Article

Educational and Psychological
Measurement
1–27
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0013164418822700
journals.sagepub.com/home/epm
SAGE

Reconsidering Cutoff Points in the General Method of Empirical Q-Matrix Validation

Pablo Nájera[1], Miguel A. Sorrel[1], and Francisco José Abad[1]

Abstract

Cognitive diagnosis models (CDMs) are latent class multidimensional statistical models that help classify people accurately by using a set of discrete latent variables, commonly referred to as attributes. These models require a Q-matrix that indicates the attributes involved in each item. A potential problem is that the Q-matrix construction process, typically performed by domain experts, is subjective in nature. This might lead to the existence of Q-matrix misspecifications that can lead to inaccurate classifications. For this reason, several empirical Q-matrix validation methods have been developed in the recent years. de la Torre and Chiu proposed one of the most popular methods, based on a discrimination index. However, some questions related to the usefulness of the method with empirical data remained open due the restricted number of conditions examined, and the use of a unique cutoff point (*EPS*) regardless of the data conditions. This article includes two simulation studies to test this validation method under a wider range of conditions, with the purpose of providing it with a higher generalization, and to empirically determine the most suitable *EPS* considering the data conditions. Results show a good overall performance of the method, the relevance of the different studied factors, and that using a single indiscriminate *EPS* is not acceptable. Specific guidelines for selecting an appropriate *EPS* are provided in the discussion.

Keywords

CDM, G-DINA, Q-matrix, validation, GDI

[1] Universidad Autónoma de Madrid, Madrid, Spain

Corresponding Author:
Miguel A. Sorrel, Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, Madrid 28049, Spain.
Email: miguel.sorrel@uam.es

In the last few years, encouraged by the growing influence of cognitive psychology, cognitive diagnosis models (CDMs) have gained interest as a suitable statistical approach to make inferences about the diverse cognitive components that are involved in the item response process. Especially in the educational field, CDMs help determine the students' strengths and weaknesses more accurately, letting the educators target their teaching efforts toward the most problematic knowledge domains for their students. This perspective differs with the one that underlies the traditional evaluation methods, in which a unique score often summarizes the students' knowledge in relative wide domains, making it difficult to know exactly where they are failing. This complicates the implementation of an educational strategy focused on strengthening the specific concepts that are not understood (de la Torre & Minchen, 2014).

Even though many articles regarding CDMs are related to the educational field (e.g., Bradshaw, Izsák, Templin, & Jacobson, 2014; Choi, Lee, & Park, 2015; Ravand, 2016; von Davier, 2005), these models are general enough to be used in other areas. In this vein, in the last few years, they have been used in the study of mental pathologies (e.g., de la Torre, van der Ark, & Rossi, 2015; Jaeger, Tatsuoka, Berns, & Varadi, 2006; Templin, & Henson, 2006), or in the organizational field of staff selection (e.g., García, Olea, & de la Torre, 2014; Sorrel et al., 2016).

CDMs are conceptualized as latent trait multidimensional models, in which the latent traits, instead of being defined as continuous—as in item response theory (IRT) or confirmatory factor analysis (CFA)—are defined as categorical/discrete. These discrete latent traits are often called *attributes*. Attributes refer to the skills or cognitive processes an examinee should possess or master to resolve an item. They may be dichotomous (''mastery'' vs. ''non-mastery'' of the attribute) or polytomous (''bad performance,'' ''regular performance,'' ''good performance''). Attributes are latent variables defined with a certain degree of preciseness that belong to a wider domain of knowledge and, thus, they are interrelated to each other but separable at the same time. The main objective of CDMs is to provide detailed information about the attributes each examinee possesses or, more technically, to classify examinees into a set of latent classes specified by attribute vectors (de la Torre & Douglas, 2004). Due to the discrete nature of attributes, there will be a restricted number of latent classes.

The number of attributes is often represented as K, so examinee i will have an *attribute profile* $\boldsymbol{\alpha}_i = \{\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK}\}$, which is a binary vector where $\boldsymbol{\alpha}_{ik}$ = 1 or 0 indicates if examinee i masters or not attribute k, respectively. This is true if attributes are defined as dichotomous, which is the most common in the literature; from now on we will focus on this case (for an application with polytomous attributes, see, e.g., Chen & de la Torre, 2013). Thus, there is a total of 2^K different attribute vectors that constitute the spectrum of latent classes. Latent classes are represented as $\boldsymbol{\alpha}_l$, being $1 \leq l \leq 2^K$. CDMs are expressed by its definition of $P(X_j = 1|\boldsymbol{\alpha}_l)$, the conditional probability of succeeding item j given latent class l (Sorrel et al., 2016). Some of these models will be presented below.

Review of Different CDMs

In the last years, several CDMs with a different degree of generalization have been proposed. The most specific models are usually known as reduced, because they are nested in the more general models. One of the most known reduced models is the *deterministic input, noise* and *gate* model (DINA; Haertel, 1984; Junker & Sijtsma, 2001), which assumes a *conjunctive* process in which the examinee must possess all the attributes involved in the item to answer it correctly. The *deterministic input, noise* or *gate* model (DINO; Templin & Henson, 2006), assumes a *disjunctive* process in which it is enough to possess one single required attribute to answer the item correctly. The *noisy input, deterministic output* and *gate* model (NIDA; Maris, 1999; Junker & Sijtsma, 2001), the *compensatory and reduced reparameterized unified model* (Hartz & Roussos, 2008), the *additive* CDM (de la Torre, 2011) and the linear logistic model (de la Torre & Douglas, 2004) are also reduced models. Although these models are easy to understand, their different structures make them difficult to compare to each other. Thus, more general and flexible CDMs have been developed, such as the general diagnosis model (von Davier, 2005), the loglinear CDM (Henson, Templin, & Willse, 2009), and the generalized DINA model (G-DINA; de la Torre, 2011). These models offer a framework to relate many of the aforementioned reduced models, in addition to estimate parameters and compare the fit of different CDMs one item at a time (de la Torre, 2011; de la Torre et al., 2015; Sorrel, de la Torre, Abad, & Olea, 2017). The G-DINA model has been previously used in some Q-matrix validation method papers (e.g., de la Torre & Chiu, 2016; Gao, Miller, & Liu, 2017; Gu, Liu, Xu, & Ying, 2018; Ma, Iaconangelo, & de la Torre, 2016) and so it will be the object of study in this article.

In the original formulation of the G-DINA model, the probability of success can be decomposed into the sum of the effects due to the presence of specific attributes and their interactions (de la Torre, 2011):

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

where $\boldsymbol{\alpha}_{lj}^{*}$ is the reduced attribute vector whose elements are relevant for solving the item j; δ_{j0} is the basal probability 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 $\delta_{12\ldots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$.

The Q-Matrix

Definition. Regardless the specific CDM formulation, all models have two main inputs: the item responses and a Q-matrix (Tatsuoka, 1983), usually binary and with dimensions J (number of items) $\times K$ (number of attributes), which determines the required attributes to correctly answer each item, making explicit the intern structure of the test and the confirmatory nature of CDMs. More specifically, being $\boldsymbol{Q} = \{q_{jk}\}$, each *q-entry* ($q_{jk}$) of the matrix points out whether attribute k is relevant or not to answer item j correctly. Thus, each item will have its own *q-vector* ($\boldsymbol{q_j}$), which indicates the relevant attributes for that item. The number of attributes specified in a q-vector is represented as K_j^*, so one item can distinguish $2^{K_j^*}$ different latent groups. The main output of CDMs is a *classification of attribute profiles*, a matrix with dimensions N (number of examinees) $\times K$, where the posterior probability of each examinee's attributes mastery is established; this probability can be later dichotomized around a cutoff point (usually 0.5) to determine mastery or non-mastery. Sometimes an uncertainty region is set up around this cutoff, in which no classification is made (Sorrel et al., 2016).

The Q-matrix construction is usually the first step in a CDM application process. First, it is necessary to decide what and how many attributes to measure (see Li & Suen, 2013, for a further explanation). Then, the initial Q-matrix is specified, determining what attributes are relevant to answer each item correctly. There are various strategies for this process, but most of them rely on subjective information. Thus, it is reasonable to think that they will lead to some misspecifications in the Q-matrix (Chiu, 2013; de la Torre & Chiu, 2016; Li & Suen, 2013). These methods are confirmatory in the sense that they assume that the proposed Q-matrix is known. However, prior research suggest that specification errors may be problematic as they can affect dramatically the estimation of the model parameters and the accuracy of the attribute classification (e.g., Chiu, 2013; de la Torre, 2008; de la Torre & Chiu, 2016; Gao et al., 2017; Romero, Ordóñez, Ponsoda, & Revuelta, 2014).

Q-Matrix Validation. Although the need of making the verification process of the Q-matrix is highly recognized, at the present time there are not as many methods to detect specification errors, especially for general CDMs (de la Torre & Chiu, 2016). Some authors like Barnes (2010); Liu, Xu, and Ying (2012); and Chiu (2013, as cited in de la Torre & Chiu, 2016) have developed different validation methods that, among other limitations, can be only applied to a restricted number of reduced models. The study of Romero et al. (2014), in which a validation method for the *Least Squares Distance* model (LSDM) was developed (Dimitrov, 2007, as cited in Romero et al., 2014), shares this limitation. Furthermore, even though the method seems adequate for detecting misspecifications, it does not suggest substitutive q-vectors. Recently, Chen (2017) has developed a very complete method to detect and modify Q-matrix misspecifications based on a sequential process and several fix indexes. The method shows a good performance, although it has some limitations, like its implementation difficulty, the great influence of the q-vector's number of attributes in the power and the need of an arbitrary cutoff point in one of the sequence steps. In addition, the number of conditions examined in the article is limited, being some of them maybe too favorable (e.g., high item discrimination). Gu et al. (2018) have also developed a Q-matrix validation procedure based on the hypothesis testing procedure. This method has some remarkable advantages, as it can be applied to all CDMs and with both complete (i.e., it can differentiate all latent attribute profiles) and incomplete Q-matrices. However, it only tests the correctness of the Q-matrix specification as a whole, and it has not been implemented a procedure to detect and modify local misspecifications yet. Last, Wang et al. (2018) have recently developed three EM-based methods for validating the Q-matrix. The methods are tested under a wide range of simulation conditions as well as with real data, and they do not depend on cutoff values or uncertain q-entries. The study focuses on two reduced models and the performance under no general models is evaluated. Moreover, the computational cost of the method rapidly increases as the number of attributes gets higher.

One of the most extended Q-matrix validation method is based on the *discrimination index*, first developed for the DINA model (de la Torre, 2008) and later extended for the G-DINA model (de la Torre & Chiu, 2016). The rationale of this method is that a correctly specified q-vector will clearly distinguish the different latent groups for that item ($2^{K_j^*}$) by their probabilities of success. On the contrary, a misspecified q-vector will lead to more homogeneous probabilities of success across the specified latent groups. Thus, the correct q-vector for each item will be the one that maximizes the variance of the probabilities of success between the different latent groups of that item (i.e., the q-vector that results in the most discriminative item possible). This will translate in a more accurate classification of people's attributes.

At this point, it is important to note the trade-off between the fit and parsimony principles. A model that differentiates more latent classes will have a better fit, because it allows for a wider probabilities of success variability; and a more complex q-vector will result in more latent groups. On the other hand, the parsimony principle dictates that, given the same variability within two different q-vectors, the simplest one should be preferred. Figure 1 provides an illustration of this rationale for the G-DINA model. This figure shows the probabilities of success for one item depending on the number of attributes specified in the q-vector (2, 3, or 4, respectively) and the different latent groups existing in each condition. The two-attribute q-vector (Figure 1a) would not be adequate, as the variability among the latent groups is lower than the one obtained with the three-attribute q-vector (Figure 1b), and so it does not fulfill the fit principle. On the other hand, the four-attribute q-vector (Figure 1c) shows the same variability than the three-attribute one. Nevertheless, it would not be adequate due to the parsimony principle. The following describes how the discrimination index is computed for the DINA and G-DINA models, and how the Q-matrix validation method tries to reach a balance between the fit and parsimony principles.

The simplest case, regarding the number of latent groups, corresponds to the DINA model. In this frame, the discrimination index for item j, named φ_j, compares the probabilities of success of two groups of examinees: group $\eta_{lj} = 1$, formed by those that possess all the required attributes for item j, and group $\eta_{lj} = 0$, formed by the rest of them. The correct q-vector will be the one that maximizes the difference in the probabilities of success between both groups. In a formal way, defining a *slip* parameter (s_j) as the probability of failing the item for $\eta_{lj} = 1$, or $P(X_j = 0|\eta_{lj} = 1)$, and a *guessing* parameter (g_j) as the probability of succeeding the item for $\eta_{lj} = 0$, or

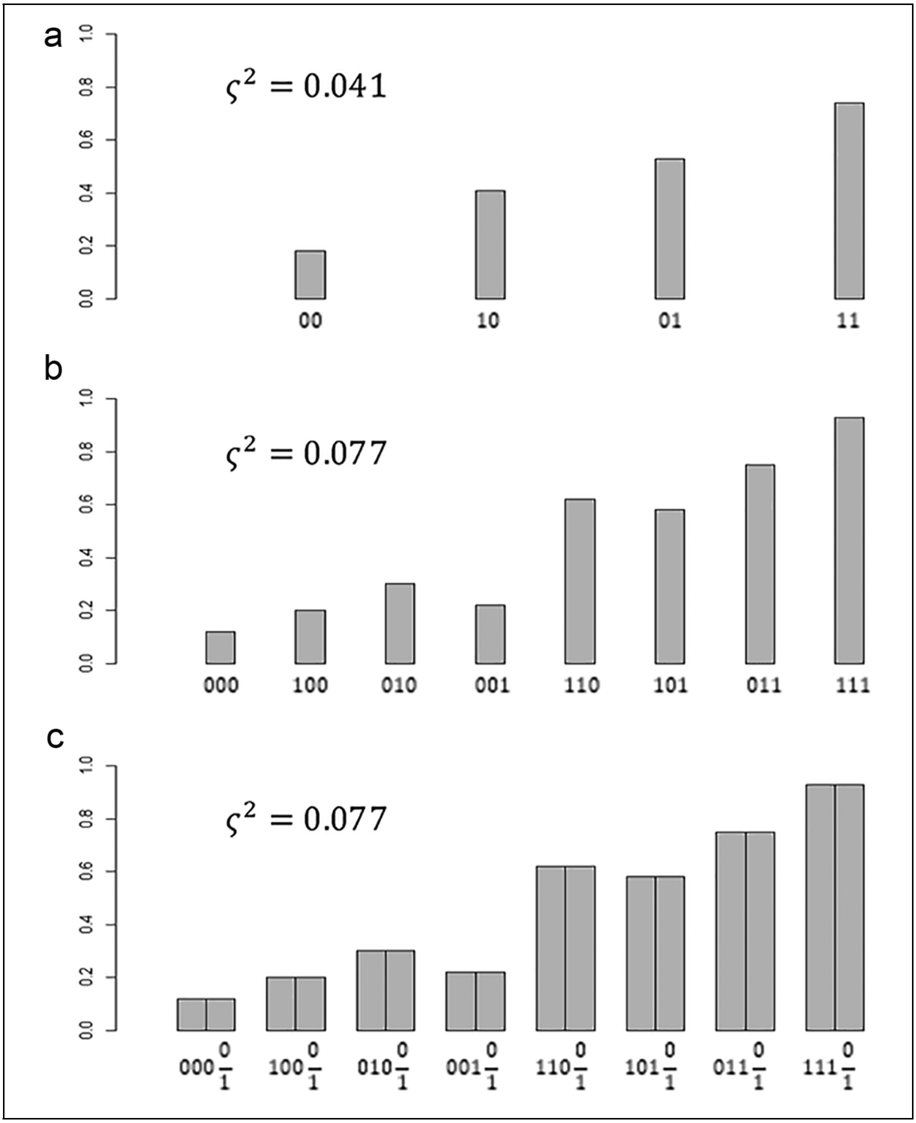


Figure 1. Probabilities of success for one item dependent on K^* and the latent class for the G-DINA framework. "a" and "b" refer to a q-vector with two and three attributes specified, respectively. "c" refers to a q-vector with four attributes specified, the last of them being irrelevant; for each latent class defined with the first three attributes, the two bars represent the probabilities of success dependent on the value of the fourth attribute that, in this case, are the same. The variance of the probabilities of success is represented by ς^2. All latent groups have the same size in this example.

$P(X_j = 1|\eta_{lj} = 0)$, then the correct q-vector will be the one that maximizes $1 - s_j - g_j$ (de la Torre, 2008).

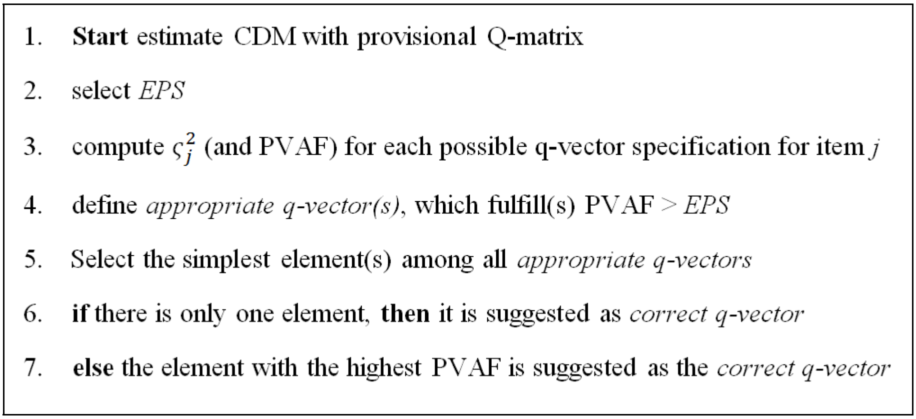
On the other hand, the most general case corresponds to the G-DINA model. In this frame, the discrimination index for item j is represented as ς_j^2, the variance of the probabilities of success of the different latent groups weighted by the posterior distribution of those groups. The formal definition of ς_j^2 is

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

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

In this case, the q-vector with all the attributes specified ($\boldsymbol{q}_{j_{1:K}}$; i.e., the one that establishes that all the attributes are relevant for succeeding the item) will always show the highest ς_j^2 (i.e., $\varsigma_{j_{1:K}}^2$), as the specification of additional attributes leads to the differentiation among more latent groups, and so to a higher variability in the probabilities of success. Nevertheless, this higher variability can be spurious. Thus, aiming to find a balance between the fit and parsimony principles, the correct q-vector will be the simplest one that can explain an important part of the highest variance (i.e., the one that has a ς_j^2 value close to $\varsigma_{j_{1:K}}^2$). For this purpose, a *proportion of variance accounted for* (PVAF) is calculated for each possible q-vector. The PVAF is defined as $\varsigma_j^2/\varsigma_{j_{1:K}}^2$. A cutoff, ϵ (also named *EPS*, for *epsilon*), is established for determining an acceptable PVAF. The suggested q-vector will be the one that, fulfilling PVAF $>$ *EPS*, has the lowest number of specified attributes, that is, is the simplest. If two q-vectors, both of them fulfilling PVAF $>$ *EPS*, have the same number of attributes specified, the suggested would be the one with higher PVAF (de la Torre & Chiu, 2016). The interested reader can find the pseudocode of the method's algorithm in Figure 2.

Figure 3 illustrates this for K = 3 and *EPS* = 0.95 with the so-called *mesaplots* (Ma & de la Torre, 2018). The mesaplots represent, in the x-axis, the different resulting q-vectors from the permutation of the K attributes and, in the y-axis, the PVAF associated to those q-vectors. The q-vectors are ordered from lowest to highest PVAF, so the plot depicts a monotonically increasing function with an absolute maximum point equal to 1 corresponding to the q-vector with all the attributes specified. Figure 3a corresponds to a fictitious item whose true q-vector is $\mathbf{q}$ = {010}, that is, only the second attribute is involved in answering the item. In this plot there is only one abrupt increase of the PVAF that divides it in two mesas by the $\mathbf{q}$ = {010} q-vector, the true one. To its left, with very low PVAF, there can be seen the q-vectors that do not have the relevant attribute; to its right lie the q-vectors that, in addition to the relevant attribute, have one or more irrelevant attributes. The slight



```
Start estimate CDM with provisional Q-matrix
select EPS
compute $\varsigma_j^2$ (and PVAF) for each possible q-vector specification for item $j$
define appropriate q-vector(s), which fulfill(s) PVAF > EPS
Select the simplest element(s) among all appropriate q-vectors
if there is only one element, then it is suggested as correct q-vector
else the element with the highest PVAF is suggested as the correct q-vector
```

Figure 2. Pseudocode of the Q-matrix validation method's algorithm (Ma & de la Torre, 2018).

increases in PVAFs of the latter ones are spurious. In this case, and with an *EPS* of 0.95, the suggested q-vector by the validation method would be clearly the **q** = {010} one, which is the simplest q-vector among those with a PVAF higher than 0.95. Figure 3b and 3c illustrate the same concept for two fictitious items whose true q-vectors are **q** = {101} and **q** = {001}, respectively. As it has been said, the chosen *EPS* determines the q-vector suggested by the validation method. For instance, in Figure 3b, an *EPS* of 0.85 would result in a misspecification **q** = {001}. Thus, here the *EPS* of 0.95 is adequate. However, this will not be always the case, and may vary according to different conditions (see the Design section in the Method): in Figure 3c, with an *EPS* of 0.95, the suggested q-vector (**q** = {101}) is not the true one.

Goals of the Present Study

In their article, de la Torre and Chiu (2016) showed the good performance of their validation method with two simulation studies. The authors chose an *EPS* of 0.95, and the results for the G-DINA model showed that the method detected an 80% of misspecified attributes and maintained a 98% of correctly specified attributes. In addition to the good performance, the method has three main advantages: first, as it has been developed in the G-DINA model framework it has a great flexibility and can be applied to other reduced models (e.g., DINA, DINO, NIDA, NIDO); second, the method not only identifies the misspecified q-vectors, but also suggests substitutive q-vectors; third, the method is available in the GDINA package (Ma & de la Torre, 2018) of the statistical software R (R Core Team, 2016) and it has a low computational cost, and so it is one of the most accessible methods.

However, as de la Torre and Chiu (2016) pointed out, their studies have some limitations. First, the explored simulation conditions may be somehow scarce and

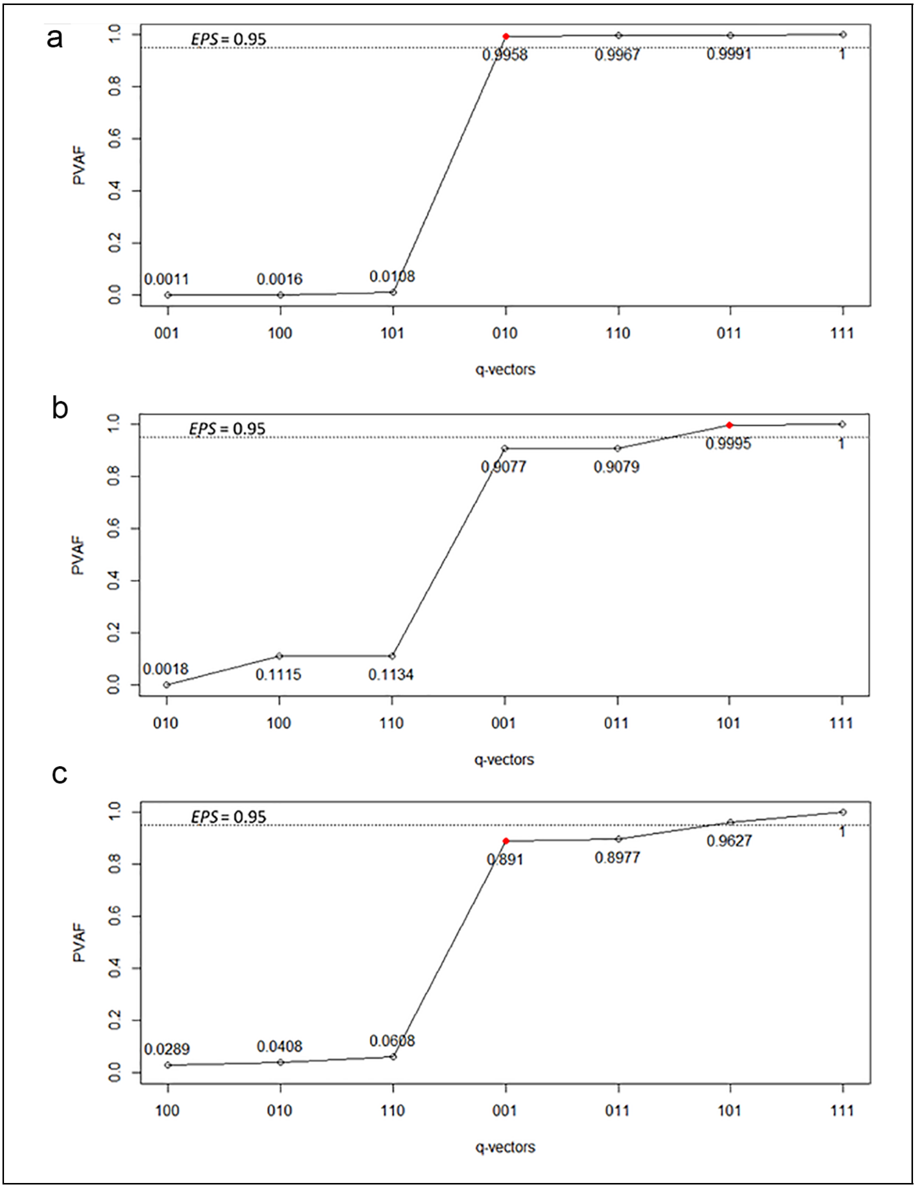


Figure 3. Mesaplots for three items (true q-vectors, respectively: {010}, {101}, {001}).

favorable. As it will be discussed in the design section of the Study 1, even though the sample size (i.e., 2,000 examinees) and item discrimination (i.e., $1 - P(\mathbf{1}) - P(\mathbf{0}) = 0.6$) used by the authors can be found in some applied studies, there are many others in which these conditions are less favorable. Second, it is not

Table 1. Summary of the Simulation Conditions.

Factors	Study 1	Study 2
Attribute structure	Uniform	
J	15, 30, 60	
N	500, 1,000, 2,000	
*IQ**	0.4, 0.6, 0.8	
EPS	0.60; 0.65; 0.70; 0.75; 0.80; 0.85; 0.90; 0.95; 0.99	
Model	G-DINA	
K	5	
% Q-matrix errors	0	10

Note. **IQ* is calculated as $1 - P(\mathbf{1}) - P(\mathbf{0})$. The probabilities of success for the three different levels of *IQ* are the following. Low $IQ{:}P(\mathbf{1}) = U(0.6; 0.8)$ and $P(\mathbf{0}) = U(0.2; 0.4)$; medium $IQ{:}P(\mathbf{1}) = U(0.7; 0.9)$ and $P(\mathbf{0}) = U(0.1; 0.3)$; high $IQ{:}P(\mathbf{1}) = U(0.8; 1.0)$ and $P(\mathbf{0}) = U(0.0; 0.2)$.

specified any PVAF cutoff point (*EPS*) selection method. The default cutoff point is set to 0.95 to guarantee that the suggested q-vector explains at least a 95% of the maximum variance. Nevertheless, it might not be an optimal *EPS* under some unfavorable conditions (e.g., small sample size, low item quality or short test length), as we will show below. Wang et al. (2018) also noticed this latter limitation, and they used different cutoff points depending on the sample size when comparing the performance of this validation method.

In light of the above, the aim of the present work is to study the performance of the empirical Q-matrix validation method based on the discrimination index for the G-DINA framework by using more extended and representative conditions than the ones explored in the original article. In relation with this objective, it will be considered if it exists a unique *EPS* that can be recommended through the different conditions, and detailed recommendations for the applied researcher will be given if this is not the case. Two simulation studies were conducted to evaluate the performance of the Q-matrix validation method when the Q-matrix is correctly specified (i.e., absence of misspecification errors) (Study 1) and in the presence of misspecification errors (Study 2).

The purpose of Study 1 is to examine whether the validation method introduces incorrect modifications in the Q-matrix when the model is correct and under what conditions this happens more frequently. Study 2 examines the performance of the validation method when misspecifications are introduced in the Q-matrix, covering the possibly typical scenario of applied contexts in which the experts make some mistakes while specifying some of the q-entries. Both studies evaluate what *EPS* shows better results under the different conditions.

Study 1: Q-Matrix Is Correctly Specified

Method

Design. Table 1 summarizes the conditions considered in both studies. The number of attributes was fixed to $K = 5$. Four factors were studied: the *EPS* (with nine levels:

from 0.60 to 0.95, in steps of 0.05, and 0.99) and test length, sample size and item quality (with three levels each). The levels of the latter factors were chosen with the purpose of having representative values of the empirical works. An attempt was made to choose a low, medium, and high level for each of them. A brief description of the rationale used for the levels selection is included in the following.

1. *Test length (J)*: With levels of 15, 30, and 60 items. In the reviewed specialized literature, it is common the use of tests with 11 to 30 items (e.g., Chen, 2017; Chiu, 2013; de la Torre, 2008, 2011; de la Torre & Chiu, 2016; Ma & de la Torre, 2016; Sorrel et al., 2016). We also find applications with more than 30 items (e.g., de la Torre et al., 2015; Templin & Henson, 2006), and even with 90 items (e.g., de la Torre, 2008).
2. *Sample size (N)*: With levels of 500, 1,000, and 2,000 participants. Even though there are studies with high sample sizes, of more than 2,000 examinees (e.g., de la Torre, 2008; de la Torre & Douglas, 2004; Romero et al., 2014), many of them have samples between 700 and 1,300 people (e.g., Chen, 2017; de la Torre, 2011; Ma & de la Torre, 2016) and even of less than 600 examinees (e.g., Chen, 2017; de la Torre & Chiu, 2016; Sorrel et al., 2016; Templin & Henson, 2006).
3. *Item quality (IQ)*: The item quality was operationalized by the discrimination, computed as the difference in the probability of success between the latent group with all relevant attributes, $P(\mathbf{1})$, and the one with none of them, $P(\mathbf{0})$ (e.g., in case of $K_j^* = 3$, $P(\mathbf{1}) = P\left(X_j = 1 | \boldsymbol{\alpha}_{lj}^* = \{111\}\right)$ and $P(\mathbf{0}) = P\left(X_j = 1 | \boldsymbol{\alpha}_{lj}^* = \{000\}\right)$). The item quality levels were set to 0.4, 0.6, and 0.8. These levels are based on the those used in previous simulation studies (e.g., Ma et al., 2016; Sorrel, Abad, Olea, de la Torre, & Barrada, 2017), and average discrimination values found in previous empirical studies. While high-quality items were typically found in the educative assessment area (e.g., Chen, 2017; de la Torre, 2008), in clinical or organizational areas it is more common to find low- or medium-item discrimination (e.g., Sorrel et al., 2016; Templin & Henson, 2006).

A Q-matrix with the same number of one-, two- and three-attribute items was specified for each test length. Table 2 shows the Q-matrix for J = 30 (***Q30***). This Q-matrix was previously used by de la Torre and Chiu (2016). The Q-matrix with J = 60 (***Q60***) was a duplicated ***Q30***, while the Q-matrix with J = 15 (***Q15***) was formed by the subset of items marked with an asterisk in Table 2.

Data Generation. The probabilities of success of the latent group with all the relevant attributes, $P(\mathbf{1})$, and the probabilities of success of the latent group with none of them, $P(\mathbf{0})$, were manipulated to generate items with different quality. For each factor level, the same uniform distribution was simulated for all the items. Specifically,

Table 2. Q-Matrix for the Simulated Data (J = 30).

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

Note. Items marked with an asterisks were chosen for $Q15$.

for low-quality items: $P(\mathbf{1}) = U(0.6, 0.8)$ and $P(\mathbf{0}) = U(0.2, 0.4)$; for medium-quality items: $P(\mathbf{1}) = U(0.7, 0.9)$ and $P(\mathbf{0}) = U(0.1, 0.3)$; and for high-quality items: $P(\mathbf{1}) = U(0.8, 1)$ and $P(\mathbf{0}) = U(0, 0.2)$. For the other latent groups (those with some of the relevant attributes) the probabilities of success were simulated so 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. The distribution of the latent classes was uniform (e.g., Sorrel, Abad, et al., 2017; Sorrel, de la Torre, et al., 2017), since there was no reason to assume a specific structure. 100 replications were generated for each of the 27 resulting conditions after combining the factor levels.

Dependent Variables. To evaluate the performance of the validation method there was used the true positive rate (TPR), which indicates the proportion of correctly specified q-entries that are retained (that is, not modified). All simulations and CDM analyses were performed in R (R Core Team, 2016), using the GDINA package (Ma & de la Torre, 2018).

Results

Figure 4 shows the TPR for each of the combined conditions ($EPS \times J \times N \times IQ$). Within each of the nine panels, result of a specific combination of item quality and test length, three sample size conditions are presented. Overall, TPR tended to increase as the sample size increased, item discrimination increased, and test length decreased. On the other hand, *EPS* of 0.99 showed clearly poorer results compared

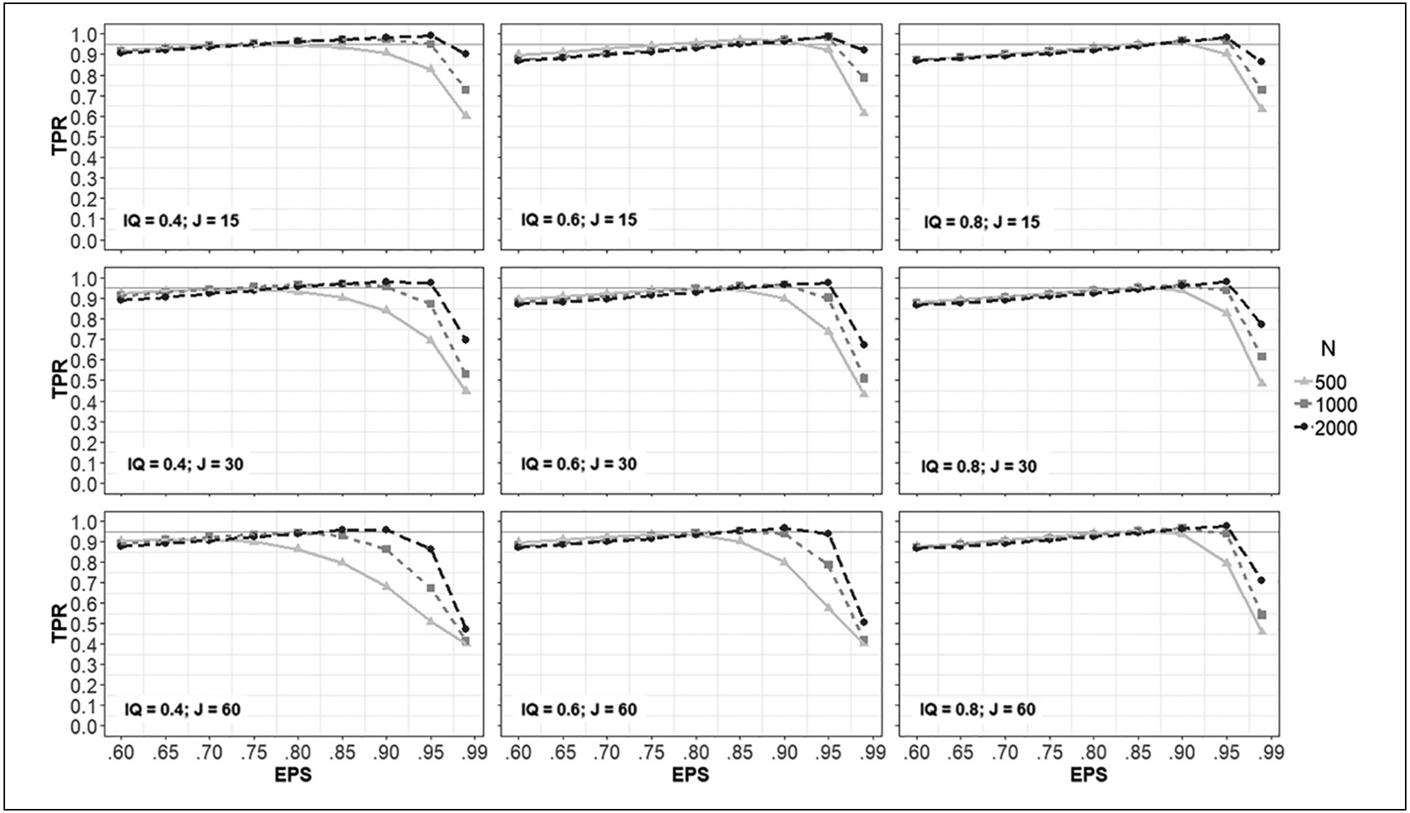


Figure 4. TPR as function of *EPS, N, J* and *IQ*. A reference horizontal line for TPR is included at 0.95 for interpretation purposes.

with the rest of *EPS* values. Omitting this value, a large sample size was associated with high TPR in all conditions at least with two different *EPS* values. The effect of the sample size got larger as the test length increased and, more especially, the item discrimination decreased. Along these lines, in low discrimination conditions the differences between the sample size levels were bigger than in high discrimination conditions. The test length also gained influence in the results as the discrimination decreased. Summarizing, all the studied factors had a relevant effect in the TPR, both independently and combined.

The chosen *EPS* had also a relevant effect in the TPR. Under large sample size conditions, TPR rates were always acceptable, regardless of the other factors (TPR $>$ 0.86). This was especially true when the item discrimination was not low and the test was not long, where higher *EPS* (ranging from 0.85 and 0.95) showed the best results overall (TPR $>$ 0.94). However, under small sample sizes, these high *EPS* showed unacceptable TPR rates, especially with long test length and low item discrimination (TPR rates ranging from 0.51 to 0.80). In these cases, the optimal *EPS* value (i.e., the one that results in the highest possible TPR) tended to be smaller (TPR $>$ 0.90 with *EPS* lower than 0.80). These results indicate the need of choosing a specific *EPS* depending on the specific conditions to make the validation method have a good performance maintaining the correctly specified q-vectors.

Study 2: Q-Matrix Has Misspecifications

Method

Design. The same factors were manipulated for Study 1 and Study 2 (see Table 1): sample size, item quality, and test length. The difference between both studies is that the Q-matrices used in Study 2 contained a 10% of misspecifications. de la Torre and Chiu (2016) worked with a 5% of misspecifications. The reason for choosing a higher misspecification rate than the one selected by the authors is to provide more generalizable results.

For each of the 100 datasets by condition, a different Q-matrix was generated, keeping always the corresponding misspecification rate. Based on the Q-matrices (***Q15***, ***Q30***, and ***Q60***) defined in Study 1, the modifications were included so the proportion of items measuring a specific number of attributes remained similar. Table 3 shows how those misspecifications were introduced for $J = 15$. ***Qtrue*** represents the true Q-matrix (i.e., the one used for simulating the examinees' responses) and ***Q*** represents the Q-matrix with misspecifications. Each table cell refers to the number of items with K_j^* attributes originally specified in ***Qtrue*** that ended up with K_j^* attributes specified in ***Q***. The diagonal cells correspond to non-modified items. In this case, for instance, from the 5 items that had one attribute originally specified in ***Qtrue***, 3 of them were not modified, while 1 of them ended up with two attributes specified in ***Q*** (i.e., for this item a 0 cell in its q-vector was randomly chosen and changed to 1), and another ended up with three attributes specified in ***Q*** (i.e., two 0 cells were randomly chosen and changed to 1). Then, for $J = 15$, a total of 6 items (8 attributes) were

Table 3. Number of One-, Two-, and Three-Attribute Items in the Generating Q-Matrix (***Qtrue***) and the Q-Matrix With Misspecifications (**Q**) for J = 15.

	K_j^* based on the misspecified **Q**			
K_j^* based on ***Qtrue***	1	2	3	Total
1	3	1	1	5
2	1	3	1	5
3	1	1	3	5
Total	5	5	5	15

Note. K_j^*: number of attributes measured.

misspecified. As it can be seen, the number of items measuring one, two, and three attributes remains the same in ***Qtrue*** and ***Q***. The same rationale was followed for J = 30 and J = 60.

Dependent Variables and Data Analysis. When misspecifications are included in the Q-matrix, the validation method should suggest correct modifications. Two dependent variables were used to evaluate the performance of the validation method: the TPR and the true negative rate (TNR). The definition of TPR is the same than the one given in Study 1 (i.e., proportion of correctly specified q-entries that are retained), whereas the TNR indicates the proportion of misspecified q-entries that are modified (corrected). It is important to note that the TPR and TNR are two complementary quality criteria of the Q-matrix validation method; that is, to consider that the method shows a good performance under a certain condition, both a high TPR and a high TNR are needed. To understand these measures the following must be noticed. From the definition of the TPR is deduced that, if no Q-matrix validation method was used, the TPR will be 1, because no q-entries (the correctly specified ones included) would be modified. Thus, the closer the TPR to 1, the better the performance of the method (i.e., from the researcher's view, the validation method would not change the correctly specified q-entries). On the other hand, the TNR represents the benefit that the validation method provides compared to not using it. The higher the TNR, the better the performance of the method (i.e., the ideal would be to correct as many misspecified q-entries as possible). Considering the previously commented results of de la Torre and Chiu (2016), in the present study there will be considered as adequate values higher than 0.95 for the TPR and higher than 0.80 for the TNR.

In addition to these measures, the proportion of correctly classified attributes (PCA) was included as an accuracy measure. In CDM literature this measure has been referred to as attribute-level classification rate (e.g., Ma & de la Torre, 2018). The PCA shows the proportion of entries (i.e., attributes) correctly classified in the $N \times K$ matrix of attribute classifications. To evaluate the effect of the Q-matrix validation method in the correct classification rates two additional indices were considered:

$$PCA_D = PCA_{Q*} - PCA_Q \tag{3}$$

and

$$PCA_{D_{max}} = PCA_{Q_{true}} - PCA_Q, \tag{4}$$

where PCA_{Q*}, PCA_Q and PCA_{Qtrue} indicate the resulting PCA after estimating the model with the Q-matrix suggested by the validation method (***Q****), the misspecified Q-matrix (***Q***), and the true Q-matrix (***Qtrue***), respectively. The PCA_D index indicates the improvement in the PCA after applying the validation method (with a specific *EPS*). A PCA_D close to zero will indicate that, for that specific condition and *EPS*, PCA_{Q*} is very similar to PCA_Q, that is, that the suggested Q-matrix (***Q****) does not provide a better classification compared with the misspecified Q-matrix (***Q***). For example, if PCA_D is equal to 0.02, then ***Q**** has correctly classified 2% more attributes than ***Q***. This 2% will correspond to a different number of attributes depending on the sample size as it is calculated over the $N \times K$ matrix. It should be noted that this value can also be negative, indicating that ***Q**** leads to a worse classification compared to ***Q***. The PCA_{Dmax} index indicates the upper-limit of improvement that can be achieved under a specific condition. As it can be deduced from the Q-matrices involved in its calculation, the PCA_{Dmax} does not depend on the *EPS* selected.

Finally, a multiple linear regression was used with the aim of finding a predictive formula for estimating an optimal value for *EPS* (the one with which a higher PCA_D is obtained) as a function of the sample size, the test length, and the item discrimination. Model fit was assessed using the corrected R^2 (R_c^2). Due to the bounded nature of *EPS* (with a minimum value of 0 and a maximum value of 1), the logit of the optimal *EPS* was used as the criterion variable (Baum, 2008):

$$\text{logit}(EPS) = \beta_0 + \beta_1 \cdot N + \beta_2 \cdot J + \beta_3 \cdot IQ, \tag{5}$$

where *logit*(*EPS*) is

$$\text{logit}(EPS) = \log\left(\frac{EPS}{1 - EPS}\right). \tag{6}$$

Results

Q-Matrix Attribute Specification Rates (TPR and TNR).

Both TPR and TNR effects are combined in Figure 5. TPR results were similar in Study 1 and Study 2. This was an expected result, as the Q-matrices, with 10% of misspecifications do not differ too much from the true ones (they share the 90% of the q-entries). Regarding the TNR, it showed unacceptable values for the conditions conformed by a short test and low-quality items, short test and medium-quality items, and medium test length and low-quality items (the three upper-left panels), regardless the *EPS* and the sample size ($\text{TNR} \leq 0.56$ for these conditions). Under short test conditions, performance was only acceptable if it was composed of high-quality items and large sample sizes, and

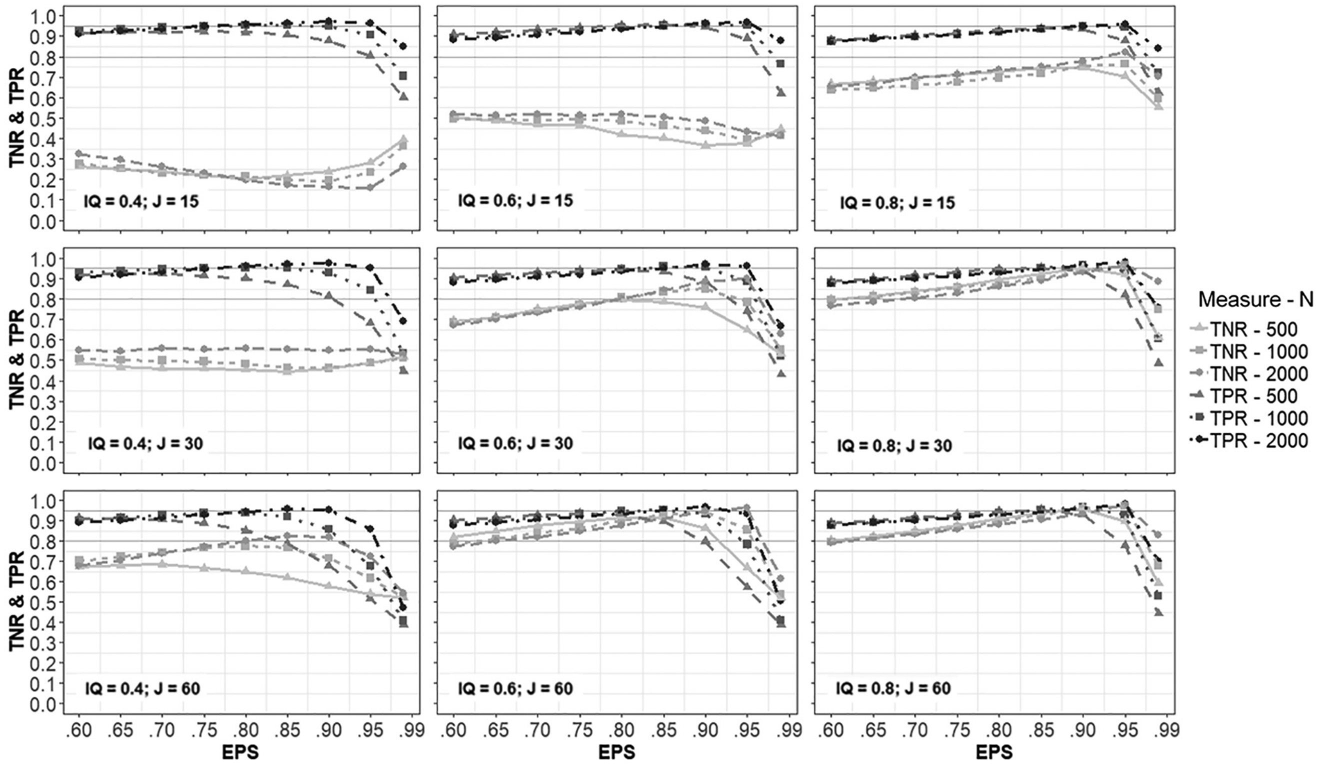


Figure 5. TPR and TNR as function of *EPS, N, J* and *IQ*. Two reference horizontal lines for TPR and TNR are included at .95 and .80, respectively, for interpretation purposes.

the validation method was applied with an *EPS* of 0.95 (see the upper-right panel; TNR = 0.82 for $N = 2{,}000$, $J = 15$, $IQ = 0.8$, and $EPS = 0.95$). Similarly, when the item quality was low, performance was only acceptable under large test length and sample sizes, given an *EPS* between 0.80 and 0.90 (see the lower-left panel; for $N = 2{,}000$, $J = 60$, $IQ = 0.4$, and these *EPS*, TNR ranged from 0.80 to 0.83). For the remaining conditions, when the average item discrimination was equal to or higher than 0.6 and the test length was equal to or higher than 30 items, the validation method showed a good general performance for some specific *EPS*. The *EPS* that tended to provide both the best TPR and TNR results in each condition depended mostly on sample size. If the sample size was large (i.e., 2,000 examinees), *EPS* of 0.90 or 0.95 showed good and similar performance. The same *EPS* were suitable for samples of medium size (i.e., 1,000 examinees), although 0.95 was more appropriate when the item discrimination and the test length were higher. If the sample size was small (i.e., 500 examinees), the optimal *EPS* was 0.85 or 0.90, depending on whether the item discrimination was medium or high, respectively.

Examinees' Attribute Classification Rates (PCA_D). The aforementioned recommendations are based on both the TPR and TNR results. Even though these measures are relevant, it is also important in the decision-making process to have an index that measures directly the impact of the different factors in the examinees' correct classification; that is, the accuracy. Figure 6 shows how the PCA_D varied as a function of the different factors and *EPS.* At this point, *EPS* of 0.99 was removed due to its bad results in the previous sections. The solid lines represent the resulting PCA_D for each condition. The dashed lines represent the maximum PCA_D (PCA_{Dmax}), with the aim of having an upper-limit reference.

The following describes the main results depicted in Figure 6. First, provided that the tests had a medium or large length (i.e., 30 or 60 items) and the items were of medium or high quality (i.e., 0.6 or 0.8), the PCA_D was close to zero. In these cases, the PCA_{Dmax} was also a small quantity, which means that the effect of the Q-matrix misspecifications was less relevant. Thus, it seems that the effect of the possible Q-matrix misspecifications gets mitigated as the item discrimination and the test length increase. That is, under conditions where a lot of information is available, slightly different Q-matrices will lead to very similar results. When conditions were less than ideal (e.g., low discrimination, short test length), Q-matrix misspecifications gained relevance, and PCA_D strongly departed from zero due to a decrease in PCA_Q. In these cases, the modifications can improve or deteriorate the performance, depending on the chosen *EPS* and the sample size. Thus, in terms of classification accuracy, Q-matrix misspecifications will be more important under unfavorable conditions (i.e., small sample size, low item quality, and short test length), as those misspecifications will have a greater influence on the results. This emphasizes the relative relevance of the *EPS* value under these unfavorable conditions.

These findings slightly differ from the TPR and TNR results where more favorable conditions led to a more accurate Q-matrix specification. That is, favorable

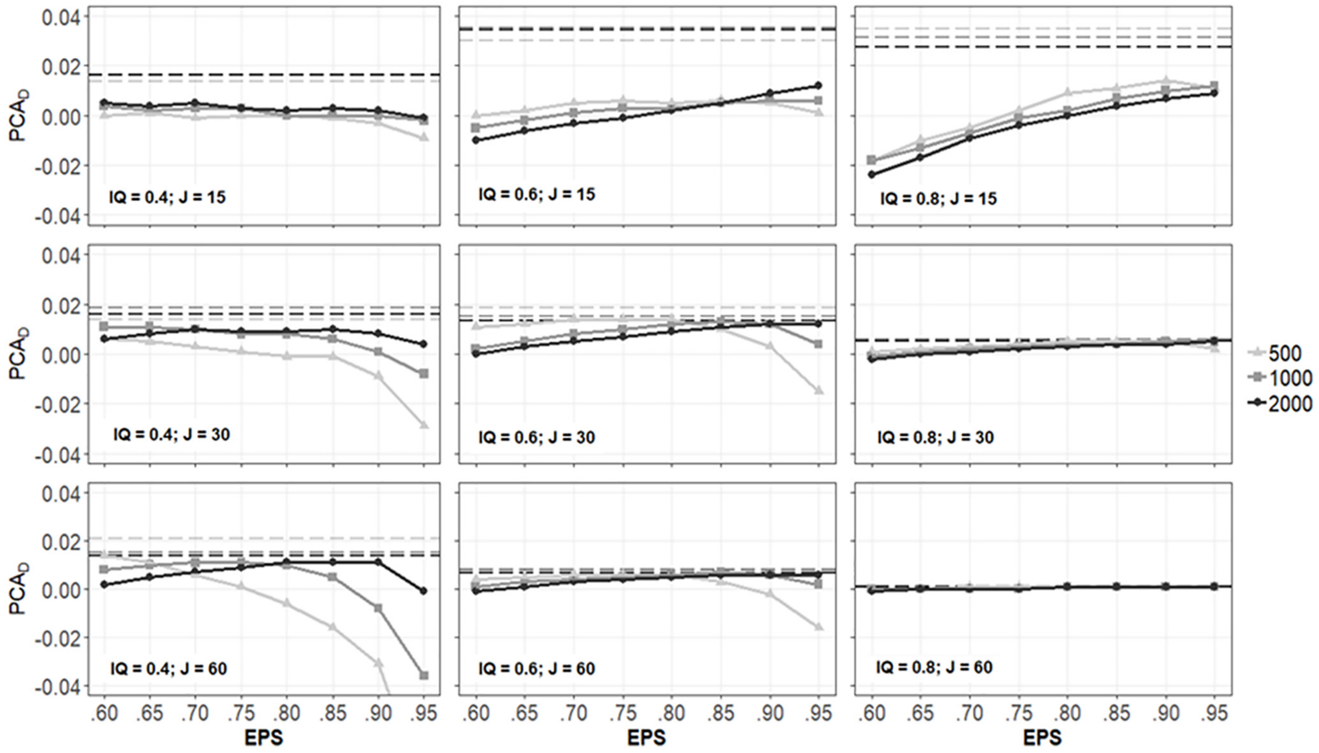


Figure 6. PCA$_D$ as function of *EPS, N, J* and *IQ*. Dashed lines represent PCA$_{Dmax}$ for each condition (*N, J*, and *IQ*). The specific value for *N* = 500, *J* = 60, *IQ* = 0.4, and *EPS* = 0.95 in the lower-left panel is −0.086.

Table 4. Main Results of the Linear Regression of the Logit of the Optimal EPS on *N, J*, and *IQ*.

	Nonstandardized coefficients	Sig.	ΔR^2
Constant	-0.405	<.001	—
IQ	2.867	<.001	0.323
N	$4.840 \cdot 10^{-4}$	<.001	0.134
J	$-3.316 \cdot 10^{-3}$	<.001	0.006

Note. The linear regression model of the *logit* of the optimal *EPS* would be the following: $\text{logit}(EPS) = -0.405 + 2.867 \cdot IQ + 4.840 \cdot 10^{-4} \cdot N - 3.316 \cdot 10^{-3} \cdot J$; thus, the predictive formula of the optimal *EPS* would be the following: $EPS = \text{inv.logit}\left(-0.405 + 2.867 \cdot IQ + 4.840 \cdot 10^{-4} \cdot N - 3.316 \cdot 10^{-3} \cdot J\right)$, where *inv.logit* is the inverse of the logit function. An illustration using R is included in the the Online Appendix.

conditions will result in a better specified Q-matrix, but also in a lower gain in terms of in attribute classification. On the other hand, even though unfavorable conditions can result in some Q-matrix misspecifications, correcting only some of the misspecified attributes can lead to a bigger gain in terms of examinees' classification if the correct *EPS* is chosen.

Regarding the optimal *EPS*, Figure 6 shows that, as the conditions got more favorable, better results were obtained with higher *EPS*. However, if conditions are not favorable (e.g., small sample size), higher *EPS* will provide worse solutions. This is consistent with the TPR ant TNR results. Still, at this point it is important to note that, regarding the correct Q-matrix specification rates (Figure 5) and provided that the conditions were not very unfavorable (i.e., low discrimination and short test length), low *EPS* (0.60-0.70) showed consistently not ideal, but acceptable results. With this information one may think that the use of those *EPS* is generally justified. However, under these same conditions and regarding the examinees' classification accuracy, if the PCA_D results are taken into consideration (Figure 6), those low *EPS* lead to really bad results. In conclusion, there is no *EPS* that performs consistently well across all conditions.

Optimal Cutoff Point Prediction. Table 4 shows the multiple linear regression results as well as the optimal *EPS* predictive formula. Using a stepwise, forward selection method, the three factors were introduced in the model ($p < .001$). According to the coefficient size and the change in R^2, the most relevant factor was the item discrimination, followed by the sample size and, finally, the test length. Model's R_c^2 was 0.462.

Discussion

The Q-matrix is one of the essential inputs in CDM. It establishes the relationship between the items and the attributes, defining the confirmatory nature of these models. An adequate item-attribute correspondence, that is, a correct Q-matrix specification, is indispensable to ensure the best possible examinees' classification. Q-matrix misspecifications can result in a biased estimation of model parameters and classification errors (Rupp & Templin, 2008), which can potentially lead to serious consequences. Thus, in the last years, it has been noticed the need of adding a posterior step to the (subjective) specification of the Q-matrix: its validation.

de la Torre and Chiu (2016) developed a Q-matrix validation method particularly advantageous. Its main benefits are its flexibility to be used with several general or reduced CDMs; its double capability for detecting and correcting misspecified q-vectors; and its high accessibility due to the G-DINA package in R, with a low computational cost (Ma & de la Torre, 2018). However, in their study, the authors explored the method performance under somehow favorable conditions. Moreover, the use of a not empirically based cutoff point may be problematic (Chen, 2017; Wang et al., 2018). Thus, the aim of the present study was to evaluate the performance of this validation method under a wider number of conditions to obtain more generalizable results and to evaluate the performance of different *EPS* across these conditions with the purpose of determining its optimal value depending on test characteristics and sample size. Two simulation studies were conducted to address these goals.

Study 1 covered the performance of the validation method when the true Q-matrix was assumed. First of all, it was pointed out that the use of a unique *EPS* is not justified. The TPR suffered great variations through the different *EPS* values and conditions examined. Thus, the optimal *EPS* value (i.e., the one that showed a greater TPR) was higher as the test length decreased, and the sample size and item discrimination increased. In general, the method showed, for each condition combination, at least one *EPS* in which the TPR obtained adequate values. This shows the good performance of the method under the correct model and the inappropriateness of choosing a unique *EPS* for a generalized use.

In Study 2 there was introduced a 10% of Q-matrix misspecifications to evaluate the performance of the method under more realistic conditions. Considering both the TPR and TNR together, the optimal *EPS* got higher as the three studied factors increased. In addition, the results tended to be better as the optimal *EPS* was higher. This is understandable as the available information increases when the three studied factors increase, helping the validation method to obtain a more correct Q-matrix specification.

de la Torre and Chiu (2016) found a TPR of 0.980 and a TNR of 0.804, using a high sample size (2,000), a medium test length (30), and a medium item discrimination (0.6). The authors used an *EPS* of 0.95. Under the same conditions, the TPR obtained in Study 1 of the present article (there is no TNR due to the absence of Q-matrix misspecifications) was 0.975; in Study 2 the TPR was 0.963 and the TNR was 0.902. The results were very similar, even though Study 2 shows slight differences with the original work. This difference may be due to the higher misspecification rate used here (10% vs. 5%), which favored the detection of more misspecified q-entries, while slightly worsened the method capability of not modifying the correctly specified q-entries.

On the other hand, a dependent variable related to accuracy (PCA_{D}) was also evaluated. This study showed that the different factors modulate the maximum gain when applying the Q-validation method, depending mainly on the baseline quality of the instrument. Interestingly, even though under unfavorable conditions, the method has a worse performing specifying the Q-matrix attributes (lower TPR and TNR), it presents higher gains in terms of attribute classification (higher PCA_{D}). The amount of information available might be a plausible explanation for this result. Favorable conditions provide more information, and a higher amount of information can help mitigate the effect of the (few) Q-matrix misspecifications. Thus, the attribute specification gains will not have a great repercussion in terms of attribute classification improvement because there is a ceiling effect. On the other hand, the resulting accuracy under unfavorable conditions will have a greater dependence on the Q-matrix specification. That is, when there is not enough information from the factors, the Q-matrix gets a greater influence. Although it has not been explored here, this may be also influenced by the Q-matrix misspecification rate. As seen in Figure 6, the PCA_{Dmax} are very low in general. It is possible that higher misspecification rates could result in bigger effects.

Altogether, the results showed that the Q-matrix validation method can be used under each of the studied conditions, but one should be cautious when they are unfavorable. When conditions are favorable, the method will correctly detect and modify Q-matrix misspecifications, while keeping the correctly specified q-entries; however, this will not result in a dramatic attribute classification improvement. On the other hand, under unfavorable conditions, the method will tend to wrongly modify some correctly specified q-entries and will only detect and modify some misspecifications; however, these little corrections will produce a large improvement in terms of attribute classification, justifying the use of the validation method in these cases. One possible exception may be when the sample size is low (i.e., 500). The maximum benefits that can be obtained under this condition, both in terms of Q-matrix specification and attribute classification, are not especially high, regardless of the item discrimination or test length values; furthermore, if the election of *EPS* is not correct, the results can get dramatically worse. Thus, as a general rule it is recommended to have as high as possible sample size. At least it should be equal of higher than 1,000 examinees to reach an acceptable performance.

On the other hand, the *EPS* should be always chosen considering the specific conditions of the study. If the *EPS* of 0.95 is indiscriminately used, serious mistakes can be made, resulting in a poorer classification. In this vein, some authors have suggested formulas to establish the optimal *EPS* depending on different factors. Liu (as cited in de la Torre & Chiu, 2017) suggested the following:

$$EPS = 1 - (\log N)^{-1}, \tag{7}$$

where *log* is the Napierian logarithm.

This formula will lead to an optimal *EPS* between 0.81 and 0.87 with the usual sample sizes (between 200 and 2,000 examinees). de la Torre and Chiu (2017) argue

Table 5. Optimal EPS dependent on *N*, *IQ*, and *J*.

N	*IQ*	*J*	Recommended *EPS*	Predicted *EPS*
2,000	0.6/0.8	–	0.90 – 0.95	0.89 – 0.94
	0.4	30/60	0.85	0.82 – 0.83
		15	0.60 – 0.70	0.84
1,000	0.8	–	0.90 – 0.95	0.90 – 0.91
	0.6	–	0.85	0.83 – 0.85
	0.40	60	0.70 – 0.80	0.74
		15/30	0.60	0.76
500	0.8	–	0.85 – 0.90	0.87 – 0.89
	0.6	–	0.70 – 0.80	0.80 – 0.82
	0.4	–	0.60	0.69 – 0.72

Note. The dashes in the *J* factor squares represent that the optimal *EPS*, under the specified conditions by *N* and *IQ*, are independent of the test length.

that these values that might be suboptimal in the context of their study. Wang et al. (2018) also used different *EPS* regarding the sample size (*EPS* = 0.995 for *N* = 300 and *EPS* = 0.9975 for *N* = 500 or *N* = 1,000). However, as it has been consistently shown through the different simulation conditions, *EPS* of 0.99 does not perform well. Still, it should be noticed that Wang et al. (2018) focused on two reduced models (DINA and rRUM), where results may be slightly better than when using a general CDM, according to the results found by de la Torre and Chiu (2016). Further research is needed to examine the possible different effects of the *EPS* when using a reduced versus a general CDM. Furthermore, we found that other factors apart from the sample size (e.g., test length or item discriminations) should be considered when predicting the optimal *EPS*. Thus, the aforementioned linear regression model proposed in this study (Equation 5) seems more adequate. The model predictions were satisfactory under the different conditions, although it tends to overestimate the optimal *EPS* when the item discrimination is low. The applied researcher may find of interest an illustrative R-script example provided in the Online Appendix in which a Q-matrix validation process is conducted using the optimal *EPS* predictive model.

Based on our findings, some *EPS* election recommendations based on the different factors are provided in Table 5. The *recommended* EPS column shows, for each condition, the range of *EPS* that have proven to provide better results both in terms of correct Q-matrix specification rates (TPR and TNR) and examinees' classification accuracy (PCA_D). On the other hand, the *predicted* EPS column shows, for illustration purposes, some of the optimal *EPS* obtained with the linear regression formula specified in Table 4 (after applying the *inverse logit* function conversion). For example, in the first line of Table 5, 0.89 is the value obtained by applying the *inverse logit* to the result of the formula with *IQ* = 0.6 and *J* = 60, while 0.94 is obtained with *IQ* = 0.8 and *J* = 15. Thus, the general recommendation is to use the *recommended* EPS when the study conditions are similar to the ones evaluated in this work. When this is not the case, the predictive formula can be used to obtain a guiding *EPS*, having in mind that, if the item discrimination is low, it will probably be overestimated. This *EPS* election procedure should be implemented in Step 2 of the algorithm illustrated in Figure 2.

The present study is not without limitations. First, it has been used only one CDM: the G-DINA model. It was chosen because it is a general model that subsumes most of the reduced ones, but its use is not always justified and it is better to use, whenever possible, a reduced model due to its higher efficiency, convergence ease, lower computational cost, lower sample size requirement, and higher parsimony (Ma et al., 2016; Rojas, de la Torre, & Olea, 2012; Sorrel, Abad, et al., 2017). Thus, even though the results under the G-DINA model give a global panoramic of the validation method performance, it would be appropriate to check the idiosyncrasies of other reduced models that can be used more often in the applied context due to their technical facilities. Second, although new relevant factors have been included in the present study, there are many others that can have a great influence and should be also considered in future studies, as the Q-matrix misspecification rate, the latent classes distribution, the numbers of attributes specified in the q-vectors, and the Q-matrix structure, among other factors. The latter one has proven to be very important for the classification accuracy (for further details see, for instance, Liu, Huggins-Manley, & Bradshaw, 2017; Madison & Bradshaw, 2015).

Finally, it would be interesting to compare this validation method with different newer and less known methods in further research. Wang et al. (2018) have been one of the few that have compared this method with others in their study, but, as it has been noted before, they might not have used the most appropriate *EPS*. The method recently suggested by Chen (2017) is also a good candidate for this comparison. In this vein, there have been also developed purely exploratory approaches for the Q-matrix specification, as in Liu et al. (2012), where the Q-matrix is derived from the data. However, the method developed by these authors has some limitations, as it is not suitable when the number of items or the number of attributes is high (Wang et al., 2018).

Improving Q-matrix validation methods can help CDMs to become more reliable and easy to use in applied contexts, especially in those were the Q-matrix specification can be more difficult. This may be the case of the clinical or organizational fields, were the constructs being measured by each item may be more abstract than in some educational contexts (e.g., measuring mathematical operations). In these situations, Q-matrix misspecifications are more likely to appear and, furthermore, item quality is expected to be lower, since it is more difficult to create good items for more abstract attributes. Regarding the Q-matrix validation method examined in this article, if this lower quality is accompanied by small sample sizes, then an *EPS* of 0.95 may lead to a considerable overspecification of the Q-matrix. This can result in a difficult theoretical interpretation of the attributes, making it difficult for the researcher to understand the classification output. By using a more suitable *EPS* for each specific condition, the method can become more useful for studying more abstract attributes under non-ideal conditions.

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 research was partially supported by Grant PSI2017-85022-P (Ministerio de Economia y Competitividad and European Social Fund).

Supplemental Material

Supplemental material for this article is available online.

References

Barnes, T. (2010). Novel derivation and application of skill matrices: The q-matrix method. In Handbook on educational data mining, 159–172. Boca Raton: CRC Press.

Baum, C. F. (2008). Modeling proportions. *Stata Journal*, *8*, 299-303.

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosis teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, *33*(1), 2-14.

Chen, J. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement*, *41*, 277-293.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, *37*, 419-437.

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

Choi, K. M., Lee, Y.-S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*, 1563-1577.

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*, 253-273.

de la Torre, J., & Chiu, C.-Y. (2017). On the consistency of Q-matrix estimation: A rejoinder. *Psychometrika*, *82*, 528-529.

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

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*, 89-97.

de la Torre, J., van der Ark, L. A., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*. Advance online publication. doi:10.1177/0748175615569110

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*, 391-403.

García, P., Olea, J., & de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema*, *26*, 372-377.

Gu, Y., Liu, J., Xu, G., & Ying, Z. (2018). Hypothesis testing of the Q-matrix. *Psychometrika*, *83*, 515-537.

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

Hartz, S., & Roussos, L. (2008, October). *The fusion model for skills diagnosis: Blending theory with practicality* (Educational Testing Service, Research Report, RR-08-71). Retrieved from https://www.ets.org/Media/Research/pdf/RR-08-71.pdf

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

Jaeger, J., Tatsuoka, C., Berns, S., & Varadi, F. (2006). Distinguishing neurocognitive functions in schizofrenia using partially ordered classification models. *Schizofrenia Bulletin*, *32*, 679-691.

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

Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnosis analyses of a reading test. *Educational Assessment*, *18*, 1-25.

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

Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement*, *77*, 220-240.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, *69*, 253-275.

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

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

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement*, *75*, 491-511.

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

R Core Team. (2016). R (Version 3.3) [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.

Ravand, H. (2016). Application of a cognitive diagnosis model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, *34*, 782-799.

Romero, S., Ordóñez, X., Ponsoda, V., & Revuelta, J. (2014). Detection of Q-matrix misspecifications using two criteria for validation of cognitive diagnosis structures under the least squares distance model. *Psicológica*, *35*, 149-169.

Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Work presented in the National Council of Measurement in Education congress, Vancouver, Canada.

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*, 78-96.

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

Sorrel, M. A., de la Torre, J., Abad, F. J., & Olea, J. (2017). Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 13*(Suppl. 1), 39-47.

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

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

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Educational Testing Service, Research Report, RR-05-16). Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.2005.tb01993.x

Wang, W., Song, L., Ding, S., Meng, Y., Cao, C., & Jie, Y. (2018). An EM-based method for Q-matrix validation. *Applied Psychological Measurement*. Advance online publication. doi:10.1177/0146621617752991