ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/346487674

Balancing fit and parsimony to improve Q-matrix validation

Article *in* British Journal of Mathematical and Statistical Psychology · November 2020

DOI: 10.1111/bmsp.12228

CITATIONS
12

READS
82

4 authors:

Pablo Nájera
Universidad Pontificia Comillas
17 PUBLICATIONS **93** CITATIONS
SEE PROFILE

Miguel A. Sorrel
Universidad Autónoma de Madrid
55 PUBLICATIONS **705** CITATIONS
SEE PROFILE

Jimmy de la Torre
The University of Hong Kong
105 PUBLICATIONS **4,675** CITATIONS
SEE PROFILE

Francisco J Abad
Universidad Autónoma de Madrid
141 PUBLICATIONS **4,073** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Developing and validating proof comprehension tests in university mathematics View project

Psychometrics: Considerations for small sample studies View project

All content following this page was uploaded by Pablo Nájera on 31 May 2023.

The user has requested enhancement of the downloaded file.

Title: Balancing Fit and Parsimony to Improve Q-matrix Validation

Short title: *FIT AND PARSIMONY IN Q-MATRIX VALIDATION*

Journal: *British Journal of Mathematical and Statistical Psychology*

Pablo Nájera*[1], Miguel A. Sorrel[1], Jimmy de la Torre[2], and Francisco José Abad[1]

[1] Department of Social Psychology and Methodology, Autonomous University of Madrid, Madrid, Spain.

[2] Faculty of Education, The University of Hong Kong, Pokfulam, Hong Kong.

* Corresponding author information: Pablo Nájera, Department of Social Psychology and Methodology, Autonomous University of Madrid, Ciudad Universitaria de Cantoblanco, Madrid 28049, Spain (e-mail: pablo.najera@uam.es).

This is the peer reviewed version of the following article:

Nájera, P., Sorrel, M. A., de la Torre, J., & Abad, F. J. (2021). Balancing fit and parsimony to improve Q-matrix validation. British Journal of Mathematical and Statistical Psychology, 74(S1), 110–130. https://doi.org/10.1111/bmsp.12228

which has been published in final form at https://doi.org/10.1111/bmsp.12228. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions. This article may not be enhanced, enriched or otherwise transformed into a derivative work, without express permission from Wiley or by statutory rights under applicable legislation. Copyright notices must not be removed, obscured or modified. The article must be linked to Wiley's version of record on Wiley Online Library and any embedding, framing or otherwise making available the article or pages thereof by third parties from platforms, services and websites other than Wiley Online Library must be prohibited.

Data availability statement:

The codes are available under request from the corresponding author.

Acknowledgements:

This research was partially supported by Ministerio de Ciencia, Innovación y Universidades, Spain (Grant PSI2017-85022-P), European Social Fund, and Cátedra de Modelos y Aplicaciones Psicométricas (Instituto de Ingeniería del Conocimiento and Autonomous University of Madrid).

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.

Balancing Fit and Parsimony to Improve Q-Matrix Validation

Abstract

The Q-matrix identifies the subset of attributes measured by each item in the cognitive diagnosis modelling framework. Usually constructed by domain experts, the Q-matrix might contain some misspecifications, disrupting classification accuracy. Empirical Q-matrix validation methods such as the GDI and Wald have shown promising results in addressing this problem. However, a cut-off point is used in both methods, which might be suboptimal. To address this limitation, the Hull method is proposed and evaluated in the present study. This method aims to find the optimal balance between fit and parsimony, and it is flexible enough to be used either with a measure of item discrimination (PVAF) or a coefficient of determination (pseudo-R^2). Results from a simulation study showed that the Hull method consistently obtained the best performance and shortest computation time, especially when used with the PVAF. The Wald method also performed very well overall, while the GDI method obtained poor results when the number of attributes was high. The absence of a cut-off point provides greater flexibility to the Hull method, and it places it as a comprehensive solution to the Q-matrix specification problem in applied settings. This proposal is illustrated using real data.

Key words: CDM, Q-matrix, validation, G-DINA, Hull method, PVAF, R-squared.

Balancing Fit and Parsimony to Improve Q-Matrix Validation

Cognitive diagnosis models (CDMs) are multidimensional discrete latent variable models. Examinees are classified into latent classes or *attribute profiles* according, for the usual case of dichotomous attributes, to their mastery or non-mastery of each attribute. CDMs require a Q-matrix (Tatsuoka, 1983) that defines the relationships between the J items and the K attributes. Each item has an associated q-vector ($\boldsymbol{q}_j$) of length K, in which each q-entry (q_{jk}) equals 1 or 0, depending on whether attribute k is relevant to correctly answer item j or not, respectively. The Q-matrix can be constructed based on theoretical knowledge or based solely on the data. The former is often performed by domain experts who discuss the theoretical structure of the test (e.g., Sorrel et al., 2016), while the latter is conducted by using empirical Q-matrix estimation methods (e.g., Liu et al., 2012). Regardless of what procedure is used, the original Q-matrix is expected to contain some misspecifications (Rupp & Templin, 2008). Q-matrix misspecifications negatively affect the estimation of CDMs parameters, disrupting the subsequent attribute profile classification (Gao et al., 2017; Rupp & Templin, 2008). In order to correct the potential misspecifications, several empirical Q-matrix validation methods have been proposed. These methods are a middle-ground solution between the confirmatory and exploratory perspectives, given that they consider the information from both the data and the original Q-matrix.

Among the Q-matrix validation methods, two are of particular interest: the *general discrimination index* method (GDI method; de la Torre & Chiu, 2016) and the Wald-based stepwise method (Wald method; Ma & de la Torre, 2020a). Apart from their good performance, these methods show some desirable features, such as their applicability to several CDMs and easy accessibility due to its inclusion in the `GDINA` package (Ma & de la Torre, 2020b). Despite these advantages, both methods share a limitation: the use of

a cut-off point in the process of selecting the suggested q-vector (a detailed explanation of the procedures is provided below). The problem associated to fixed cut-off points is that they cannot take into account the variations that the performance of a method suffers as a function of different data conditions (Nájera et al., 2019). In relation to this, the robustness of the GDI and Wald methods should be further evaluated. Although both methods have been examined under a wide range of conditions, some important factors have not been explored in previous simulation studies. For instance, their performance has been tested under 5 attributes; however, the average number of attributes found in applied studies is 8 (Sessoms & Henson, 2018). Moreover, the Wald method has been only evaluated under a fixed test length and the sequential G-DINA model (Ma & de la Torre, 2016).

Considering all of the above, the purpose of the present paper is twofold. First, propose an empirical Q-matrix validation method that, without requiring a cut-off point, can achieve an optimal fit-parsimony balance. Second, examine how the performance of the new proposal compares to that of the GDI and Wald methods under a wide range of realistic conditions by means of a simulation study. The remainder of the paper is laid out as follows: First, a general CDM model is introduced; Second, the rationale of the conventional methods (GDI and Wald methods) and the new proposal (Hull method) is detailed; Third, a simulation study is conducted to compare the performance of the methods; Fourth, a real-data example is presented for illustration purposes; Finally, the implications for applied studies and future research lines are discussed.

Review of the G-DINA model

The *generalized deterministic input, noisy and gate* model (G-DINA; de la Torre, 2011) is a general CDM that subsumes most of the reduced CDMs, whose calibration is a required preliminary step for Q-matrix validation. Let $L = 2^K$ be the number of latent

classes, and $\boldsymbol{\alpha}_l$ be the attribute pattern for latent class l. Then, let $\boldsymbol{q}_j$ be the q-vector of item j, and $\boldsymbol{q}^{(m)}$ be the m q-vector. There are $M = L - 1$ possible q-vectors, since the null q-vector (i.e., with no attributes required) is not plausible. Table 1 shows the $\boldsymbol{\alpha}_{1:L}$ attribute profiles and $\boldsymbol{q}^{(1:M)}$ q-vectors for the case of $K = 3$. Let $\boldsymbol{k}^{(m)}$ and $K^{(m)}$ be the positions and the number of attributes specified in $\boldsymbol{q}^{(m)}$, respectively, and $L^{(m)} = 2^{K^{(m)}}$ be the number of latent groups given $\boldsymbol{q}^{(m)}$. Then, $\boldsymbol{\alpha}_l^{(m)}$ would be the attribute pattern for latent group l given a q-vector with the attributes included in $\boldsymbol{k}^{(m)}$ specified. Note that we refer to a latent class when all K attributes are considered, and to a latent group when considering only the $\boldsymbol{k}^{(m)}$ attributes specified in $\boldsymbol{q}^{(m)}$. Consider again the case of $K = 3$, where $L = 8$ and $M = 7$. Then, for instance, a latent class would be $\boldsymbol{\alpha}_7 = \{0,1,1\}$, which indicates that the seventh latent class possesses both the second and third attributes, but not the first one. Now consider that $\boldsymbol{q}_j = \boldsymbol{q}^{(5)} = \{1,0,1\}$, where $\boldsymbol{k}^{(5)} = \{1,3\}$, $K^{(5)} = 2$, and $L^{(5)} = 4$. A latent group would be $\boldsymbol{\alpha}_2^{(5)} = \{1,0\}$, which indicates that the second latent group given the fifth q-vector possesses the first attribute, but not the third.

According to the item response function of the G-DINA model, the probability of a latent group correctly answering an item, given a certain q-vector, is modelled by the sum of the effects of the attributes involved and their interactions:

$$P_j\left(\boldsymbol{\alpha}_l^{(m)}\right) = \delta_{j0}^{(m)} + \sum_{k=1}^{K^{(m)}} \delta_{jk}^{(m)} \alpha_{lk}^{(m)} + \sum_{k'>k}^{K^{(m)}} \sum_{k=1}^{K^{(m)}-1} \delta_{jkk'}^{(m)} \alpha_{lk}^{(m)} \alpha_{lk'}^{(m)} \ldots + \delta_{j(12\ldots K^{(m)})}^{(m)} \prod_{k=1}^{K^{(m)}} \alpha_{lk}^{(m)}, (1)$$

where $\delta_{j0}^{(m)}$ is the intercept of item j given $\boldsymbol{q_j} = \boldsymbol{q}^{(m)}$, $\delta_{jk}^{(m)}$ is the main effect due to $\alpha_{lk}^{(m)}$, $\delta_{jkk'}^{(m)}$ is the interaction effect due to $\alpha_{lk}^{(m)}$ and $\alpha_{lk'}^{(m)}$, and $\delta_{j(12\ldots K^{(m)})}^{(m)}$ is the interaction effect due to $\alpha_{l1}^{(m)}, \ldots, \alpha_{lK^{(m)}}^{(m)}$.

It is important to emphasise the relevance of the δ parameters. Let $\boldsymbol{q}_j = \{1, 1, 0\}$ be the true q-vector for item j. $\boldsymbol{q}_j$ indicates *what* attributes are being measured by item j,

but not *to what extent*; δ parameters provide that information. Here, if $\delta_{j2}^{(m)}$ and $\delta_{j12}^{(m)}$ were low, none of the effects related to the second attribute would be substantive. This being the case $\boldsymbol{q}_j = \{1, 0, 0\}$ would be the most appropriate q-vector *de facto*. This topic will be brought back again in the Method section.

Conventional Q-matrix validation methods

The GDI method of Q-matrix validation. The GDI method (de la Torre & Chiu, 2016) is based on the concept of item discrimination. Specifically, the GDI (ς_{jm}^2) is defined as the variance of the probabilities of success of the different possible latent groups weighted by the size of those latent groups:

$$\varsigma_{jm}^2 = \sum_{l=1}^{L^{(m)}} \pi\left(\boldsymbol{\alpha}_l^{(m)}|\boldsymbol{X}\right)\left[P_j\left(\boldsymbol{\alpha}_l^{(m)}\right) - \bar{P}_j\left(\boldsymbol{\alpha}_l^{(m)}\right)\right]^2, \tag{2}$$

where

$$\pi\left(\boldsymbol{\alpha}_l^{(m)}|\boldsymbol{X}\right) = \sum_{i=1}^{N} \pi\left(\boldsymbol{\alpha}_l^{(m)}\middle|\boldsymbol{X}_i\right) \tag{3}$$

and

$$\bar{P}_j\left(\boldsymbol{\alpha}_l^{(m)}\right) = \sum_{l=1}^{L^{(m)}} \pi\left(\boldsymbol{\alpha}_l^{(m)}|\boldsymbol{X}\right) P_j\left(\boldsymbol{\alpha}_l^{(m)}\right), \tag{4}$$

where $\pi\left(\boldsymbol{\alpha}_l^{(m)}\middle|\boldsymbol{X}_i\right)$ is the posterior probability of examinee i having attribute pattern $\boldsymbol{\alpha}_l^{(m)}$. Since the ς_{jm}^2 index lacks a known metric, a relative ratio is considered. The maximum variance is obtained by the fully specified q-vector ($\boldsymbol{q}^{(M)} = \{\mathbf{1}\}$; de la Torre & Chiu, 2016). All M q-vectors are ranked-order based on the *proportion of variance accounted for* (PVAF): $\text{PVAF}_{jm} = \varsigma_{jm}^2/\varsigma_{jM}^2$. The PVAF is then enclosed between 0 and 1.

A common way to represent the PVAF is the *mesaplot* (Ma & de la Torre, 2020b). The PVAF is shown in the *y-axis*, and the *candidate q-vectors* are shown in the *x-axis*. The candidate q-vectors are the ones with the highest PVAF among those with a given $K^{(m)}$. The function is monotonically increasing to 1. A reduction in the growing rate is expected after all the relevant attributes have been correctly specified. Thus, the q-vector on the edge of the mesa is likely to be the most appropriate q-vector. An illustration of a mesaplot is shown in Figure 1.

A critical question is how to define that a mesa is "sharp enough". The GDI method tries to solve this problem by setting a cut-off point. Specifically, among the candidate q-vectors, the set of *appropriate q-vectors* is formed by those who fulfil $\text{PVAF}_{jm} > \varphi$.[1] In pursuit of parsimony, the *suggested q-vector* will be the simplest one among the appropriate q-vectors. In the original formulation of the GDI method, φ was fixed to .95. Nájera et al. (2019) noted that the optimal φ value depended on data characteristics and proposed a predictive formula based on the sample size (N), test length (J), and average item discrimination (IQ):

$$\varphi = \text{inv.logit}(-0.405 + 2.867 \cdot IQ + 4.840 \cdot 10^{-4} \cdot N - 3.316 \cdot 10^{-3} \cdot J), \qquad (5)$$

where *inv.logit* is the inverse of the logit function. Even though this resulted in very accurate results, the inclusion of all potentially relevant factors in a predictive formula is not practicable. Hence, the usefulness of such predictions might remain limited to the specific simulation conditions under they were generated.

The Wald method of Q-matrix validation. The stepwise Q-matrix validation method or Wald method (Ma & de la Torre, 2020a) evaluates the statistical significance of the attribute effects in model fit using the Wald test (Wald, 1943). Specifically, the method examines whether an attribute can be excluded from an item without a significant

[1] Note that the cut-off φ has been denoted as ϵ in previous studies (e.g., de la Torre & Chiu, 2016).

loss of fit. An attribute is said to be *statistically necessary* if it cannot be removed from an item without a significant loss. The Wald method always compares a q-vector with $K^{(m)}$ attributes specified with a nested q-vector with $K^{(m)} - 1$ attributes specified. A restriction matrix $\boldsymbol{R}^{(km)}$ is required to evaluate whether attribute *k* is statistically necessary in $\boldsymbol{q}^{(m)}$. Under the null hypothesis (i.e., attribute *k* is not statistically necessary), $\boldsymbol{R}^{(km)} \times \boldsymbol{P}_j^{(m)} = 0$. Here, $\boldsymbol{P}_j^{(m)}$ denotes the vector of success probabilities for the $L^{(m)}$ latent groups to item *j* (see Eq. 1). The Wald statistic for testing whether attribute *k* is statistically necessary for item *j* given a provisional q-vector *m* is:

$$W_{jk}^{(m)} = \left[\boldsymbol{R}^{(km)} \times \boldsymbol{P}_j^{(m)}\right]' \left[\boldsymbol{R}^{(km)} \times \boldsymbol{V}_j^{(m)} \times \boldsymbol{R}^{(km)\prime}\right]^{-1} \left[\boldsymbol{R}^{(km)} \times \boldsymbol{P}_j^{(m)}\right], \qquad (6)$$

where $\boldsymbol{V}_j^{(m)}$ is a $L^{(m)} \times L^{(m)}$ submatrix of the covariance matrix $\boldsymbol{V}(\boldsymbol{P}) = \mathcal{I}(\boldsymbol{P})^{-1}$, being $\mathcal{I}(\boldsymbol{P})$ the information matrix. The Wald statistic is asymptotically χ^2 distributed with $2^{K^{(m)}-1}$ degrees of freedom.

The Wald method implements the Wald test with a stepwise algorithm. For each item, the first attribute to be included in $\boldsymbol{q}_j$ is the one with the highest PVAF among the q-vectors with $K^{(m)} = 1$. If $\text{PVAF}_{jm} > \varphi$, the process terminates and $\boldsymbol{q}_j$ becomes the suggested q-vector. If not, all the q-vectors with $K^{(m)} = 2$ that subsume the simpler provisional $\boldsymbol{q}_j$ are evaluated using the Wald test. If none of the attributes under test are statistically necessary, the process terminates. Otherwise, the attribute with the highest PVAF among all the statistically necessary attributes is included in $\boldsymbol{q}_j$. Then, the Wald test is conducted again to examine whether the previously included attributes are still statistically necessary. Non statistically necessary attributes are removed from $\boldsymbol{q}_j$. The algorithm continues in a stepwise fashion until $\text{PVAF}_{jm} > \varphi$ or no more attributes are found to be statistically necessary.

The Wald method uses both a statistical test (Wald) and a sort of effect size measure (PVAF) to choose the suggested q-vector. Ma & de la Torre (2020a) used $\varphi = .95$ as a cut-off point for the PVAF. This value is arbitrary to some extent and might be suboptimal under different conditions. For instance, a φ down to .75 has been shown to perform better under demanding scenarios (e.g., low quality items; Nájera et al., 2019). It should be noted that the GDI and Wald methods use the cut-off point for different purposes. While the GDI entirely relies on it for the selection of the suggested q-vectors, in the Wald method the statistical test has a primary role for determining what attributes are retainable, while the 0.95 cut-off point is only a high-enough upper-bound for PVAF to prevent for the inclusion of potentially irrelevant, although statistically necessary, attributes. Given all this, it is expected that the disruptive effects associated to the use of a cut-off point might be more notable for the GDI method than for the Wald method.

The Hull method for Q-matrix validation

The Q-matrix validation method proposed in the present paper does not require the use of a cut-off point and aims to find the best fit-parsimony balance. This is achieved by accounting for the complexity of the q-vectors. It is built upon the Hull method developed by Lorenzo-Seva et al. (2011) in the context of factor retention methods in the exploratory factor analysis framework. This method compares different models, from 0 to K factors, in terms of fit and parsimony. All the solutions are depicted in a two-dimensional graph, called *hull plot*, which represents the number of parameters in the *x-axis* and a fit index in the *y-axis*. Different fit indices can be used for this purpose (e.g., CFI, RMSEA, SRMR). The hull plot forms a monotonically increasing curve. The Hull method takes its name after the *convex hull* concept. For the two-dimensional case, it implies that all the solutions placed below a segment connecting any two other solutions are removed; the convex hull is formed by the solutions contained in the most possible

upper curve. After the convex hull is achieved, the optimal solution is defined as the one placed on the most pronounced elbow (i.e., preceded by a big jump and followed by a small jump). To quantify the magnitude of the elbow of each solution, the *st* index (Ceulemans & Kiers, 2006) is computed as

$$st_k = \frac{(f_k - f_{k-1})/(np_k - np_{k-1})}{(f_{k+1} - f_k)/(np_{k+1} - np_k)}, \quad (7)$$

where f_k and np_k denote the fit-index value and the number of parameters associated to the solution with k factors, respectively. The larger the st_k, the bigger the gain in fit per degree of freedom between solution k and the previous solution in comparison with the next solution and solution k. The solution that maximizes *st* is the one retained. Note that the first (i.e., the model with 0 factors) and Kth solutions cannot be selected, since either the previous or posterior solution is not available.

The Hull method for Q-matrix validation follows the same rationale, but some considerations must be made to adapt the method for the CDM framework. First, while the original method evaluates the number of factors underlying a set of variables, the new proposal evaluates the number of attributes to be included in an item's q-vector. Thus, the hull plot varies accordingly. As in the mesaplot, only the best q-vector for each $K^{(m)}$, from 1 to K, is considered as a candidate q-vector and depicted in the plot. The explanation on how the candidate q-vectors are chosen is provided below. The *x-axis* represents the number of parameters associated to each q-vector. For general CDMs, the number of parameters is $np_{K^{(m)}} = 2^{K^{(m)}}$. On the other hand, the *y-axis*, as in the original method, can represent different indices. Two indices are considered in the present study. The first one is the previously defined PVAF. This variant (Hull$_P$) then evaluates the balance between fit, understood as item discrimination, and parsimony. While the relevant attributes are expected to produce an increase in the PVAF, irrelevant attributes

are expected to produce a negligible effect. This will result in the most appropriate q-vector being preceded by a steep slope and followed by a sharp mesa.

The second index considered is the McFadden pseudo-R^2 (McFadden, 1974), which is an absolute model-fit index that measures the proportion of variance accounted for the observed responses. It is a coefficient of determination used in logistic regression models as an analogous index to the squared multiple-correlation coefficient in linear statistical models. In the context of CDM, where probabilities of correctly answering an item are estimated for each examinee, McFadden pseudo-R^2 can be used to obtain a measure of fit between these estimates and observed responses. It is computed as:

$$R^2_{McFadden} = 1 - \frac{\log(L_M)}{\log(L_0)}, \tag{8}$$

where L_M denotes the likelihood of the model, and L_0 denotes the likelihood of the null model. In the Q-matrix validation framework, the model is conditional to the item and q-vector specification. Let x_{ij} be the response (i.e, 0 or 1) of examinee i in item j, $\bar{x}_j$ be the observed mean of item j across the N examinees, and $P_j^{(m)}(\boldsymbol{X_i})$ be the estimated success probability of examinee i in item j given q-vector m:

$$P_j^{(m)}(\boldsymbol{X_i}) = \pi\left(\boldsymbol{\alpha}_l^{(m)}\middle|\boldsymbol{X}_i\right) \times P_j\left(\boldsymbol{\alpha}_l^{(m)}\right). \tag{9}$$

Then:

$$R^2_{jm} = 1 - \frac{\log\left(L_{jm}\right)}{\log\left(L_{j0}\right)}, \tag{10}$$

where

$$L_{j0} = \prod_{i=1}^{N} \bar{x}_j{}^{x_{ij}}\left[1 - \bar{x}_j\right]^{1-x_{ij}} \tag{11}$$

and

$$L_{jm} = \prod_{i=1}^{N} P_j^{(m)}(\boldsymbol{X_i})^{x_{ij}} \left[1 - P_j^{(m)}(\boldsymbol{X_i})\right]^{1-x_{ij}}. \quad (12)$$

Just as in the case before, adding relevant attributes to an item's q-vector is expected to produce an improvement in the prediction of the observed response, while adding irrelevant attributes is expected to result in a negligible increase. This variant (Hull_R) aims to identify the q-vector that obtains the best balance between model fit, understood as a coefficient of determination, and parsimony.

The *st* of Eq. 7 is reformulated as

$$st_{jK^{(m)}} = \frac{(f_{jK^{(m)}} - f_{jK^{(m)}-1})/(np_{K^{(m)}} - np_{K^{(m)}-1})}{(f_{jK^{(m)}+1} - f_{jK^{(m)}})/(np_{K^{(m)}+1} - np_{K^{(m)}})}, \quad (13)$$

where $f_{jK^{(m)}}$ and $np_{K^{(m)}}$ represent the index and number of parameters of the $K^{(m)}$ candidate q-vector for item j, respectively. Index f can be either the PVAF or pseudo-R^2. The $K^{(m)}$ candidate q-vector for item j is the one with the highest PVAF or pseudo-R^2 among the q-vectors with $K^{(m)}$ attributes specified. The algorithm of the Hull method for Q-matrix validation is set as (see Figures 2 and 3 for graphical illustrations):

Step 1. Create a hull plot by representing the number of parameters (np) in the *x-axis* and the index (PVAF or pseudo-R^2) in the *y-axis*. The candidate q-vectors are depicted in the hull plot.

Step 2. Set the *origin* of the hull plot at $np_0 = f_{j0} = 0$, so that the q-vector with $K^{(m)} = 1$ is suitable for election.

Step 3. Remove all the q-vectors that are not part of the convex hull (i.e., remove the q-vectors whose index stays below the line segment connecting any two other q-vectors).

Step 3a. If only the origin and the fully specified q-vector remain, then select the fully specified q-vector.

Step 3b. If two or more q-vectors remain, go to Step 4.

Step 4. Compute the *st* index for each q-vector and retain the one that maximizes *st*.

Simulation Study

A simulation study was conducted to examine the performance of the two variants of the proposed method (Hull_P and Hull_R) under realistic conditions and extend the insights on the performance of the conventional methods (GDI and Wald). Since both Hull_P and Hull_R variants share a lot of common characteristics, we will allude to the Hull method to refer to both of them.

Method

Design. A mixed factorial design was employed. The four Q-matrix validation procedures formed the within-subject factor. Six between-subject factors were manipulated: Q-matrix misspecification rate (*QM*), number of attributes (*K*), item quality (*IQ*), sample size (*N*), ratio of number of items to attribute (*JK*), and attribute distribution (*AD*). Factor levels, represented in Table 2, were elected in pursuit of representativeness of applied settings. For instance, $K = 4$ is the most common scenario in the applied literature, while $K = 8$ is the average (Sessoms & Henson, 2018). The ratio of number of items to attribute indicates the test length as a function of *K*. Hence, four test structures were considered in the present study: $J = 16$ $(K = 4 \times JK = 4)$, $J = 32$ $(K = 4 \times JK = 8)$, $J = 32$ $(K = 8 \times JK = 4)$, $J = 64$ $(K = 8 \times JK = 8)$. The levels for *QM*, *IQ*, *N*, and *AD* are also considered representative of applied settings (Ma & de la Torre, 2020a; Nájera et al., 2019). A total of 144 between-subject conditions resulted after combining the factor levels.

Regarding the implementation of the validation methods, a distinction must be made between *stepwise* and *iterative* procedures. The Wald method was implemented in

a stepwise manner (Ma & de la Torre, 2020a). This method considers the inclusion or exclusion of one attribute at a time. No re-estimation of the CDM is made after introducing the modifications in the Q-matrix. The GDI and Hull methods were implemented iteratively (Nájera et al., 2020). This means that the CDM is re-estimated after each modification. In order to provide a fair comparison to the Wald test, the entire Q-matrix was modified in each iteration introducing or removing the smallest possible number of attributes. This is referred to as *test-attribute* iterative implementation. For instance, consider that the original q-vector for item j in Figures 1, 2, and 3 is $\boldsymbol{q}_j = \{11111\}$. The methods would select $\boldsymbol{q}_j = \{11001\}$ as the suggested q-vector. However, the test-attribute iterative implementation would first suggest $\boldsymbol{q}_j = \{11011\}$, since it is closer to the original q-vector. If $\boldsymbol{q}_j = \{11001\}$ is indeed the most appropriate q-vector, it will be likely modified in a subsequent iteration. This implementation provides greater stability to the validation methods, while giving more weight to the original Q-matrix specification. All validation methods were conducted after estimating a G-DINA model using the original Q-matrix.

Data generation. Examinees' responses were simulated under the G-DINA model. Attribute patterns were generated following either a uniform distribution or a higher-order distribution (de la Torre & Douglas, 2004). For the latter, examinees' continuous latent trait (i.e., $\boldsymbol{\theta}$) were drawn from a standardized normal distribution, attribute discrimination parameters were drawn from a uniform distribution (i.e., $Unif(1, 2)$), and attribute difficulty parameters were given equidistant values between –1.5 and 1.5 (Ma & de la Torre, 2020a).

Item quality (IQ) is defined as $IQ = \sum_{j=1}^{J} \left(P_j(\mathbf{1}) - P_j(\mathbf{0}) \right) / J$, where J is the number of items, and $P_j(\mathbf{1})$ and $P_j(\mathbf{0})$ are the probabilities of correctly answering item j for the latent group that possesses all or none of the attributes involved in item j,

respectively. The three levels of IQ were defined as: high IQ: $P_j(\mathbf{0}) \sim U(0,.2)$ and $P_j(\mathbf{1}) \sim U(.8,1)$; medium IQ: $P_j(\mathbf{0}) \sim U(.1,.3)$ and $P_j(\mathbf{1}) \sim U(.7,.9)$; and low IQ: $P_j(\mathbf{0}) \sim U(.2,.4)$ and $P_j(\mathbf{1}) \sim U(.6,.8)$. This resulted in $IQ \approx .8,.6,.4$ for high, medium, and low IQ, respectively. The probabilities of success of all the other latent groups were randomly generated with two constraints. First, the item response function had to be monotonic on the number of attributes. Second, the sum of the δ parameters associated to an attribute was constrained to be higher than .15. This guarantees that all attributes have a non-negligible effect, as stated on the Review of the G-DINA model section.

True Q-matrices were randomly generated with the following constraints: a) each Q-matrix contained, at least, two identity matrices; b) the composition of the Q-matrix comprised 50% of one-attribute q-vectors, 25% of two-attribute q-vectors, and 25% of three-attribute q-vectors; c) apart from the two identity matrices, each attribute was measured at least by another item; d) the maximum correlation between attributes in the Q-matrix was .3, to avoid attribute overlapping. This was expected to allow for a certain degree of randomness in the original Q-matrices, thus increasing the generalizability of the results, while controlling for extreme scenarios. This Q-matrix generation procedure is also in agreement with the recommendations for identifiability by Xu and Shang (2018). We then proceeded to include misspecifications in these Q-matrices for the conditions of $QM > 0$. Misspecifications were introduced randomly with two constraints. First, all items had to be measured by at least one attribute. Second, one of the identity matrices was always retained.

For each of the 144 conditions, 500 datasets were generated. A new set of true Q-matrix and delta parameters were generated for each dataset. All simulations and analyses were conducted in R software (R Core Team, 2019), using the `GDINA` package and self-

developed functions. All figures were created with the `ggplot2` package (Wickham, 2016). The codes are available under request.

Dependent variables. The Q-matrix recovery rate (*QRR*) was the main dependent variable. For each replication, it was calculated by

$$QRR = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} I\left(q_{jk}^{(s)} = q_{jk}^{(t)}\right)}{J \times K}, \tag{14}$$

where $I(\cdot)$ is the indicator function, and $q_{jk}^{(s)}$ and $q_{jk}^{(t)}$ are the suggested and true q-entries, respectively, for item j and attribute k. This variable provides information on overall performance. Additional variables were calculated to gather more specific information. The true positive rate (*TPR*; i.e., specificity) and true negative rate (*TNR*; i.e., sensitivity) were computed as

$$TPR = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} I\left(q_{jk}^{(s)} = q_{jk}^{(t)} \middle| q_{jk}^{(o)} = q_{jk}^{(t)}\right)}{\sum_{j=1}^{J}\sum_{k=1}^{K} I\left(q_{jk}^{(o)} = q_{jk}^{(t)}\right)} \tag{15}$$

and

$$TNR = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} I\left(q_{jk}^{(s)} = q_{jk}^{(t)} \middle| q_{jk}^{(o)} \neq q_{jk}^{(t)}\right)}{\sum_{j=1}^{J}\sum_{k=1}^{K} I\left(q_{jk}^{(o)} \neq q_{jk}^{(t)}\right)}, \tag{16}$$

where $q_{jk}^{(o)}$ is the original q-entry for item j and attribute k. The number of over-specifications (*OS*; i.e., 0 to 1 misspecifications) and under-specifications (*US*; i.e., 1 to 0 misspecifications) were also considered:

$$OS = \sum_{j=1}^{J}\sum_{k=1}^{K} I\left(q_{jk}^{(s)} > q_{jk}^{(t)}\right) \tag{17}$$

and

$$US = \sum_{j=1}^{J}\sum_{k=1}^{K} I\left(q_{jk}^{(s)} < q_{jk}^{(t)}\right). \tag{18}$$

Two dependent variables were registered to evaluate the classification accuracy derived from each method: the proportion of correctly classified attributes (*PCA*) and the proportion of correctly classified vectors (*PCV*):

$$PCA = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} I(\hat{\alpha}_{ik} = \alpha_{ik})}{N \times K} \tag{19}$$

and

$$PCV = \frac{\sum_{i=1}^{N} I(\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i)}{N}, \tag{20}$$

where $\hat{\alpha}_{ik}$ and α_{ik} are the estimated (with the suggested Q-matrix) and true *k* attribute for examinee *i*, respectively; and $\hat{\boldsymbol{\alpha}}_i$ and $\boldsymbol{\alpha}_i$ are the estimated and true attribute profile for examinee *i*, respectively. Finally, the computation time in seconds and convergence rate for each method were recorded. Convergence issues for the iteratively implemented methods (i.e., GDI and Hull) occurred when they achieved the maximum number of iterations (i.e., 150). When this happened, the suggested Q-matrix in the last iteration was taken as the definitive one. Finally, both independent and repeated measures ANOVA were conducted to examine the factor interactions. Interactions with a partial eta-squared (η_p^2) higher than .14 were considered as relevant (Cohen, 1988).

Results

Medians instead of means are provided for *QRR*, *TPR*, *TNR*, *PCA*, and *PCV* due to the presence of asymmetry. Table 3 gives the overall results for the four validation methods. The Hull method provided the best results in almost all dependent variables, with a consistent, slightly better performance obtained when the PVAF was used rather than the pseudo-R^2. Both variants obtained a very high overall Q-matrix recovery of .953 and .945, respectively. The Wald method also obtained a very good performance, similar to the Hull method for most dependent variables. It showed a tendency to under-specify ($US = 14.0$) but committed very few over-specification errors ($OS = 3.9$). Finally, the

GDI method obtained the worst performance in all dependent variables, with the only exception of under-specification errors ($US = 4.1$).

Table 4 gives the results split by factor levels. A separate ANOVA for each method was conducted to better understand the effect of each factor on each method, using the *QRR*, *TPR*, and *TNR* as dependent variables. The *QRR* of the Wald and Hull methods was most influenced by the misspecification rate ($\eta_p^2 \geq .591$) and item quality ($\eta_p^2 \geq .587$). The more demanding levels of both factors (i.e., $QM = .3$ and $IQ = .4$) led to the worst results for these methods. On the other hand, the GDI method was greatly influenced by the number of attributes ($\eta_p^2 = .946$), in such a way that its poor overall performance was primarily due to $K = 8$. Regarding the *TPR* and *TNR*, the *TPR* remained more stable than the *TNR* under the different factor levels for the Wald and Hull methods. Thus, the *TPR* was always high, even with low quality items ($TPR \geq .926$), while the *TNR* obtained lower values ($TNR \geq .421$). Finally, the sample size, ratio of number of items to attribute, and attribute distribution showed small to moderate effects on all methods.

The best performance across the different factor levels and dependent variables was generally obtained by the Hull method. Both variants, Hull_P and Hull_R, showed an almost identical pattern on the results, with the former being better in most of the conditions. Thus, unless otherwise indicated, results for Hull_P will be the ones described. The Hull method obtained the best *QRR*, *TPR*, and *TNR* under almost all factor levels. It showed a $TPR > .950$ and a TNR closed to or above .800 under all conditions, except for low-quality items ($TNR = .429$) and low ratio of number of items to attribute ($TNR = .658$). The best *PCA* and *PCV* results were always obtained by the Hull method, except for the conditions with $QM = 0$ and $IQ = .4$. Under the conditions in which the Hull method did not performed the best, it obtained very close results to the best performing

method. The Wald method also obtained a good performance under most conditions. It showed a very low over-specification tendency, but the highest under-specification tendency, especially with $K = 8$. On the contrary, the GDI method showed an extreme tendency to over-specify under $K = 8$. However, with $K = 4$, the GDI method performed as well as the other methods, with even a slightly better *PCA* and *PCV* compared to the Wald method. The computation time for all methods was very short with $K = 4$ ($CT \leq 3$ seconds), but dramatically increased with $K = 8$. Under these conditions, the Hull method was the fastest, six minutes below the Wald method. Finally, convergence rates were very high and consistent for the Wald and Hull methods ($CR \geq .972$). The GDI method obtained a convergence rate close to 0 with $K = 8$.

In order to further examine whether either the Hull or Wald method should be preferred under different scenarios, a repeated measures ANOVA was conducted to examine the factor level interactions regarding the *QRR*. The Hull$_P$ variant was used for this analysis. The most relevant interactions were $Method \times K \times IQ$ $\left(\eta_p^2 = .214\right)$ and $Method \times QM \times IQ$ $\left(\eta_p^2 = .205\right)$. Both interactions are summarized and depicted in Figure 4. The performance of both methods tended to be more similar as *IQ* increased and *QM* decreased. In line with previous results, the Hull method provided an overall better performance than the Wald method. The only notable exception was the conditions of $IQ = .4$, $K = 8$, and $QM > 0$, where the Wald method obtained better results. An additional analysis revealed that, under these conditions, the differences in median *QRR* between the Wald and Hull$_P$ methods were pronounced with $JK = 4$ and $N = 500$ (a difference of .117), but much lower with $JK = 8$ and $N = 1000$ (a difference of .012). Thus, the better performance of the Wald method occurred mainly under the most demanding scenarios of $IQ = .4$, $K = 8$, $JK = 4$, $N = 500$, and $QM > 0$.

Real Data Analysis

The Hull method in conjunction with the PVAF has been shown to have better overall performance using simulated data. To further examine its practical viability, an illustration using a real data is provided. The dataset employed for the illustration was previously analysed by Chen and de la Torre (2014). It consists of the dichotomized responses of 2012 students from the United Kingdom who had answer at least half of the 26 items from Booklets 8 and 9 of the PISA 2000 reading assessment (OECD, 2006). Only fully correct answers were considered as successes, whereas partially correct and incorrect answers were treated as failures. The Q-matrix employed in their study is represented in Table 5. It involves six attributes, namely: *locating information* (α_1), *forming a broad general understanding* (α_2), *developing a logical interpretation* (α_3), *evaluating a number-rich text with number sense* (α_4), *evaluating the quality or appropriateness of a text* (α_5), and *related to test speediness* (α_6). Additional details about sample characteristics and attribute definitions can be found in Chen and de la Torre (2014).

As in the simulation study, the Hull$_P$ procedure was applied in a *test-attribute* iterative manner. Only two iterations were required to achieve a stable solution. Modifications were suggested for five items – an attribute was suggested to be added for two items, whereas an attribute was suggested to be dropped for the other three items. For illustration purposes, consider Item 2, which refers to a graph that represents the depth of Lake Chad through several millennia. Specifically, the item asks: "In about which year does the graph in Figure 1 start?" In the original Q-matrix, this item was specified to measure attributes α_1 (i.e., locating information) and α_4 (i.e., evaluating a number-rich text with number sense). Figure 5 depicts the hull plot for this item. The plot shows a very sharp mesa after the first q-vector, $\boldsymbol{q}_j = \{000100\}$, with had a PVAF of .9826. The original q-vector, $\boldsymbol{q}_j = \{100100\}$, with a PVAF of .9841, was not even the best among

the q-vectors with two required attributes. The best two-attribute q-vector, $\boldsymbol{q}_j = \{001100\}$, had a PVAF of .9930, which is very close to that of the first q-vector. As a result, the *st* indices of $\boldsymbol{q}_j = \{000100\}$ and $\boldsymbol{q}_j = \{001100\}$ were 94.5 and 4.8, respectively. Item 1, which is very similar to Item 2, asks, "What is the depth of Lake Chad today?" The original specification of Item 1, which remained unchanged after the implementation of the Hull$_P$ procedure, did not require α_1. Taken together, these two items suggest that α_1, which focuses on finding textual keys, such as enumerations or examples, is not as relevant in measuring the recovery of graphical or numeric information, which is already covered by α_4.

Regarding model fit, the G-DINA model based on the suggested Q-matrix obtained a better relative fit ($AIC = 50249.7$ and $BIC = 51298.2$) compared to that of the original Q-matrix ($AIC = 50328.1$ and $BIC = 51365.4$). This provides additional empirical evidence to support the use of the suggested Q-matrix. Note, however, the results from the Hull$_P$ procedure cannot guarantee that the suggested Q-matrix is correct or appropriate. To arrive at a theoretically and empirically defensible Q-matrix, domain experts need to review the suggested modifications to decide which changes to adopt.

Discussion

CDMs rely on correctly specified Q-matrices to achieve accurate attribute profile classifications (Gao et al., 2017; Rupp & Templin, 2008). Among the empirical Q-matrix validation methods, the GDI method (de la Torre & Chiu, 2016) and the Wald method (Ma & de la Torre, 2020a) have shown promising results and desirable features. A common drawback of both methods is that they require a cut-off point, which can be either selected arbitrarily or by means of unexhaustive predictive formulas. This is the case of the predictive formula developed for the GDI method by Nájera et al. (2019). In this case, the formula was developed under the condition of $K = 5$; hence, its

performance under a higher number of attributes remained uncertain. On the other hand, the cut-off point in the Wald method, which has been fixed to .95, was not expected to excessively disrupt the performance of the method. However, the performance of the Wald method remained also uncertain under some realistic conditions, such as varying test lengths or high number of attributes (Sessoms & Henson, 2018).

The present study proposes the Hull method for Q-matrix validation, which does not require a cut-off point. This feature is expected to provide a greater robustness to the method under varying conditions. The Hull method was implemented with two different indexes: a measure of item discrimination (PVAF) and a coefficient of determination (McFadden's pseudo-R^2). A simulation study was conducted with the aim of examining the performance of the aforementioned methods under realistic conditions.

The Hull method obtained the best results in almost all the dependent variables and factor levels considered in the simulation study. The Hull$_P$ variant obtained an overall $QRR = .953$, with a high overall specificity and sensitivity, which resulted in the highest classification accuracy. Furthermore, these results were achieved while having the lowest computation time (an average of 66 seconds) and a convergence rate of approximately 1. The Hull$_R$ variant obtained very similar but slightly poorer results. It should be noted that an adjusted McFadden pseudo-R^2 has been proposed in the literature (McFadden, 1979), which considers the complexity of the model. By using the adjusted pseudo-R^2, no *st* index computation would be required, since the suggested q-vector would be the one with the highest value. This method was examined and dismissed in a pilot simulation study because the Hull$_R$ procedure provided much better results. The Wald method also performed very well under most conditions, with an overall performance that was close to that of the Hull method. Finally, the GDI method performed very well under $K = 4$, with its results being comparable to the Wald method. These results are in line with Nájera

et al. (2019). However, the predictive formula led to an over-estimation of the cut-off point under $K = 8$, leading to the non-convergence of the GDI method. A more comprehensive predictive formula for the cut-off point could be developed, with the precaution of noticing that it might provide suboptimal results under unconsidered scenarios.

Regarding the factors under study, item quality and Q-matrix misspecification rate were the most relevant ones for the Hull and Wald methods. The great effect of both factors in the final Q-matrix recovery emphasizes the importance of the original item and Q-matrix development process. This is good news for the applied researcher, since the factors that are under her/his control are the most influential ones. Other than that, the Hull and Wald methods were robust to different scenarios of sample size and attribute distribution.

Another important finding is that the lower *QRR* achieved by the Hull and Wald methods under the most demanding conditions (e.g., $IQ = .4$, $QM = .3$) was due to a decrease in the *TNR*, while the *TPR* remained high. Again, this is good news for the practitioner: one can have reasonable expectations that these methods will not incorrectly modify the q-entries that were correctly specified in the Q-matrix, even under challenging conditions. In return, it should be expected that, under these demanding conditions, the methods will not detect all the potentially misspecified q-entries.

According to the comparison between Hull and Wald methods, the Wald method is only recommended when the conditions are particularly unfavourable (i.e., high number of attributes, low item quality, short test length, and low sample size). Other than that, the Hull method is recommended given its great overall performance across different dependent variables. As it has been shown, the Hull method has the advantage that it can be applied iteratively at a reasonable computational cost. Although the Wald method

could be also implemented iteratively, the computation time would grow exponentially as the number of attributes increases. Considering $K = 8$, the Wald method was already more than six minutes slower than the Hull method on the average. If the Wald method were performed iteratively, the computation time would increase to about 8.5 minutes per iteration. This can be unfeasible for even higher-dimensional datasets. On another note, there is still an unsolved limitation in both empirical Q-matrix estimation and validation methods, and the Hull method is no exception: they require the number of attributes to be set in advance. This potential source of misspecification, widely studied in the exploratory factor analysis framework, has been only tentatively addressed in few CDM applied studies (Robitzsch & George, 2019; Xu & Shang, 2018). A systematic evaluation on how to empirically determine the number of attributes should be conducted to provide more practical usefulness to empirical Q-matrix estimation and validation methods. Furthermore, the GDI and Wald methods have been recently evaluated in the context of estimating the q-vector of new items given a partially specified Q-matrix, which is of special relevance for computerized adaptive testing (Wang et al., 2020). The performance of the Hull method could be also further evaluated in this context. Finally, in line with the pseudo-R^2 index, the residual-based approach has been previously considered for Q-matrix validation. The statistic proposed by Yu & Cheng (2019) for the reduced DINA model could be further developed for the G-DINA model and applied within the Hull method.

In conclusion, the use of the *st* index to obtain a fit-parsimony balance in q-vector suggestions without the requirement for a cut-off point has been proven to provide good results under different conditions. The Hull method is a simple, yet powerful method that can serve as a comprehensive solution for the Q-matrix misspecification problem in many applied scenarios. Given its slightly, but consistently better performance, the PVAF is

recommended rather than the pseudo-R^2 index. As was shown in the empirical illustration, the modifications suggested by the method led to a better model fit. However, the theoretical interpretation and adequacy of such suggestions should be considered by domain experts. In this line, we would like to emphasize that Q-matrix validation methods should not be blindly trusted. As indicated in the caveats and recommendations made by Nájera et al. (2020), Q-matrix validation methods should not be understood as a substitute for experts' judgment, but a complement to it.

References

Ceulemans, E., & Kiers, H. A. L. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical and Statistical Psychology*, *59*, 133–150. DOI: 10.1348/000711005X64817

Chen, J., & de la Torre, J. (2014). A procedure for diagnostically modeling extant large-scale assessment data: The case of the Programme for International Student Assessment in reading. *Psychology*, *5*, 1967–1978. DOI: 10.4236/psych.2014.518200

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199. DOI: 10.1007/S11336-011-9207-7

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 528–529. DOI: 10.1007/s11336-015-9467-8

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

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. DOI: 10.21031/epod.332712

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

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*, 340–364. DOI: 10.1080/00273171.2011.564527

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. DOI: 10.1111/bmsp.12070

Ma, W., & de la Torre, J. (2020a). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology*, *73*, 142–163. DOI: 10.1111/bmsp.12156

Ma, W., & de la Torre, J. (2020b). *GDINA: The generalized DINA model framework. R Package version 2.7.8*. https://cran.r-project.org/package=GDINA

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Economics* (pp. 105–142). Academic Press.

McFadden, D. (1979). Quantitative methods for analysing travel behavior of individuals: Some recent developments. In D. A. Hensher & P. R. Stopher (Eds.), *Behavioural travel modelling* (pp. 279–318). Croom Helm.

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*, 727–753. DOI: 10.1177/0013164418822700

Nájera, P., Sorrel, M. A., de la Torre, J., & Abad, F. J. (2020). Improving robustness in Q-matrix validation using an iterative and dynamic procedure. *Applied Psychological Measurement*, *44*, 431–446. DOI: 10.1177/0146621620909904

OECD (2006). PISA released items: Reading. http://www.oecd.org/pisa/38709396.pdf

R Core Team (2019). R (Version 3.6) [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.

Robitzsch A., & George A. C. (2019). The R Package CDM for Diagnostic Modeling. In: von Davier M., Lee Y.-S. (Eds.). *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages*. Springer. DOI: 10.1007/978-3-030-05584-4_26

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. DOI: 10.1177/0013164407301545

Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, *16*, 1–17. DOI: 10.1080/15366367.2018.1435104

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. DOI: 10.1177/1094428116630065

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

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

Wang, D., Cai, Y., & Tu, D. (2020). Q-matrix estimation methods for cognitive diagnosis models: Based on partial known Q-matrix. *Multivariate Behavioral Research*. DOI: 10.1080/00273171.2020.1746901

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag. Retrieved from https://ggplot2.tidyverse.org

Xu, G., & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, *113*, 1284–1295. DOI: 10.1080/01621459.2017.1340889

Yu, X., & Cheng, Y. (2019). Data-driven Q-matrix validation using a residual-based statistic in cognitive diagnosis assessment. *British Journal of Mathematical and Statistical Psychology*. DOI: 10.1111/bmsp.12191

Table 1. Latent classes and q-vectors for $K = 3$

Latent class	q-vector	Pattern
$\boldsymbol{\alpha}_1$		{0,0,0}
$\boldsymbol{\alpha}_2$	$\boldsymbol{q}^{(1)}$	{1,0,0}
$\boldsymbol{\alpha}_3$	$\boldsymbol{q}^{(2)}$	{0,1,0}
$\boldsymbol{\alpha}_4$	$\boldsymbol{q}^{(3)}$	{0,0,1}
$\boldsymbol{\alpha}_5$	$\boldsymbol{q}^{(4)}$	{1,1,0}
$\boldsymbol{\alpha}_6$	$\boldsymbol{q}^{(5)}$	{1,0,1}
$\boldsymbol{\alpha}_7$	$\boldsymbol{q}^{(6)}$	{0,1,1}
$\boldsymbol{\alpha}_8$	$\boldsymbol{q}^{(7)}$	{1,1,1}

Table 2. Summary of the factors in the simulation study

Factors	Factor levels
Q-matrix misspecification rate (*QM*)	0, .15, .30
Number of attributes (*K*)	4, 8
Average item quality (*IQ*)	.4, .6, .8
Sample size (*N*)	500, 1000
Ratio of number of items to attribute (*JK*)	4, 8
Attribute distribution (*AD*)	Uniform, Higher-order
Q-matrix validation method	GDI, Wald, Hull$_P$, Hull$_R$

Table 3. Overall results

	QRR	*TPR*	*TNR*	*OS*	*US*	*PCA*	*PCV*	*CT*	*CR*
GDI	.711	.800	.400	127.0	**4.1**	.813	.334	955	.389
Wald	.928	.955	.766	**3.9**	14.0	.888	.551	254	.990
Hull$_P$	**.953**	**.965**	**.800**	9.6	8.9	**.895**	**.591**	**66**	.994
Hull$_R$	.945	.963	.789	10.4	9.1	.894	.584	80	**.996**

Note. Best result by dependent variable is highlighted in bold. *QRR* = Q-matrix recovery rate; *TPR* = true positive rate; *TNR* = true negative rate; *OS/US* = number of over- and under-specifications, respectively; *PCA/PCV* = proportion of correctly classified attributes and vectors, respectively; *CT* = computation time (in seconds); *CR* = convergence rate.

Table 4. Results across the different factor levels

		QM			*K*		*IQ*			*N*		*JK*		*AD*	
DV	*Method*	0	.15	.30	4	8	.4	.6	.8	500	1000	4	8	Unif	H-O
QRR	GDI	.898	.766	.641	.922	.328	.688	.750	.703	.625	.750	.719	.661	.688	.719
	Wald	.971	.922	.863	.922	.938	**.871**	.930	**.969**	.914	.945	.922	.938	.938	.922
	$Hull_P$	**.973**	**.938**	**.879**	**.945**	**.959**	.859	**.953**	**.969**	**.938**	**.961**	**.938**	**.955**	**.961**	**.938**
	$Hull_R$	.969	.930	.867	.938	.957	.852	.945	**.969**	.936	.959	.934	.953	.955	**.938**
TPR	GDI	.898	.833	.672	.944	.335	.811	.822	.694	.644	.844	.822	.688	.756	.822
	Wald	.971	.959	.933	.945	.963	.926	.961	**.977**	.938	.969	.961	.949	.963	.950
	$Hull_P$	**.973**	**.963**	**.955**	**.961**	**.969**	**.953**	**.967**	.969	**.954**	**.975**	**.968**	**.963**	**.971**	**.960**
	$Hull_R$	.969	**.963**	.944	.954	.968	.940	.963	.970	.950	.972	.963	.961	.969	.954
TNR	GDI	–	.400	.383	.757	.289	.289	.390	.558	.370	.421	.364	.461	.403	.389
	Wald	–	.800	.737	.684	**.816**	**.526**	.786	.909	.779	.763	**.658**	.844	.792	.737
	$Hull_P$	–	**.816**	**.779**	**.789**	**.816**	.429	**.842**	**.948**	**.789**	**.811**	**.658**	**.909**	**.842**	**.763**
	$Hull_R$	–	.800	.740	**.789**	.805	.421	.800	**.948**	.766	.792	.610	.895	.817	.737
OS	GDI	122.1	127.7	131.1	**1.8**	252.1	135.1	132.9	112.9	131.8	122.2	78.4	175.5	127.2	126.7
	Wald	**0.3**	**3.7**	**7.8**	2.1	**5.8**	**7.0**	**3.0**	1.8	**3.3**	**4.5**	**4.4**	**3.4**	**3.6**	**4.2**
	$Hull_P$	3.1	9.3	16.5	2.8	16.4	22.8	5.1	**0.9**	12.0	7.3	9.7	9.6	9.1	10.1
	$Hull_R$	3.7	10.0	17.6	3.3	17.6	24.3	5.9	1.0	13.0	7.9	10.1	10.7	9.9	11.0
US	GDI	**2.7**	**4.0**	**5.6**	7.8	**0.4**	**5.4**	**4.1**	**2.8**	**4.2**	**4.0**	**3.0**	**5.2**	**3.3**	**4.9**
	Wald	7.0	14.1	21.1	7.6	20.6	20.4	13.0	8.7	16.9	11.1	9.2	18.9	12.8	15.3
	$Hull_P$	6.0	8.7	12.0	**5.3**	12.6	10.4	8.6	7.7	9.9	7.9	6.1	11.7	7.6	10.2
	$Hull_R$	6.1	9.0	12.3	5.5	12.7	10.8	9.3	7.3	10.0	8.2	6.3	12.0	7.6	10.7

Table 4. Results across the different factor levels (*Cont.*)

		QM			*K*		*IQ*			*N*		*JK*		*AD*	
DV	*Method*	0	.15	.30	4	8	.4	.6	.8	500	1000	4	8	Unif	H-O
PCA	GDI	.840	.820	.781	.890	.686	.670	.800	.944	.792	.834	.776	.841	.806	.820
	Wald	**.920**	.900	.855	.888	.888	**.756**	.896	.976	.884	.891	.856	.942	.874	.896
	$Hull_P$	**.920**	**.903**	**.869**	**.895**	**.894**	.755	**.903**	**.980**	**.891**	**.898**	**.857**	**.946**	**.882**	**.901**
	$Hull_R$	**.920**	.901	.866	.893	**.894**	.752	.901	**.980**	.889	.896	.856	**.946**	.881	.899
PCV	GDI	.391	.346	.281	.632	.054	.138	.330	.664	.286	.382	.268	.428	.318	.343
	Wald	**.638**	.584	.432	.626	.384	**.227**	.576	.892	.536	.567	.388	.674	.520	.582
	$Hull_P$	.636	**.608**	**.490**	**.648**	**.411**	.206	**.608**	**.910**	**.576**	**.605**	**.396**	**.705**	**.568**	**.624**
	$Hull_R$	.633	.600	.478	.641	.402	.204	.602	**.910**	.569	.597	.391	.698	.561	.612
CT	GDI	889	918	1059	3	1908	1140	934	792	1052	859	280	1631	927	984
	Wald	**34**	125	613	**2**	512	238	242	282	202	306	**53**	460	273	235
	$Hull_P$	**34**	**61**	**102**	**2**	**130**	**133**	**47**	**18**	**54**	**78**	58	**74**	**64**	**68**
	$Hull_R$	44	79	117	**2**	159	165	54	22	63	98	67	94	78	83
CR	GDI	.485	.391	.292	.762	.017	.335	.396	.438	.372	.407	.410	.369	.390	.389
	Wald	**1.000**	**.998**	.972	**1.000**	.980	.986	.986	**.997**	.980	**.999**	**1.000**	.980	.990	.989
	$Hull_P$	.996	.994	.991	.998	.990	.998	.994	.989	.991	.997	.993	.994	.995	.992
	$Hull_R$	.998	.997	**.994**	.996	**.997**	**.999**	**.999**	.991	**.996**	.997	.997	**.996**	**.997**	**.996**

Note. Best result by dependent variable and factor level is highlighted in bold. *DV* = dependent variable; *QRR* = Q-matrix recovery rate; *TPR* = true positive rate; *TNR* = true negative rate; *OS/US* = number of over- and under-specifications, respectively; *PCA/PCV* = proportion of correctly classified attributes and vectors, respectively; *CT* = computation time (in seconds); *CR* = convergence rate; *QM* = Q-matrix misspecification rate; *K* = number of attributes; *IQ* = average item quality; *N* = sample size; *JK* = ratio of number of items to attribute; *AD* = attribute distribution; *Unif* = uniform distribution; *H-O* = higher-order distribution.

Table 5. Original Q-matrix (Chen & de la Torre, 2014)

		Attributes					
Item	Item ID	α_1	α_2	α_3	α_4	α_5	α_6
1	R040Q02	0	0	1	1	0	0
2	R040Q03A†	<u>1</u>	0	0	1	0	0
3	R040Q03B	0	0	0	1	1	0
4	R040Q04	0	1	0	1	0	0
5	R040Q06	0	0	1	1	0	0
6	R077Q02	1	0	0	0	0	0
7	R077Q03	0	1	0	0	1	0
8	R077Q04	0	0	1	0	1	0
9	R077Q05	0	1	0	0	1	0
10	R077Q06	1	0	1	0	0	0
11	R088Q01	0	1	0	1	0	0
12	R088Q03†	1	0	<u>0</u>	1	0	0
13	R088Q04T	0	0	1	1	0	0
14	R088Q05T	0	1	0	1	0	0
15	R088Q07	0	1	0	1	1	0
16	R110Q01	0	1	0	0	1	0
17	R110Q04	1	0	1	0	0	0
18	R110Q05	1	0	1	0	0	0
19	R110Q06	1	0	1	0	0	0
20	R216Q01	0	1	0	0	0	1
21	R216Q02	0	0	1	0	1	1
22	R216Q03T†	1	0	1	0	<u>0</u>	1
23	R216Q04	0	0	1	0	0	1
24	R216Q06†	<u>1</u>	0	1	0	0	1
25	R236Q01†	1	0	1	0	0	<u>1</u>
26	R236Q02	0	0	1	0	0	1

Note. † = item with modified q-vector; underlined q-entries = entries to be modified based on the Hull$_P$ method.

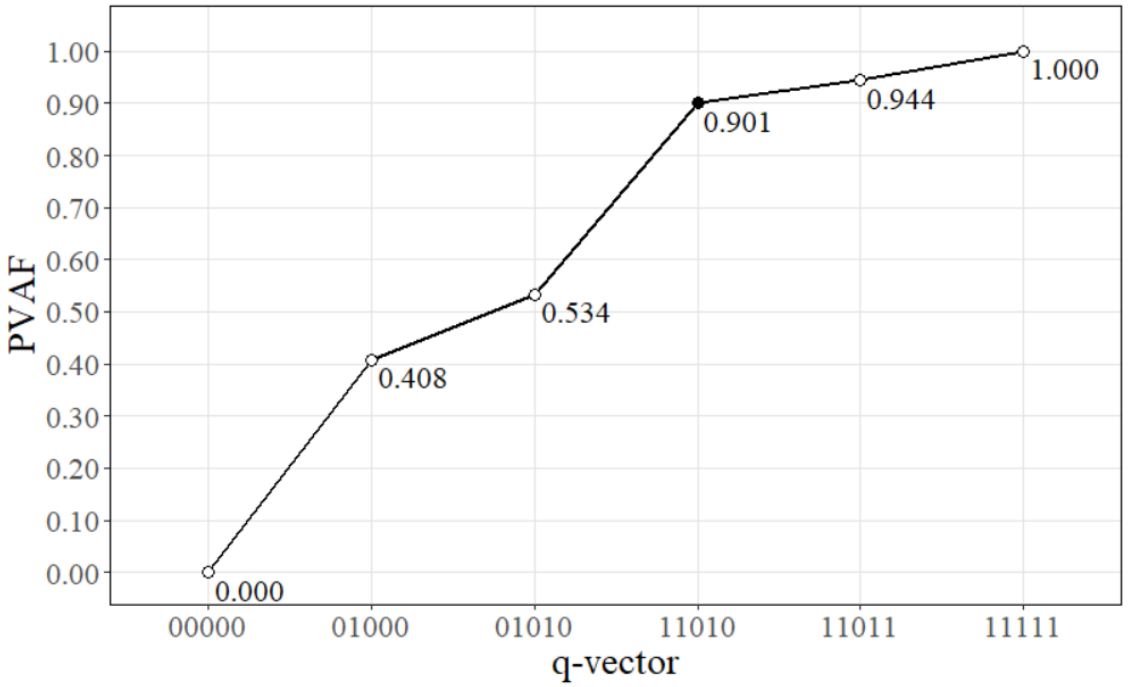

Figure 1. Illustration of a mesaplot for item j. $\boldsymbol{q}_j = \{11010\}$ is likely to be the most appropriate q-vector.

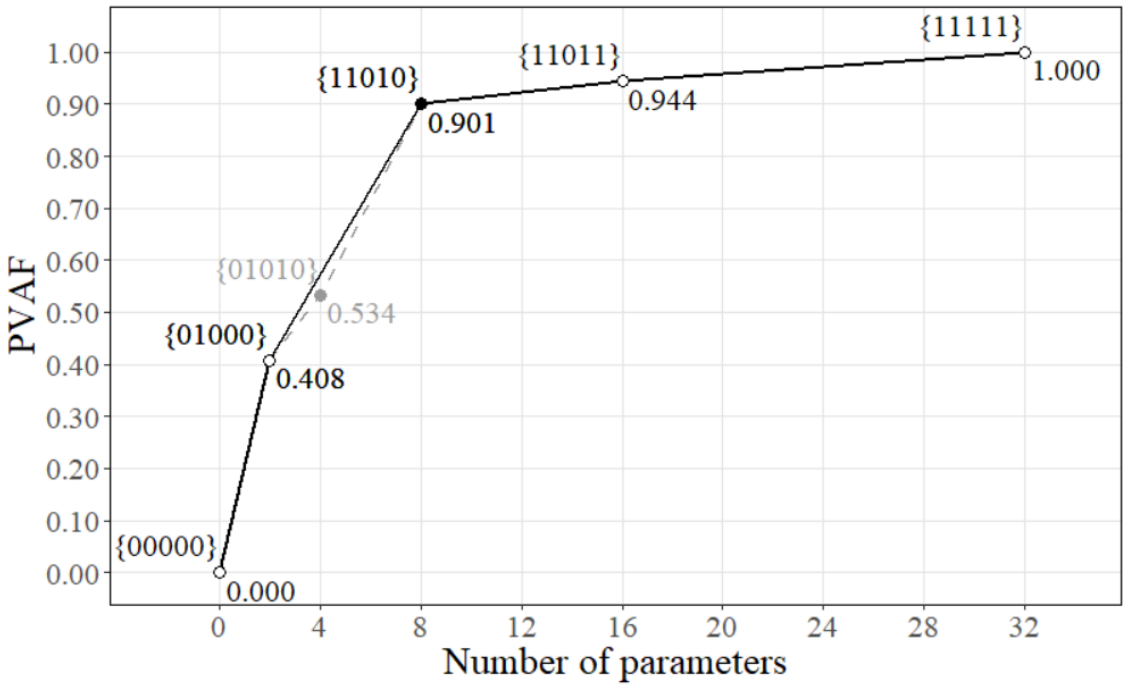

Figure 2. Illustration of a hull plot for item j with the PVAF on the *y-axis*. The convex hull is represented with the black line. The grey line shows the non-candidate q-vector. Q-vector specifications are shown in curly brackets. In this example, $st_{j1} = 2.48$, $st_{j3} = 15.34$, and $st_{j4} = 1.53$ for $\boldsymbol{q}_j = \{01000\}$, $\{11010\}$, and $\{11011\}$, respectively. The suggested q-vector, $\boldsymbol{q}_j = \{11010\}$, is represented by a black point.

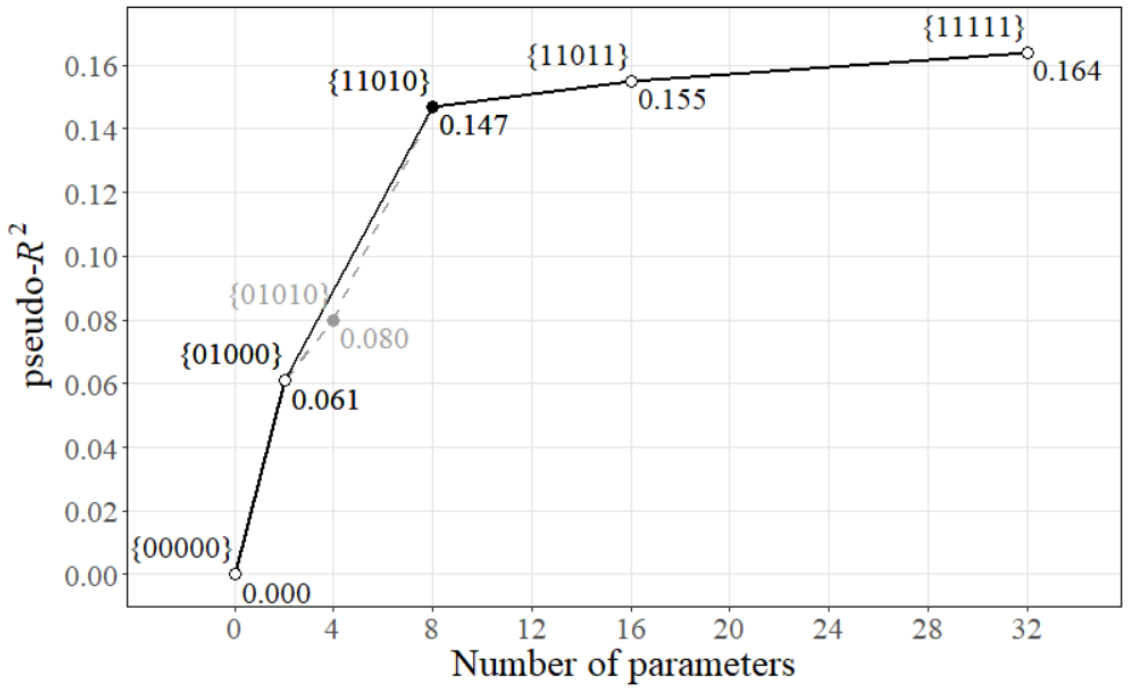

Figure 3. Illustration of a hull plot for item *j* with the pseudo-R^2 on the *y-axis*. The convex hull is represented with the black line. The grey line shows the non-candidate q-vector. Q-vector specifications are shown in curly brackets. In this example, $st_{j1} = 2.11$, $st_{j3} = 14.76$, and $st_{j4} = 1.75$ for $\boldsymbol{q}_j = \{01000\}$, $\{11010\}$, and $\{11011\}$, respectively. The suggested q-vector, $\boldsymbol{q}_j = \{11010\}$, is represented by a black point.

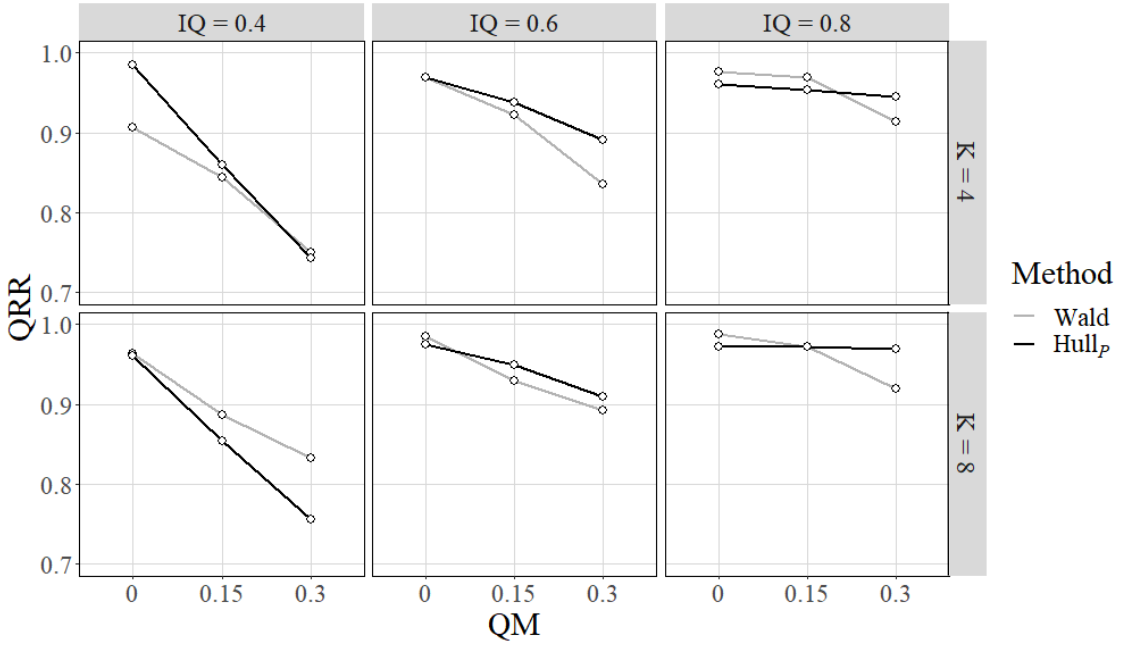

Figure 4. *QRR* medians for the interaction *Method* (Wald and Hull$_P$) × *IQ* × *K* × *QM*. *QRR* = Q-matrix recovery rate; *IQ* = average item quality; *K* = number of attributes; *QM* = Q-matrix misspecification rate.

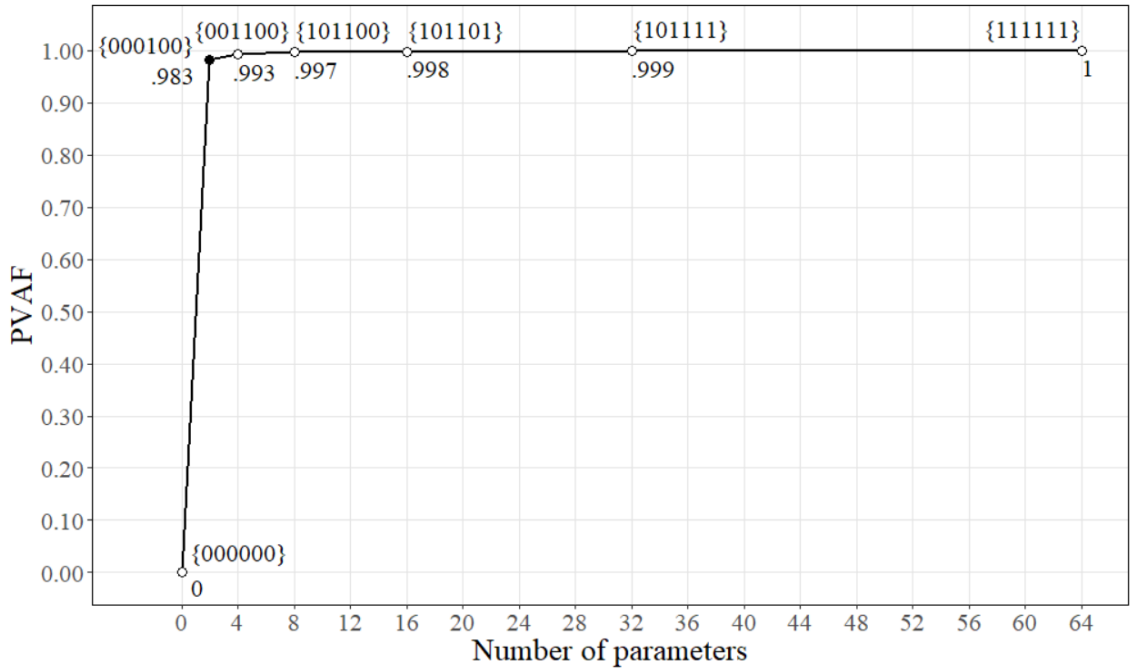

Figure 5. Hull plot for item R040Q03A. Q-vector specifications are shown in curly brackets. For this item, $st_{j1} = 94.5$, $st_{j2} = 4.8$, $st_{j3} = 6.6$, $st_{j4} = 2.4$, and $st_{j5} = 7.3$, for $\boldsymbol{q}_j = \{000100\}, \{001100\}, \{101100\}, \{101101\}$, and $\{101111\}$, respectively. The suggested q-vector, $\boldsymbol{q}_j = \{000100\}$, is represented by a black point.

View publication stats