0.687 (1, 46)
0.4	1	0.600 (26, 34)	0.740 (21, 18)	**0.893** (13, 3)	0.627 (8.5, 47)	0.610 (0.5, 58)

Note. MR = misspecification rate; $\mathbf{Q}_{true}$ = true Q-matrix; $\mathbf{Q}_{miss}$ = misspecified Q-matrix; std = standard procedure; ite = iterative procedure. A grayscale has been used for interpretation purposes. Highest QRRs among the validation methods for each MR are shown in bold.

Median values for the number of under- and over-specified q-entries, respectively, are shown in brackets. Q-matrices are formed by 150 q-entries.

Table 5

Medians for the Proportion of Correctly Classified Attributes (PCA) Results

			Predicted *EPS*		*EPS* = 0.95	
MR	$\mathbf{Q}_{true}$	$\mathbf{Q}_{miss}$	std	ite	std	ite
0.1	0.910	0.895	**0.907**	**0.907**	0.900	0.894
0.2	0.911	0.867	0.901	**0.906**	0.894	0.889
0.3	0.911	0.813	0.862	**0.903**	0.868	0.880
0.4	0.910	0.764	0.822	**0.895**	0.807	0.864

Note. MR = misspecification rate; $\mathbf{Q}_{true}$ = true Q-matrix; $\mathbf{Q}_{miss}$ = misspecified Q-matrix; std = standard procedure; ite = iterative procedure. A grayscale has been used for interpretation purposes. Highest PCAs among the validation methods for each MR are shown in bold.

Table 6

Medians for the Proportion of Correctly Classified Vectors (PCV) Results

			Predicted *EPS*		*EPS* = 0.95	
MR	$\mathbf{Q}_{true}$	$\mathbf{Q}_{miss}$	std	ite	std	ite
0.1	0.637	0.583	**0.625**	**0.625**	0.603	0.581
0.2	0.642	0.484	0.604	**0.623**	0.586	0.560
0.3	0.643	0.325	0.457	**0.613**	0.492	0.531
0.4	0.639	0.227	0.337	**0.579**	0.335	0.483

Note. MR = misspecification rate; $\mathbf{Q}_{true}$ = true Q-matrix; $\mathbf{Q}_{miss}$ = misspecified Q-matrix; std = standard procedure; ite = iterative procedure. A grayscale has been used for interpretation purposes. Highest PCVs among the validation methods for each MR are shown in bold.

The following comments can be made regarding the manipulated factors. First, as it was expected, for both application procedures (standard vs. iterative) and *EPS* values (predicted *EPS* vs. *EPS* = 0.95), results were worse as the misspecification rate increased. Second, for both the standard and iterative procedures, and in line with the conclusions of Nájera et al. (2019), the predicted *EPS* provided better results than the *EPS* of 0.95. Third, regarding the interaction between the application procedure and the *EPS* value, the iterative procedure showed a better performance than the standard procedure only when the predicted *EPS* was used. Results were very similar for both procedures when the misspecification rate was low (*MR* = 0.1), but, as the misspecification rate was higher, the differences between both procedures substantially increased favoring the iterative procedure. On the contrary, when the *EPS* of 0.95 was used, the QRR of the iterative procedure was lower for all misspecification rates. As previously stated, these results were expected, since an inappropriate *EPS* increases the probability of selecting an incorrect suggested q-vector, enlarging the distance between the provisional Q-matrix and the true Q-matrix, disrupting the calculation of ς_j^2. However, regarding the PCA and the PCV, the iterative procedure, in conjunction with the *EPS* of 0.95, showed slightly worse results when the misspecification rate was low (*MR* = 0.1 or 0.2), but outperformed the standard procedure when the misspecification rate was high (*MR* = 0.3 or 0.4). All this reflects the fact that both an iterative procedure and a dynamic optimal *EPS* value are required in order to achieve optimal results.

Real Data Example

Data and Analysis

In order to facilitate a direct comparison between the proposed procedure and the original GDI method, we used the same dataset as de la Torre and Chiu (2016). It consists of 536 examinees' responses to 11 fraction-subtraction items (Tatsuoka, 1990) measuring four attributes (see strategy *b* in Mislevy, 1996): (1) performing basic fraction-subtraction operation, (2) simplifying/reducing, (3) separating whole number from fraction, and (4) borrowing one from whole number to a fraction. Table 7 shows the initial Q-matrix for these data, which is the same as the one used by de la Torre and Chiu (2016). A higher-order G-DINA model (de la Torre & Douglas, 2004) was used to fit the data.

Results

Table 7 shows the Q-matrix suggested by the iterative procedure. Six q-entries modifications were proposed, all of them switching from 1 to 0, and all of them involving attribute 2, with the exception of attribute 1 in Item 1. These results are somewhat congruent with those found by de la Torre and Chiu (2016), who reported three modifications in attribute 2 (Items 4, 5, and 11). According to the results found in the simulation results, the iterative procedure suggested a less complex Q-matrix (i.e., less attributes specified) than the original GDI method (see Table 4).

Regarding the original Q-matrix, attribute 2 (simplifying/ reducing) seems to have theoretical relevance to solve the modified items. However, it is important to note that it shows a great collinearity with attributes 3 and 4; that is, almost every time attribute 2 is required, attributes 3 and 4 are also required. The only time that attribute 2 appears without attributes 3 or 4 is in Item 6, which is the only one that retains attribute 2 in the suggested Q-matrix. Thus, even though this attribute makes theoretical sense and seems to be correctly specified in the original Q-matrix, it cannot be properly separated from other attributes. Since it cannot provide any additional value, it becomes an irrelevant attribute and almost disappeared in the suggested Q-matrix.

Table 7

Original and suggested Q-matrices for the fraction-subtraction data

Item		Original Q-matrix				Suggested Q-matrix			
		α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4
1	$3\frac{1}{2}-2\frac{3}{2}$	1	1	1	1	0*	0*	1	1
2	$\frac{6}{7}-\frac{4}{7}$	1	0	0	0	1	0	0	0
3	$3\frac{7}{8}-2$	1	0	1	0	1	0	1	0
4	$4\frac{4}{12}-2\frac{7}{12}$	1	1	1	1	1	0*	1	1
5	$4\frac{1}{3}-2\frac{4}{3}$	1	1	1	1	1	0*	1	1
6	$\frac{11}{8}-\frac{1}{8}$	1	1	0	0	1	1	0	0
7	$3\frac{4}{5}-3\frac{2}{5}$	1	0	1	0	1	0	1	0
8	$4\frac{5}{7}-1\frac{4}{7}$	1	0	1	0	1	0	1	0
9	$7\frac{3}{5}-\frac{4}{5}$	1	0	1	1	1	0	1	1
10	$4\frac{1}{10}-2\frac{8}{10}$	1	1	1	1	1	0*	1	1
11	$4\frac{1}{3}-1\frac{5}{3}$	1	1	1	1	1	0*	1	1

Note. Q-entries modifications are highlighted with an asterisk.

Regarding Item 1, the first attribute is also removed in the suggested Q-matrix. This item can be correctly solved by following different strategies:

(a) $3\frac{1}{2}-2\frac{3}{2}=\frac{7}{2}-\frac{7}{2}=0$ (attributes 1 and 4);

(b) $3\frac{1}{2} - 2\frac{3}{2} = 2\frac{3}{2} - 2\frac{3}{2} = 0$ (attributes 1, 3, and 4).

A *mesaplot* (Ma & de la Torre, 2018), which shows the PVAF related to each possible q-vector specification, for Item 1 is presented in Figure 1. Four q-vectors (0011, 0111, 1011, 1111) clearly show a higher PVAF than the rest. Since their PVAF is higher than the *EPS* (0.903), they form the set of appropriate q-vectors. The q-vector of 0011 is chosen as the suggested q-vector because it is the simplest one. This attribute specification is related to strategy (b), although attribute 1 is missing. A possible explanation to this could be that the subtraction required in Item 1 may be a very easy operation that almost every examinee can solve, since it involves two identical elements. As a consequence, attribute 1 would no longer provide additional information. Nevertheless, these are modification suggestions, and domain experts can seek among the appropriate q-vector in order to find the most suitable specification. The last decision about the Q-matrix specification should rely on the judgment of domain experts (de la Torre & Chiu, 2016).

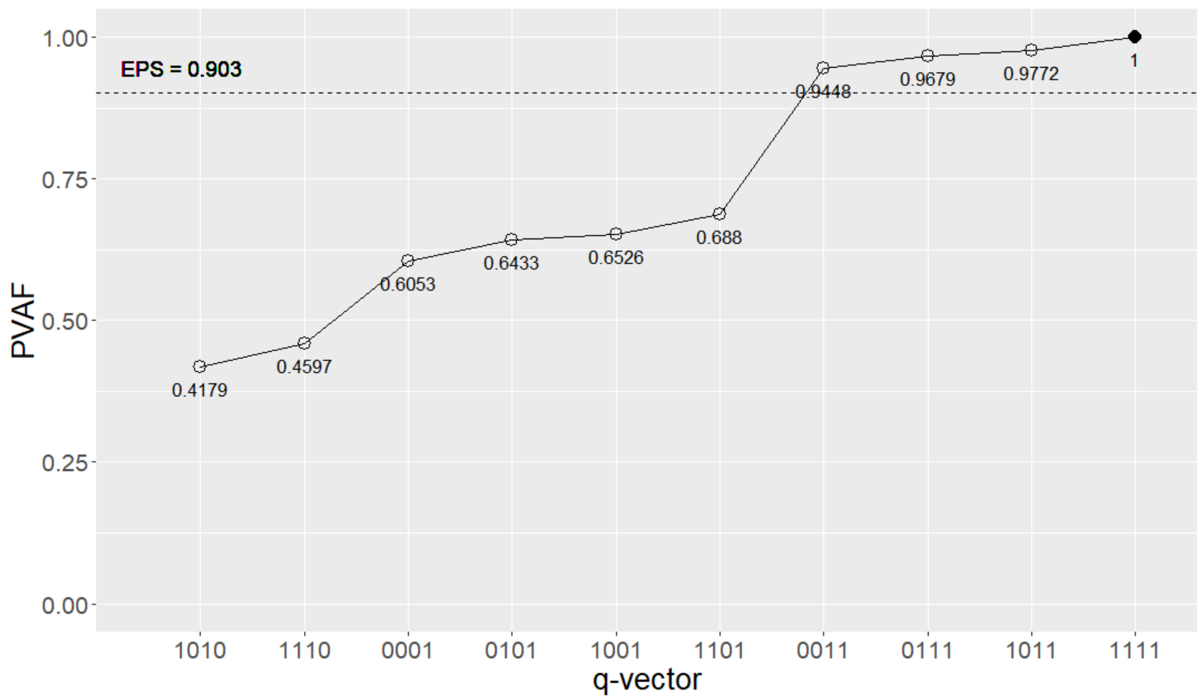

Figure 1. Mesaplot for Item 1 of Tatsuoka's fraction-subtraction dataset included in Table 7. The black dot represents the original q-vector specification (1111). The PVAF represents the ratio of the GDI associated to a q-vector to the highest possible GDI that is obtained when all the attributes are specified.

Discussion

CDMs rely on a correctly specified Q-matrix to provide an accurate classification of examinees' attribute profiles. Domain experts are expected to specify the Q-matrix along with a theoretical background, but they may commit some errors while doing so, especially when the knowledge domain is particularly complex and ambiguous (e.g., mental pathologies, reading comprehension, students' competencies). In this context, among the many Q-matrix validation methods that have been developed in the last few years, de la Torre and Chiu (2016) proposed the GDI method, which has some important advantages, such as its great flexibility to be used with several reduced or general CDMs, its good performance at modifying incorrectly specified q-vectors, and its low computational cost (Ma & de la Torre, 2018). Despite its benefits, the GDI method relies on the original Q-matrix, which may not be correctly specified in most applied contexts. Although the method seemed robust to the violation of this assumption when the Q-matrix misspecification rate was low, it is expected to show a poorer performance when validating Q-matrices with more misspecifications.

The present paper evaluated an item-level iterative with dynamic *EPS* implementation for the GDI method (this approach can be referred to as "ILD-GDI"). Considering past research (e.g., Chiu, 2013; Liu et al., 2012; Nájera et al., 2019; Terzi & de la Torre, 2018ab), we hypothesized that this implementation would lead to better results compared to the existing procedures, especially when the misspecification rate is high. A simulation study was conducted to test this hypothesis. Results showed that the new implementation did provide better results. The gain obtained increased as the misspecification rate was higher.

The iterative procedure was hypothesized to have a poorer performance than the standard procedure when used in conjunction with an inappropriate *EPS*. However, even though the iterative-0.95 *EPS* (*ite95*) obtained a lower QRR than the standard-0.95 *EPS* (*std95*), it provided better attribute profile classification results when the misspecification rate was high (*MR* = 0.3 or 0.4). A tentative explanation of this result could be related to the type of misspecification error. Some prior studies in the field (e.g., Gao, Miller, & Liu, 2017; Choi, Templin, Cohen, & Atwood as cited in Kunina-Habenicht, Rupp, & Wilhelm, 2012) have found that under-specifications have a greater impact in attribute profiles classification than over-specifications. This effect is logically expected, since removing a parameter with a substantive effect from a model might dramatically disrupt the probabilities of success of the affected item; on the other hand, a spurious parameter added to the model may obtain a marginal effect estimate, mitigating its impact (as long as the sample size is big enough to produce stable parameter estimates).

This effect can explain the aforementioned results regarding *ite95* and *std95*. Table 4 shows the information regarding the Q-matrix recovery, disaggregated by specification error type. On one hand, when *MR* = 0.1 or 0.2, *std95*'s QRR was higher than *ite95*'s. *Std95*'s PCA and PCV were also higher than *ite95*'s. However, PCA differences were not as big as QRR differences, since the higher amount of misspecifications in *ite95* were mainly over-specifications, and both procedures had a similar number of under-specifications. On the other hand, when *MR* = 0.3 or 0.4, *std95*'s QRR was still higher than *ite95*'s. However, *ite95*'s PCA and PCV were higher than *std95*'s. Here, the QRR differences between both procedures were smaller than those obtained with *MR* = 0.1 or 0.2. In addition, the higher amount of misspecifications in *ite95* were mainly over-specifications, while *std95* obtained more under-specifications. As previously stated, the latter might provoke a bigger disruption in the posterior probabilities estimates, causing a worse attribute classification.

The explanation given above is certainly conditioned by the total number of misspecifications. Under-specifications may have a bigger impact than over-specifications as long as the total number of misspecifications remains at a similar range. The validation procedure proposed in the present work (iterative in conjunction with the predicted EPS) showed a higher number of under-specifications than *std95* and *ite95*; however, it showed a much better performance in terms of Q-matrix specification recovery, which resulted in a higher classification accuracy. It is important to note that other factors may have a relevant role in modulating the relation between Q-matrix specification and attribute classification, such as the number of different q-vectors represented in the Q-matrix (Rupp & Templin, 2008) and the identifiability of the Q-matrix (Gu & Xu, in press a, in press b).

Finally, a reviewer proposed examining whether the proposed procedure performs also well when the underlying attribute's distribution is non-uniform. The performance of the procedures under a multivariate normal distribution ($\rho = 0.25$; see Xu & Shang, 2018) and a higher-order distribution ($\lambda_0 = (-1, -0.5, 0, 0.5, 1)$, $\lambda_{1k} = 1.5$; see de la Torre & Chiu, 2016) are provided in the Online Appendix. It was observed that the pattern of results was very similar to the ones obtained with the uniform distribution. Thus, the interpretation of the findings do not differ according to the underlying attribute distribution, and the proposed procedure still showed the best Q-matrix recovery and classification accuracy.

In conclusion, the ILD-GDI method proposed in this paper outperformed the original method developed by de la Torre and Chiu (2016), as well as the method with the optimized *EPS* value election (Nájera et al., 2019). The proposed procedure showed good performance at detecting and modifying the Q-matrix even with a high misspecification rate (QRR $\geq$ 0.893) and also at classifying attribute profiles (PCA $\geq$ 0.895; $\text{PCA}_{\mathbf{Q}_{\text{true}}} \approx 0.910$), being the only procedure that achieved a PCV higher than 0.5 under the worse misspecification rate scenario (PCV $\geq$ 0.579; $\text{PCV}_{\mathbf{Q}_{\text{true}}} \approx 0.640$). The iterative procedure's computation time was short. On a laptop computer with four 2.2-GHz processors and 7 GB of RAM memory, the average replica computation time under the worst condition (*MR* = 0.4) was 111 seconds.

The performance of the ILD-GDI method was also illustrated with Tatsuoka's fraction-subtraction data. De la Torre and Chiu (2016) found that the standard GDI method with an *EPS* of 0.95 proposed three modifications. These modifications were congruent with the ones suggested by the ILD-GDI method. The suggestions of the ILD-GDI should be considered rather than the GDI method's ones, since it provides a better recovery of the Q-matrix, as shown in the simulation study. However, two consideration should be noticed. First, even though Q-matrix validation methods are helpful in the search for the best possible specified Q-matrix, some misspecifications may remain after their application. Second, attribute positions in the Q-matrix are arbitrary just as factors are in a factor analysis; thus, when two attributes (i.e., Q-matrix columns) have a similar specification through the items and / or the number of misspecifications in the original Q-matrix is high, there exists the possibility that the suggested Q-matrix shows interchanged positions for these attributes with respect to the original Q-matrix. These considerations emphasize the role of domain experts in the review of the validated Q-matrix. They should reject those suggested modifications that lack a theoretical interpretation and check that the attributes maintain their original meaning. Also, if they consider that several strategies can be followed to answer the items, multiple-strategy models may be of help (e.g., de la Torre & Douglas, 2008; Ma & Guo, 2019). These considerations may provide the most useful Q-matrix specification, since a tradeoff between theoretical interpretation and data fit can be more easily achieved.

Further research is needed to extend the applicability of the ILD-GDI method. Even though the performance of the GDI method was deeply studied under a wide range of conditions by Nájera et al. (2019), the performance of the ILD-GDI method has only been tested under a limited set of conditions. Further research would help to know whether it is robust when the conditions are less favorable (e.g., small sample size, short test length, low item quality). In this sense, other factors can be added to the study design, such as the number of attributes or the underlying CDM (e.g., DINA).

Furthermore, it would be interesting to study whether the inclusion of model fit indices to the iterative procedure could improve its performance. For instance, Kang et al. (2019) used the item-level version of the RMSEA, which provided good results under the DINA model. For the general CDMs framework, the Akaike's information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarzer, 1976), which have been previously used as fit indices in CDMs (e.g., Chen, de la Torre, & Zhang, 2013), could be good candidates at selecting the *suggested q-vector*. One important drawback of this approach would be the dramatic computational cost increment, since one additional model should be estimated for each q-vector for each *hit item*. In this vein, the Wald test for model comparison has also been recently used for Q-matrix validation under the sequential G-DINA model (Ma & de la Torre, 2019).

Final remarks

The authors want to emphasize that empirical validation methods *suggest* modifications, and cannot derive a *true* Q-matrix in empirical settings. The suggested Q-matrix represents a model with empirical support. The purpose of Q-matrix validation should not be to replace experts from the Q-matrix specification process, but to "provide supplemental information for improving model-data fit, and consequently, increasing the validity of inference from cognitive diagnosis assessments" (de la Torre & Chiu, 2016, p. 268). Especially in those contexts in which there is a certain degree of uncertainty involving the Q-matrix, modification suggestions may help to understand which cognitive processes are involved in responding each item. Also, as has been shown in the real data illustration, validation methods can help detecting problems regarding the structure of the Q-matrix (e.g., attributes collinearity). Thus, we recommend applying three steps during the Q-matrix specification process. First, construct the original Q-matrix with the help of domain experts. In this step, the Delphi methodology can be of great help, facilitating the debate and subsequent agreement between the judges (see Sorrel et al., 2016). It is also useful to track the degree of uncertainty involved in each q-entry during the process. Second, apply an empirical Q-matrix validation method, in order to detect any possible misspecifications made in the first step. Third, gather again the panel of experts to debate the theoretical viability of the suggested modifications and the meaning of the attributes after the process is completed. The degree of uncertainty involving each q-entry recorded in the first step can be of help at this point; a q-entry in which all experts showed a total agreement should probably not be modified even though the validation method suggests the opposite. In conclusion, the authors are of the opinion that the theory should be the main guide in the Q-matrix specification process. Empirical validation methods' role should be to support the domain experts' judgements.

References

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automated Control*, *19,* 716–723.

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.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*(8), 598–618.

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. (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*(2), 253–273.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: an analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595–624.

Gao, M., Miller, M. D., & Liu, R. (2017). The impact of Q-matrix misspecification and model misuse on classification accuracy in the generalized DINA model. *Journal of Measurement and Evaluation in Education and Psychology, 8*(4), 391–403.

Gu, Y., & Xu, G. (in press a). Partial identifiability of restricted latent classes models. *Annals of Statistics*. Retrieved from arXiv:1803.04353.

Gu, Y., & Xu, G. (in press b). Sufficient and necessary conditions for the identifiability of the Q-matrix. *Statistica Sinica*. Retrieved from arXiv:1810.03819.

Haertel, E. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement*, *8*(3), 333–346.

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric IRT. *Applied Psychological Measurement*, *25*, 258–272.

Kang, C., Yang, Y., & Zeng, P. (2019). Q-matrix refinement based on item fit statistic RMSEA. *Applied Psychological Measurement*, *43*(7), 527–542.

Köhn, H.-F., & Chiu, C.-Y. (2017). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika*, *82*(1), 112–132.

Köhn, H.-F., & Chiu, C.-Y. (2018). How to build a complete Q-matrix for a cognitively diagnostic test. *Journal of Classification*, *35*, 273–299.

Kunina-Habenicht, O., Rupp, A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59–81.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, *36*(7), 548–564.

Ma, W., & de la Torre, J. (2018). GDINA: The generalized DINA model framework. R Package Version 2.0.8. Retrieved from https://cran.r-project.org/package=GDINA

Ma, W., & de la Torre, J. (2019). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12156

Ma, W., & Guo, W. (2019). Cognitive diagnosis models for multiple strategies. *British Journal of Mathematical and Statistical Psychology*, *72*, 370–392.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187–212.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*(4), 379–416.

Nájera, P., Sorrel, M. A., & Abad, F. J. (2019). Reconsidering cutoff points in the general method of empirical Q-matrix validation. *Educational and Psychological Measurement*, *79*(4), 727–753.

R Core Team (2018). R (Version 3.4) [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.

Rupp, 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*(1), 78–96.

Schwarzer, G. (1976). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, *19*(3), 506-532.

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

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

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305.

Terzi, R. (2017). *New Q-matrix validation procedures* (Unpublished doctoral dissertation). Rutgers, The State University of New Jersey, New Jersey, USA.

Terzi, R., & de la Torre, J. (2018a). An iterative method for empirically-based Q-matrix validation. *International Journal of Assessment Tools in Education*, *5*, 248–262.

Terzi, R., & de la Torre, J. (2018b, April). *Two general iterative Q-matrix validation procedures*. Paper presented at the meeting of the National Council of Measurement in Education, New York, NY.

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

Wang, W., Song, L., Ding, S., Meng, Y., Cao, C., & Jie, Y. (2018). An EM-based method for Q-matrix validation. *Applied Psychological Measurement*, 1–14.

Xu, G., & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, *113*(523), 1284–1295.