Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

Nonparametric Calibration of Item-by-Attribute Matrix in Cognitive Diagnosis

Youn Seon Lim & Fritz Drasgow

To cite this article: Youn Seon Lim & Fritz Drasgow (2017) Nonparametric Calibration of Item-by-Attribute Matrix in Cognitive Diagnosis, Multivariate Behavioral Research, 52:5, 562-575, DOI: 10.1080/00273171.2017.1341829

To link to this article: https://doi.org/10.1080/00273171.2017.1341829

Published online: 17 Jul 2017.

MULTIVARIATE BEHAVIORAL RESEARCH
2017, VOL. 52, NO. 5, 562–575
https://doi.org/10.1080/00273171.2017.1341829

Nonparametric Calibration of Item-by-Attribute Matrix in Cognitive Diagnosis

Youn Seon Lim [a] and Fritz Drasgow [b]

[a] Hofstra University; [b] University of Illinois Urbana Champaign

ABSTRACT
A nonparametric technique based on the Hamming distance is proposed in this research by recognizing that once the attribute vector is known, or correctly estimated with high probability, one can determine the item-by-attribute vectors for new items undergoing calibration. We consider the setting where Q is known for a large item bank, and the q-vectors of additional items are estimated. The method is studied in simulation under a wide variety of conditions, and is illustrated with the Tatsuoka fraction subtraction data. A consistency theorem is developed giving conditions under which nonparametric **Q** calibration can be expected to work.

KEYWORDS
Cognitive diagnosis; nonparametric classification; online calibration

Introduction

Cognitive diagnosis has received much attention in recent years because it aims to provide more diagnostic information about each examinee's abilities than a single score can, which could lead to finer classification and more efficient remediation. Many models for cognitive diagnosis have been introduced, but they all tend to share a common feature, a matrix that details which attributes are required for each item, usually referred to as the Q-matrix (Tatsuoka, 1983). The Q-matrix is a binary $J \times K$ matrix, in which rows represent the J items, and the columns refer to the K attributes. Each entry q_{jk} in the matrix indicates whether the kth attribute is involved in the solution of the jth item. Correctly specifying Q is perhaps the most crucial part of cognitive diagnosis, because its misspecification leads to inaccurate estimation of an examinee's knowledge state (e.g., Im & Corter, 2011; Rupp & Templin, 2008).

Typically, expert knowledge is used to construct Q-matrices. However, it has been shown that Q-matrices constructed by content experts do not always precisely reflect the pattern of examinee thought (Hubal, 1992). Another shortcoming is that it is unlikely that all content experts would arrive at the same set of attributes, in spite of following the same procedures. Another complication is that problem-solving strategies of examinees and content experts might differ substantially (Sweller, 1988). In spite of the known weaknesses, statistical estimation of Q has not been widely explored because of inherent unidentifiability and computational burdens. As Liu et al. (2011) indicated, the discrete entries of Q and the nonlinearity of parametric models make estimation complicated.

Studies have been conducted to validate and estimate Q under some regularity conditions. DeCarlo (2012) proposed a Bayesian approach in which some elements of Q elements are unknown, but others are known. In this work, the posterior distributions of hyperparameters modeling the probability of a 1 for matrix entries were used to estimate the elements of Q. Some researchers have tried to obtain Q from a small set of candidates by fitting a model with each candidate Q and then comparing likelihood-based fit statistics (e.g., Rupp & Templin, 2008). When the models are fitted, the indices of relative goodness of fit such as the Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC) are used. Often likelihood ratio statistics are used when nested Q-matrices are compared.

The difficulties of this approach include determining the correct number of attributes and identifying a reasonable subset of the many possible Q-matrices. De la Torre (2008) proposed an empirically based method of validating Q in the DINA model (e.g., Junker & Sijtsma, 2001) by maximizing the difference in success probabilities between a group of examinees possessing all required attributes and a group not possessing any of them when the item parameters are known. Huo and de la Torre (2013) extended the approach to estimate Q in the setting in which examinees' latent attribute vectors are known. They obtained correct q-vectors by comparing the weighted variance ς^2 of the success probabilities of

CONTACT Youn Seon Lim younseon.lim@gmail.com School of Medicine, Hofstra University, Hempstead, NY 11549, USA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hmbr.

different latent classes (de la Torre & Chiu, 2010, 2015), which was maximized for differences between success probabilities for the two groups with the correct q-vector.

The hill-climbing algorithm has been used to estimate Q from item response data (e.g., Brewer, 1996; Barnes, 2003). This is an iterative algorithm that refines the item by attribute relationships by adjusting randomly assigned q-vector until the total error associated with clustering examinees by the q-vector is minimized. More specifically, it starts by setting the number of attributes K to 1 and generating a random Q. With the Q, each examinee's latent attribute vector α is obtained from 2^K possible attribute vectors. Then ideal item response for each examinee i and each item j is estimated by using the estimated α and the random $\mathbf{q}_j$. If an α includes all required attributes for an item $\mathbf{q}_j$, the ideal response is set to 1, and otherwise to 0. The observed item responses are compared with the ideal responses. Their Hamming distance (HD) is simply the number of differences. The total error is obtained by summing the Hamming distances over all items and all examinees. The algorithm is repeated by adding more attributes K, and the Q with minimum error is saved until a stopping criterion is satisfied: the overall Q error rate does not fall below a preset threshold such as an error rate less than 1 per examinee, or an increase in K does not lead to a significant decrease in the error. The primary weakness of this approach is that it is purely exploratory and cannot complete an exhaustive search through all possible Q-matrices, making its convergence properties questionable.

Chiu and Douglas (2013) utilized the Hamming distance technique to estimate α for a known Q. Like Barnes (2003), their model posits a conjunctive relationship among the attributes, but the ideal response for an examinee i to an item j is defined as $\eta_{ij} = \prod_{k=1}^{K} {\alpha_{ik}}^{q_{jk}}$. Then like Barnes (2003), 2^K ideal responses are assembled for each examinee by using the known Q. The estimate $\hat{\alpha}_i$ is obtained by minimizing the Hamming distance between ideal responses and observed item responses over each examinee. Wang and Douglas (2013) showed the consistency of this approach for underlying true conjunctive models. Their theoretical justification is that the true attribute vector α_i minimizes the expected distance between item response pattern and ideal response pattern, under some general conditions for the underlying model.

Chiu (2013) proposed a Q refinement method to modify inaccurately specified q-entries of a Q-matrix by using an extension of the method of Chiu and Douglas (2013). This method starts by identifying the item having the highest Residual Sum of Squares (RSS) between the observed responses and the ideal responses across all examinees. Then this method replaces that q-vector with a different q-vector minimizing the RSS. Examinees' class memberships are re-estimated using the new Q-matrix before moving to the next item. The primary limitation of this method is that the ideal responses rely solely on an unrealistic noncompensatory attribute relationship assumption. Her published R program (NPCD, 2015) for this algorithm requires users to specify either a conjunctive or disjunctive relationship. In practice, it cannot be specified a priori. Furthermore, the RSS can be inflated by incorrect estimates of examinees' class membership due to incorrect q-vectors for different items and the compensatory attribute relationships.

A statistically consistent but computationally burdensome solution to this problem was suggested by Liu et al. (2011, 2012) in the case of a fixed K. Unlike hill-climbing algorithms, this technique offers a global solution to a loss function, which is critical for asymptotic theory. A matrix T is introduced which essentially describes expected scores that would result, given population proportions of attribute patterns and item parameters. The matrix $T(Q)$ has $2^K - 1$ columns, each of which represents one of the possible q-vector patterns excluding the all zero pattern. Each row of $T(Q)$ corresponds to the number of examinees who correctly answered each of all possible $2^J - 1$ combinations of J items. Then let $\hat{\boldsymbol{p}}_c$, $c = 1, 2, \ldots, 2^K$ equal the proportion of examinees possessing the cth attribute patterns, and obtain $T(Q)\hat{\boldsymbol{p}}$ which is expected to be equal to the corresponding β. The β is a vector containing probabilities of observed empirical distribution - the proportions of examinees answering the combinations of J items correctly given α. Then Q is taken by minimizing the differences between $T(Q)\ \hat{\boldsymbol{p}}$ and β over all competing matrices and vectors. The established theoretical framework is implemented primarily with the DINA model, and the item parameters are considered together with the estimation of Q. This approach has some limitations in practice because it involves an exhaustive search that cannot be conducted for medium and large problems. Liu et al. (2012) provided some algorithms for approximating the optimal solution.

The primary purpose of this study was to estimate Q for new items that are undergoing calibration. We propose a nonparametric technique based on the Hamming distance (e.g., Barnes, 2003; Chiu & Douglas, 2013) by recognizing that correctly estimated attribute vectors enable estimation of q-vectors for new items that are being introduced without a known or expertly chosen q-vector. The approach we consider may be implemented with any method that yields consistent attribute classifications, whether doing parametric or nonparametric modeling, although we stick with the nonparametric method. We begin with the brief overview of cognitive diagnosis models. Following this, the presentation of the estimation procedure and then the theoretical foundation for the proposed method will be given. Next, a simulation study is provided and an analysis of real data is given. Finally, a nonparametric algorithm to determine K is proposed before concluding with a discussion of the results and remarks on some promising research directions.

Cognitive diagnosis models

Let Y_{ij} denote the binary item response of the ith examinee to the jth item, $i = 1, \ldots, I,\ j = 1, \ldots, J$, with 1 = correct and 0 = incorrect. Cognitive diagnosis models are completely defined by formulating the conditional distribution of item responses Y_{ij} given latent attributes $\alpha_i = \{\alpha_{ik}\}$, for $k = 1, \ldots, K$, and parameterizing the joint distribution of α_i. Each entry α_{ik} indicates whether the ith examinee has mastered the kth attribute (i.e., the kth knowledge or skill), with $\alpha_{ik} = 1$ indicates examinee i has mastered attribute k and 0 otherwise. A key element in all cognitive diagnosis models is the binary $J \times K$ Q-matrix. The Q-matrix has a row for each item, $j = 1, \ldots, J$, and a column for each attribute, $k = 1, \ldots, K$. Each entry q_{jk} in the matrix indicates whether the kth attribute is relevant for the solution of the jth item: $q_{jk} = 1$ if the attribute is germane, 0 is not.

Noncompensatory CDMs

Noncompensatory models include conjunctive and disjunctive models. Henson, Templin and Willse (2009) discussed their differences thoroughly. Conjunctive models refer to the models for which all nonzero elements of an attribute vector are required in the item solving process, because the lack of any needed attribute cannot be compensated by the mastery of other attributes. Disjunctive models require at least one non-zero element of an attribute vector to solve an item. A common conjunctive model is the deterministic input, noisy and gate (DINA) model (e.g., Junker & Sijtsma, 2001). In this model, an ideal response η_{ij} is used to indicate whether all needed attributes for the jth item are possessed by the ith examinee. The ideal response can be written as $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$, where 0^0 in taken as 1.

This characterizes the conjunctive feature of the DINA model because when a required attribute ($q_{ik} = 1$) is not mastered ($\alpha_{ik} = 0$), we have a $0^1 = 0$ term in the product, and hence $\eta_{ij} = 0$. The item response function (IRF) for the DINA model is

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i, s_j, g_j) = (1 - s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})} \tag{1}$$

where $s_j = P(Y_j = 0 \mid \eta_j = 1)$ and $g_j = P(Y_j = 1 \mid \eta_j = 0)$. In this equation, the guessing parameter g_j functions as a lower asymptote, akin to the c_j parameter of the three-parameter logistic model. The slipping parameter, s_j is used to create an upper asymptote: if a test taker possesses all the attribute required by an item, then $\eta_{ij} = 1$ and then $P(Y_{ij} = 1|\alpha_i, s_j, g_j) = 1 - s_{ij}$.

Another example of conjunctive model is a reduced version of the reparameterized unified model (R-RUM) (R-RUM, Roussos, DiBello, Stout, Hartz, Henson, & Templin, 2007; Hartz, & Roussos, 2008). The IRF in this model is

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i) = \pi_j \prod_{k=1}^{K} b_{jk}^{q_{jk}(1-\alpha_{ik})} \tag{2}$$

where π_j denotes the probability of a correct response to the jth item for the ith examinee who possesses and correctly applies all attributes required for that item, and b_{jk} indicates the penalty to the probability of correct response to item j for not having mastered the kth attributes. More specially, when either $q_{jk} = 0$ or $\alpha_{ik} = 1$, $(1 - \alpha_{ik})$ $q_{jk} = 0$. Thus, no penalty is applied for the attribute because $r_{jk}^0 = 1$.

The Deterministic Input, Noisy Output Or gate (DINO) model is an example of a disjunctive model (Templin & Henson, 2006). This model shares similarities with the DINA model although they are conceptually different. Unlike the DINA model, the DINO model assumes that a test taker can answer item j correct if he/she possesses one attribute relevant to that item. An algebraically compact expression for this condition is given by $w_{ij} = 1 - \prod_{k=1}^{K}(1 - \alpha_{ik})^{q_{jk}}$. For example, if the first two attributes are relevant ($q_{j1} = q_{j2} = 1$) and an examinee has mastery of the first, but not the second, $w_{ij} = 1 - [(1-1)^1(1-0)^1] = 1 - 0 = 1$. The model has the following IRF:

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = (1 - s_j)^{w_{ij}} g_j^{(1-w_{ij})} \tag{3}$$

where s_j and g_j are again the slipping and guessing parameters, respectively.

Compensatory CDMs

In compensatory models, the absence of any element of an attribute vector can be compensated by the remaining elements. Many of these models contain linear combinations of elements from a latent attribute vector rather than products of them, and the linear combinations are transformed via a link function to yield the probability of a correct item response (e.g., Rupp, Templin, & Henson, 2010).

von Davier's general diagnostic model (GDM) (von Davier, 2005, 2008) is an example of a compensatory model. By defining the link function in different ways, the GDM easily generalizes to many latent variable models for both categorical and continuous latent variables. The special case of the GDM for Compensatory Reparameterized Unified Model (C-RUM, e.g., Templin, 2006) can be represented as the following IRF:

$$P(Y_{ij} = 1 \mid \alpha_i) = \frac{\exp[\lambda_{j0} + \sum_{k=1}^{K} \lambda_{jk}\alpha_{ik}q_{jk}]}{1 + \exp[\lambda_{j0} + \sum_{k=1}^{K} \lambda_{jk}\alpha_{ik}q_{jk}]} \tag{4}$$

where the additional parameters λ_{j0} and λ_{jk}, respectively, represent jth item difficulty and K-dimensional slope.

The above are some popular cognitive diagnosis models. In the simulation study below, DINA, R-RUM, DINO, and C-RUM are used for data simulation.

Calibration of the $\mathbf{Q}_{\text{new}}$ matrix

Throughout this article, $\mathbf{Q}$ is used to denote a known $J \times K$ matrix, and $\mathbf{Q}_{\text{new}}$ denotes an unknown $J_{\text{new}} \times K$ matrix, in which J_{new} represents the number of new items being calibrated. Let $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_I$ be estimated latent attribute vectors of I examinees and $\hat{\alpha}_i = (\hat{\alpha}_{i1}, \hat{\alpha}_{i2}, \ldots, \hat{\alpha}_{iK})$, where each $\hat{\alpha}_{ik}$ indicates whether examinee i has mastered attribute k. The proposed approach is built on the information contained in these I examinees' estimated latent attribute patterns, which may be obtained through either nonparametric classification or parametric modeling of the J assessment items. Let $\boldsymbol{y}_{Ij_{\text{new}}}$ be a response vector of I examinees to the j_{new} calibration item. In the following parts, $\boldsymbol{y}_{Ij_{\text{new}}}$ is denoted as $\boldsymbol{y}_{j_{\text{new}}}$ for convenience. The item response vector of the ith examinee is denoted as $\boldsymbol{y}_{iJ_{\text{new}}} = (y_{i1_{\text{new}}}, y_{i2_{\text{new}}}, \ldots, y_{iJ_{\text{new}}})$. Let $\mathbf{Q}_{\text{new}}$ indicate the relationship between the J_{new} items and the K attributes as does $\mathbf{Q}$. We assume that $\mathbf{Q}_{\text{new}}$ is unknown, and the purpose is to estimate the individual rows of $\mathbf{Q}_{\text{new}}$.

$\mathbf{Q}_{new}$ *estimator*

Let $\mathbf{q}_{j_{\text{new}}}$ denote each row vector of $\mathbf{Q}_{\text{new}}$, and $\mathbf{q}_{j_{\text{new}}} = (q_{j_{\text{new}}1}, q_{j_{\text{new}}2}, \ldots, q_{j_{\text{new}}K})$, where each entry $q_{j_{\text{new}}k} \in \{0, 1\}$, according to whether the kth attribute is required for a correct response. Each true $\mathbf{q}_{j_{\text{new}}}$ is one of the $2^K - 1$ possible q-patterns, denoted as $\mathbf{q}_{j_{\text{new}}c}$, for $c = 1, \ldots, 2^K - 1$. It is assumed that all J_{new} calibration items require at least one attribute, and the vector of all zeroes is excluded.

Let $\eta_{ij_{\text{new}}c}$ denote one of ideal responses $\eta_{ij_{\text{new}}c1}, \eta_{ij_{\text{new}}c2}$, and $\eta_{ij_{\text{new}}c3}$ given $\mathbf{q}_{j_{\text{new}}c}$. Let $\eta_{ij_{\text{new}}c_1}$ denote the noncompensatory situation in which all of the necessary q-attributes for a calibration item j_{new} are required to answer the item correctly, and

$$\eta_{ij_{\text{new}}c_1} = \prod_{k=1}^{K} \hat{\alpha}_{ik}^{q_{j\text{new}ck}} \tag{5}$$

Let $\eta_{ij_{\text{new}}c_2}$ denote the disjunctive attribute relationship where at least one attribute is required to be mastered among the necessary attributes, here one mastered attribute can compensate for any unmastered attributes

$$\eta_{ij_{\text{new}}c_2} = 1 - \prod_{k=1}^{K} (1 - \hat{\alpha}_{ik})^{q_{j\text{new}ck}} \tag{6}$$

Let $\eta_{ij_{\text{new}}c_3}$ denote the compensatory case where a low value on one attribute can be compensated for by a high value on another latent variables

$$\eta_{ij_{\text{new}}c_3} = \text{round}\left(\sum_{k=1}^{K} (\hat{\alpha}_{ik} \times q_{j_{\text{new}}ck})/K\right) \tag{7}$$

Then $\eta_{Ij_{\text{new}c}}$ denotes a matrix with rows consisting of ideal response vectors of I examinees to the j_{new} calibration item for $\mathbf{q}_{j_{\text{new}}c}$. In the following parts, $\eta_{Ij_{\text{new}c}}$ is denoted as $\eta_{j_{\text{new}c}}$ for convenience.

Like previous studies (e.g., Barnes, 2003; Chiu & Douglas, 2013), the estimator of $\mathbf{Q}_{\text{new}}$ is based on the minimum Hamming distance between observed item responses and their ideal responses across all possible q-patterns $\boldsymbol{q}_{j_{\text{new}}c}$, $c = 1, 2, \ldots, 2^K - 1$. The distance $D(\boldsymbol{y}_{j_{\text{new}}}, \boldsymbol{q}_{j_{\text{new}}c})$ is measured by counting the number of times the observed item responses $\mathbf{y}_{j_{\text{new}}}$ match the ideal responses $\eta_{j_{\text{new}c}}$. However, unlike the previous studies, $\eta_{j_{\text{new}}c} = \arg\min_{t \in \{1,2,3\}} D(\boldsymbol{y}_{j_{\text{new}}}, \eta_{j_{\text{new}}c_t})$.

Then for $\boldsymbol{q}_{j_{\text{new}}c}$

$$D(\boldsymbol{y}_{j_{\text{new}}}, \boldsymbol{q}_{j_{\text{new}}c}) = \sum_{i=1}^{I} \mid y_{ij_{\text{new}}} - \eta_{ij_{\text{new}}c} \mid = \sum_{i=1}^{I} d(\boldsymbol{q}_{j_{\text{new}}c}) \tag{8}$$

Formally, minimizing the distance over all possible values of $\boldsymbol{q}_{j_{\text{new}}}$ produces the estimator for this study

$$\hat{\boldsymbol{q}}_{j_{\text{new}}} = \arg \min_{c \in \{1,2,\ldots,2^K-1\}} D(\boldsymbol{y}_{j_{\text{new}}}, \boldsymbol{q}_{j_{\text{new}}c}) \tag{9}$$

This $\mathbf{Q}_{\text{new}}$ classifier can result in ties, especially when the sample size I is small. In practice, ties may be broken by implementing a random selection method. However, the probability of a tie converges to 0 as $I \to \infty$ based on the following discussion of asymptotic consistency of the estimator, under a general class of cognitive diagnosis models. The asymptotic behavior of the $\mathbf{Q}_{\text{new}}$ estimator is discussed in Appendix A.

An illustrative example

To help the understanding of this proposed approach, one simple example is provided. Suppose that two attributes $K = 2$ are tested. Then all examinees are classified into $2^K = 4$ classes associated with a particular attribute pattern α. Assume that examinees $I = 4$ and their α are

Table 1. An example of generating ideal item responses with $K = 2$.

	$\alpha = (0, 0), y_{j_{\text{new}}1} = 0$			$\alpha = (1, 0), y_{j_{\text{new}}1} = 0$			$\alpha = (0, 1), y_{j_{\text{new}}1} = 1$			$\alpha = (1, 1), y_{j_{\text{new}}1} = 1$			
q_c	Equation (1)	Equation (2)	Equation (3)	Equation (1)	Equation (2)	Equation (3)	Equation (1)	Equation (2)	Equation (3)	Equation (1)	Equation (2)	Equation (3)	HD
1, 0	0	0	0	1	1	1	0	0	0	1	1	1	2, 2, 2
0, 1	0	0	0	0	0	0	1	1	1	1	1	1	0, 0, 0
1, 1	0	0	0	0	1	1	0	1	1	1	1	1	2, 2 ,2

obtained through either nonparametric or parametric modeling of assessment items J. All items measure at least an attribute, possible q vectors are $2^K - 1 = 3$. Assume that the q-vector for an calibration item j_{new} needs to be estimated.

Let four examinees' item responses for the calibration item j_{new} be {0, 0, 1, 1}. Given each of the possible q vectors, examinees' ideal responses are obtained from Equations (5) to (7) as shown in Table 1. The HD between the ideal responses and observed item responses is counted over all examinees. Then the estimated $\hat{q}$ vector for the calibration item j_{new} is taken as q vector $= (0, 1)$ for which any of the three corresponding ideal response most closely matched the observed item response.

Simulation study

To investigate how accurately the true $\boldsymbol{q}_{j_{\text{new}}}$ can be estimated by the proposed method, several simulations were conducted by crossing the number of examinees, the length of the assessment test, the number of attributes K, the distribution of α, and the values of item parameters for four different cognitive diagnosis models.

Study design

For each condition, item response data were simulated for assessment items and calibration items. As the first step, we simulated the assessment item response data in which sample sizes I of 250, 1000, or 5000 were drawn from two discretized multivariate normal distributions $MVN(0_K, \Sigma)$, where the covariance matrix Σ has unit variance and common correlation $\rho = .3$, or .6. Like Chiu and Douglas (2013), the K-dimensional continuous vectors $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK})'$ were dichotomized by

$$\alpha_{ik} = \begin{cases} 1, & \text{if } \theta_{ik} \geq \Phi^{-1}\frac{k}{K+1} \\ 0, & \text{otherwise} \end{cases}$$

Assessment test lengths J of 20, 40, and 60 were studied with attribute vectors of length K= 3, 5, and 7. For the calibration data set, 1000 calibration items were investigated with the same length K as the assessment exam to obtain stable results. The q-vectors for both assessment and calibration items were generated randomly from a discrete uniform distribution on the maximum $2^K - 1$ possible q-vectors. For the study with misspecified $\mathbf{Q}$, 10% of q_{jk} entries for J items were randomly misspecified.

The data were generated from two commonly used noncompensatory models: the DINA model and the R-RUM, one disjunctive model and one compensatory model: the DINO model and C-RUM. For the DINA and DINO models, three different levels of guessing g_j and slipping s_j parameter values were sampled from uniform distributions with a left endpoint value of 0 and a right endpoint value of *Max*, where *Max* was .1, .3 or .5. For R-RUM and C-RUM, π_j and λ_{j0} were generated from the uniform distribution (.6, 1), and r^{*jk} and $\lambda_{jk}$from a uniform distribution(0, .4).

As indicated, this method begins by obtaining estimates of α from the assessment items to construct the $\mathbf{Q}_{\text{new}}$ for calibration items.

For the C-RUM model, an EM algorithm as described in de la Torre (2011) was used to estimate $\hat{\alpha}_i$. For the other models, the reverse of this proposed technique was employed to estimate examinees' attribute profile $\hat{\alpha}_i$. More specifically, given $\mathbf{Q}$, minimizing the distance over all possible values of α_m, $m = 1, \ldots, 2^K$ produces the estimator

$$\hat{\alpha}_i = \arg \min_{m \in \{1,2,\ldots,2^K\}} D(\boldsymbol{y}_j, \alpha_m) \qquad (10)$$

Results

The proportions of the times in which the true $\boldsymbol{q}_{j_{\text{new}}}$ and the estimated $\hat{\boldsymbol{q}}_{j_{\text{new}}}$ agreed were obtained for each condition and summarized in two different ways: one is the Component-wise Agreement Rate (CAR) $= (\sum_{j_{\text{new}}=1} \sum_{k=1} |q_{j_{\text{new}}k} = \hat{q}_{j_{\text{new}}k}|)/(J_{\text{new}} \times K)$, and the other one is the Vector-wise Agreement Rate (VAR) $= (\sum_{j_{\text{new}}=1} |\mathbf{q}_{j_{\text{new}}} = \hat{\mathbf{q}}_{j_{\text{new}}}|)/J_{\text{new}}$. The probabilities of tied $\mathbf{q}_{j\text{new}}$ vector were obtained by determining the proportions of ties in each condition under each model, and then averaging them by the size of sample size I.

The results summarized in Table 2 support the results of Proposition 1 that the probability of ties decreases as the sample size increases.

Table 3 documents the agreement rates when data sets were simulated from the DINA model. In support

Table 2. Average rates of ties among $2^K - 1$ q-vectors.

	$\alpha \sim MVN(0, \Sigma)$ with $\rho = .3$				$\alpha \sim MVN(0, \Sigma)$ with $\rho = .6$			
I	DINA	R-RUM	DINO	C-RUM	DINA	R-RUM	DINO	C-RUM
	True Q							
250	0.243	0.288	0.270	0.058	0.372	0.383	0.382	0.111
1000	0.096	0.164	0.104	0.021	0.200	0.192	0.194	0.090
5000	0.022	0.026	0.022	0.004	0.060	0.081	0.063	0.085
	10% misspecified Q							
250	0.247	0.266	0.273	0.066	0.341	0.454	0.393	0.106
1000	0.109	0.213	0.095	0.024	0.168	0.205	0.194	0.089
5000	0.022	0.032	0.023	0.008	0.048	0.069	0.049	0.085

of the theoretical results, the proposed method performs well when the slipping and guessing parameters were less than .1, and deteriorate overall at larger levels of the item parameters. Nevertheless, the CAR remains above .664 when the attributes were generated from $MVN(0, \Sigma)$ with $\rho = .3$, and .607 when they are from $MVN(0, \Sigma)$ with $\rho = .6$, even at the largest noise level (Max = .5).

Consistent with the asymptotic theory, the agreement rates increase as sample size increases. However, for larger K, results suffer when there are few assessment items. This is due to less accurate estimation of examinees' latent attribute profiles, which are taken as known when estimating the $\mathbf{q}_{\text{new}}$ vectors. For example, for the condition of $K = 7, J = 20$, Max $= .1$, the agreement rates decrease from .857 to .846 for CAR, and .352 to .302 for VAR as the accuracy of examinees' latent attribute α estimates do .857 to .745 at CAR, and .496 to .270 at VAR as sample size increases from 1000 to 5000. When the number of attributes and sizes of slip and guess parameters are all large, correctly classifying the latent attribute vector as well as the q-vector is more difficult.

This approach produces nearly perfect estimation rates when the number of attributes was $K = 3$, the upper bound of item parameters was less than .3, and the number of assessment items was larger than 40. Additionally, when the attributes are less correlated, the agreement rates appear to be more consistent and higher. The results for the DINO model in Table 4, and the R-RUM in Table 5 are

Table 3. Agreement rates between Q_{new} and $\hat{Q}_{\text{new}}$ under the DINA model.

			$\alpha \sim MVN(0, \Sigma)$ with $\rho = .3$						$\alpha \sim MVN(0, \Sigma)$ with $\rho = .6$					
			Max = .1		Max = .3		Max = .5		Max = .1		Max = .3		Max = .5	
K	J	I	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR
3	20	250	0.999	0.996	0.879	0.692	0.778	0.475	0.956	0.867	0.911	0.744	0.766	0.465
		1000	1.000	1.000	0.993	0.980	0.901	0.768	1.000	1.000	0.982	0.945	0.735	0.390
		5000	1.000	1.000	0.953	0.896	0.816	0.535	1.000	1.000	0.949	0.846	0.755	0.418
	40	250	0.999	0.996	0.973	0.921	0.853	0.624	1.000	1.000	0.917	0.759	0.855	0.641
		1000	1.000	1.000	0.990	0.970	0.890	0.711	0.999	0.997	0.982	0.947	0.897	0.728
		5000	1.000	1.000	1.000	1.000	0.949	0.863	1.000	1.000	1.000	0.999	0.937	0.831
	60	250	0.998	0.995	0.990	0.971	0.899	0.756	0.986	0.957	0.874	0.625	0.851	0.630
		1000	1.000	1.000	1.000	0.999	0.931	0.822	1.000	1.000	0.994	0.982	0.875	0.668
		5000	1.000	1.000	1.000	1.000	0.968	0.914	1.000	1.000	0.997	0.992	0.940	0.838
5	20	250	0.936	0.735	0.852	0.489	0.714	0.204	0.950	0.782	0.829	0.383	0.675	0.168
		1000	0.965	0.823	0.855	0.436	0.724	0.202	0.891	0.489	0.817	0.309	0.663	0.141
		5000	1.000	0.999	0.877	0.518	0.695	0.156	0.863	0.440	0.840	0.408	0.756	0.261
	40	250	0.991	0.956	0.911	0.614	0.746	0.242	0.906	0.631	0.856	0.455	0.734	0.197
		1000	0.999	0.996	0.955	0.802	0.845	0.467	0.994	0.970	0.882	0.489	0.792	0.320
		5000	1.000	1.000	0.959	0.794	0.780	0.305	0.959	0.800	0.910	0.580	0.841	0.430
	60	250	0.935	0.681	0.919	0.672	0.808	0.410	0.923	0.642	0.862	0.462	0.756	0.261
		1000	0.969	0.852	0.959	0.806	0.851	0.499	0.954	0.769	0.918	0.648	0.761	0.263
		5000	0.989	0.946	0.945	0.740	0.843	0.478	0.984	0.919	0.923	0.621	0.848	0.466
7	20	250	0.894	0.479	0.819	0.279	0.664	0.059	0.779	0.144	0.720	0.091	0.607	0.019
		1000	0.857	0.352	0.805	0.179	0.686	0.071	0.803	0.188	0.750	0.107	0.647	0.053
		5000	0.846	0.302	0.811	0.208	0.734	0.110	0.803	0.162	0.756	0.131	0.662	0.058
	40	250	0.829	0.287	0.851	0.349	0.677	0.083	0.825	0.247	0.801	0.182	0.676	0.061
		1000	0.957	0.742	0.901	0.468	0.779	0.195	0.885	0.397	0.817	0.220	0.758	0.165
		5000	0.941	0.680	0.881	0.430	0.750	0.127	0.901	0.441	0.852	0.272	0.742	0.091
	60	250	0.912	0.528	0.854	0.297	0.720	0.105	0.868	0.383	0.818	0.240	0.692	0.090
		1000	0.938	0.623	0.935	0.638	0.746	0.132	0.889	0.421	0.844	0.265	0.770	0.155
		5000	0.952	0.684	0.907	0.494	0.823	0.247	0.929	0.581	0.833	0.232	0.763	0.150

Table 4. Agreement rates between Q_{new} and $\hat{Q}_{\text{new}}$ under the DINO model.

			$\alpha \sim MVN(0, \Sigma)$ with $\rho = .3$						$\alpha \sim MVN(0, \Sigma)$ with $\rho = .6$					
			Max = .1		Max = .3		Max = .5		Max = .1		Max = .3		Max = .5	
K	*J*	*I*	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR
3	20	250	1.000	1.000	0.979	0.940	0.797	0.503	0.946	0.838	0.959	0.880	0.707	0.350
		1000	1.000	0.999	0.990	0.969	0.801	0.514	0.996	0.988	0.888	0.663	0.818	0.563
		5000	1.000	1.000	0.811	0.557	0.886	0.697	0.974	0.921	0.929	0.788	0.847	0.607
	40	250	1.000	1.000	0.974	0.924	0.854	0.633	0.995	0.984	0.978	0.938	0.802	0.535
		1000	1.000	1.000	0.980	0.941	0.900	0.751	1.000	1.000	0.967	0.900	0.907	0.754
		5000	1.000	1.000	1.000	1.000	0.952	0.873	1.000	1.000	0.989	0.966	0.889	0.716
	60	250	0.999	0.996	0.992	0.977	0.880	0.695	0.954	0.863	0.977	0.931	0.859	0.638
		1000	1.000	1.000	0.997	0.990	0.941	0.850	0.995	0.986	0.994	0.983	0.892	0.731
		5000	1.000	1.000	1.000	1.000	0.957	0.889	1.000	1.000	0.999	0.997	0.929	0.807
5	20	250	0.839	0.369	0.852	0.422	0.638	0.110	0.875	0.456	0.839	0.406	0.674	0.148
		1000	0.941	0.718	0.892	0.509	0.670	0.142	0.940	0.721	0.834	0.387	0.686	0.153
		5000	0.970	0.906	0.942	0.743	0.714	0.205	0.878	0.506	0.829	0.427	0.745	0.243
	40	250	0.931	0.687	0.855	0.450	0.780	0.326	0.938	0.702	0.873	0.469	0.732	0.219
		1000	0.970	0.851	0.957	0.795	0.835	0.437	0.988	0.940	0.895	0.539	0.805	0.364
		5000	0.992	0.958	0.907	0.600	0.839	0.451	0.981	0.906	0.920	0.641	0.802	0.320
	60	250	0.984	0.921	0.968	0.855	0.834	0.467	0.926	0.687	0.860	0.413	0.762	0.266
		1000	0.999	0.995	0.960	0.803	0.885	0.595	0.948	0.740	0.916	0.612	0.805	0.334
		5000	0.990	0.951	0.978	0.895	0.870	0.520	0.973	0.865	0.966	0.834	0.859	0.464
7	20	250	0.797	0.203	0.701	0.106	0.615	0.035	0.809	0.167	0.738	0.100	0.610	0.035
		1000	0.862	0.341	0.795	0.175	0.662	0.053	0.829	0.281	0.788	0.163	0.630	0.042
		5000	0.894	0.398	0.787	0.163	0.705	0.082	0.814	0.186	0.756	0.153	0.623	0.032
	40	250	0.877	0.421	0.815	0.242	0.662	0.057	0.799	0.181	0.746	0.127	0.646	0.049
		1000	0.905	0.462	0.854	0.326	0.727	0.116	0.889	0.403	0.837	0.249	0.734	0.118
		5000	0.924	0.565	0.901	0.481	0.770	0.154	0.909	0.488	0.834	0.225	0.749	0.115
	60	250	0.921	0.529	0.824	0.295	0.708	0.129	0.840	0.312	0.774	0.154	0.678	0.087
		1000	0.956	0.708	0.877	0.367	0.780	0.217	0.910	0.503	0.843	0.258	0.738	0.117
		5000	0.979	0.884	0.942	0.659	0.824	0.285	0.925	0.562	0.873	0.368	0.764	0.137

Table 5. Agreement rates between Q_{new} and $\hat{Q}_{\text{new}}$ under the R-RUM model.

			Reduced RUM				Compensatory RUM			
			$\rho = .3$		$\rho = .6$		$\rho = .3$		$\rho = .6$	
K	*J*	*I*	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR
3	20	250	0.917	0.763	0.853	0.591	0.711	0.361	0.658	0.294
		1000	0.951	0.865	0.891	0.680	0.731	0.413	0.720	0.388
		5000	0.989	0.966	0.989	0.969	0.736	0.447	0.718	0.410
	40	250	0.933	0.807	0.948	0.848	0.707	0.370	0.688	0.324
		1000	0.988	0.966	0.949	0.848	0.724	0.405	0.705	0.382
		5000	0.994	0.982	0.960	0.881	0.729	0.434	0.725	0.415
	60	250	0.975	0.925	0.945	0.840	0.694	0.345	0.704	0.354
		1000	0.975	0.925	0.992	0.976	0.715	0.403	0.722	0.411
		5000	1.000	0.999	0.957	0.872	0.753	0.469	0.724	0.426
5	20	250	0.802	0.343	0.753	0.211	0.648	0.165	0.639	0.118
		1000	0.830	0.408	0.771	0.273	0.704	0.256	0.674	0.173
		5000	0.869	0.487	0.855	0.458	0.713	0.275	0.709	0.271
	40	250	0.879	0.544	0.824	0.419	0.650	0.120	0.641	0.108
		1000	0.937	0.715	0.884	0.522	0.696	0.226	0.678	0.167
		5000	0.956	0.814	0.859	0.432	0.716	0.278	0.704	0.252
	60	250	0.908	0.610	0.893	0.559	0.651	0.144	0.656	0.130
		1000	0.950	0.779	0.888	0.585	0.707	0.234	0.685	0.199
		5000	0.972	0.864	0.918	0.653	0.721	0.299	0.714	0.251
7	20	250	0.759	0.155	0.724	0.101	0.646	0.087	0.625	0.051
		1000	0.776	0.160	0.758	0.127	0.679	0.110	0.670	0.102
		5000	0.787	0.144	0.793	0.173	0.714	0.219	0.676	0.133
	40	250	0.802	0.244	0.726	0.101	0.651	0.068	0.620	0.053
		1000	0.856	0.320	0.797	0.165	0.705	0.129	0.648	0.088
		5000	0.906	0.494	0.838	0.251	0.698	0.169	0.685	0.130
	60	250	0.820	0.260	0.819	0.226	0.641	0.062	0.619	0.037
		1000	0.880	0.405	0.877	0.362	0.689	0.111	0.666	0.087
		5000	0.926	0.597	0.866	0.349	0.714	0.188	0.697	0.144

Table 6. Agreement rates between Q_{new} and $\hat{Q}_{\text{new}}$ under the DINA model with 10% misspecified Q.

			$\alpha \sim MVN(0, \Sigma)$ with $\rho = .3$						$\alpha \sim MVN(0, \Sigma)$ with $\rho = .6$					
			Max = .1		Max = .3		Max = .5		Max = .1		Max = .3		Max = .5	
K	*J*	*I*	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR
3	20	250	0.999	0.998	0.958	0.882	0.743	0.407	0.988	0.964	0.849	0.587	0.757	0.422
		1000	1.000	1.000	0.835	0.528	0.842	0.620	0.997	0.992	0.879	0.676	0.736	0.378
		5000	1.000	1.000	0.864	0.683	0.885	0.712	0.870	0.609	0.842	0.529	0.848	0.603
	40	250	1.000	0.999	0.989	0.968	0.795	0.494	0.858	0.575	0.896	0.704	0.727	0.368
		1000	1.000	1.000	0.995	0.984	0.853	0.621	0.968	0.905	0.941	0.824	0.897	0.730
		5000	1.000	1.000	1.000	1.000	0.886	0.725	1.000	1.000	0.998	0.994	0.856	0.648
	60	250	1.000	1.000	0.991	0.974	0.879	0.704	0.988	0.963	0.956	0.875	0.848	0.644
		1000	1.000	1.000	1.000	1.000	0.916	0.789	0.999	0.998	0.956	0.879	0.899	0.747
		5000	1.000	1.000	1.000	1.000	0.950	0.864	1.000	1.000	0.997	0.992	0.872	0.651
5	20	250	0.882	0.496	0.823	0.409	0.664	0.136	0.910	0.575	0.734	0.212	0.613	0.090
		1000	0.877	0.583	0.778	0.254	0.661	0.129	0.854	0.422	0.847	0.393	0.647	0.131
		5000	0.869	0.522	0.826	0.346	0.652	0.134	0.848	0.413	0.823	0.345	0.655	0.125
	40	250	0.957	0.785	0.870	0.495	0.760	0.298	0.893	0.526	0.792	0.330	0.697	0.172
		1000	0.944	0.766	0.898	0.590	0.768	0.283	0.892	0.500	0.845	0.431	0.709	0.170
		5000	0.963	0.919	0.939	0.751	0.814	0.396	0.914	0.690	0.904	0.584	0.739	0.237
	60	250	0.991	0.954	0.839	0.373	0.714	0.201	0.872	0.465	0.848	0.390	0.695	0.143
		1000	0.988	0.941	0.933	0.675	0.836	0.442	0.931	0.658	0.871	0.458	0.795	0.311
		5000	0.993	0.964	0.913	0.646	0.828	0.403	0.971	0.859	0.901	0.620	0.819	0.377
7	20	250	0.793	0.211	0.728	0.099	0.590	0.027	0.745	0.112	0.694	0.076	0.608	0.031
		1000	0.871	0.373	0.792	0.179	0.614	0.037	0.810	0.190	0.747	0.116	0.612	0.036
		5000	0.853	0.308	0.767	0.150	0.638	0.051	0.806	0.156	0.722	0.077	0.653	0.044
	40	250	0.817	0.215	0.723	0.095	0.652	0.052	0.741	0.079	0.732	0.099	0.643	0.038
		1000	0.896	0.422	0.804	0.209	0.684	0.055	0.842	0.244	0.797	0.155	0.691	0.071
		5000	0.934	0.624	0.894	0.462	0.737	0.142	0.842	0.300	0.771	0.154	0.704	0.074
	60	250	0.855	0.308	0.834	0.251	0.697	0.090	0.826	0.239	0.799	0.240	0.655	0.049
		1000	0.920	0.530	0.868	0.342	0.744	0.132	0.866	0.314	0.826	0.226	0.728	0.097
		5000	0.925	0.562	0.862	0.372	0.789	0.193	0.911	0.491	0.825	0.242	0.726	0.098

similar to those found for the DINA model. As noticed in the simulation of the DINA model, the performance of the proposed method seems to depend on the size of K, the size of the sample, and the levels of items parameters, in precisely the same ways for the DINO model. The results of C-RUM in Table 5 have the same tendencies with those of the other models. The agreement rates deteriorate overall, but the CAR remains larger than .616 for all conditions.

In order to investigate the robustness of the proposed method in the cases in which the assessment item q-vectors were incorrectly specified, 10% of entries in the assessment Q-matrices were randomly misspecified before estimating each examinee's $\hat{\alpha}_i$. Both CAR and VAR remained similar to what was observed in the simulations above in which correct Q-matrices were implemented when $K = 3$, $MVN(0, \Sigma)$ with $\rho = .3$, and the values of item parameters were smaller than 0.1 under the DINA model as Table 6 shows. In other conditions for the DINA model as well as in the DINO model in Table 7, R-RUM and C-RUM in Table 8, both CAR and VAR slightly decreased. Although it requires theoretical proof, we assume that the consistency theorems may be applicable for a certain incorrect classification range of $\hat{\alpha}_i$, and it is possible that some proportion of $\mathbf{Q}$ may be altered while still retaining consistency.

Fraction subtraction data

Data

We analyzed the Tatsuoka fraction subtraction data that include the item responses to 20 items with 8 necessary attributes for 536 examinees to investigate the agreement of the $\mathbf{Q}_{\text{new}}$ estimator with what has been derived through expert opinion. The data, which were originally collected and analyzed by Tatsuoka (1990), have been analyzed in numerous studies. Here we use the Q-matrix for the data that appeared in de la Torre and Douglas (2004) shown in Table 9. The specified attributes are (1) convert a whole number to a fraction, (2) separate a whole number from fraction, (3) simplify before subtracting, (4) find a common denominator, (5) borrow from whole number part, (6) column borrow to subtract the second numerator from the first, (7) subtract numerators, and (8) reduce answers to simplest form.

Results

The q-vector for each individual item was estimated based on the information obtained from the other 19 items. In particular, each item took its turn as the studied item, while the remaining 19 items comprised the assessment test. Once attribute vectors were estimated

Table 7. Agreement rates between Q_{new} and $\hat{Q}_{\text{new}}$ under the DINO model with 10% misspecified Q.

			$\alpha \sim MVN(0, \Sigma)$ with $\rho = .3$						$\alpha \sim MVN(0, \Sigma)$ with $\rho = .6$					
			Max = .1		Max = .3		Max = .5		Max = .1		Max = .3		Max = .5	
K	*J*	*I*	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR
3	20	250	1.000	1.000	0.961	0.886	0.767	0.483	0.998	0.995	0.855	0.604	0.809	0.526
		1000	1.000	1.000	0.954	0.868	0.825	0.564	0.964	0.893	0.853	0.601	0.756	0.422
		5000	1.000	1.000	0.996	0.989	0.901	0.739	1.000	1.000	0.985	0.954	0.809	0.507
	40	250	0.994	0.983	0.981	0.945	0.816	0.559	0.991	0.972	0.918	0.767	0.794	0.481
		1000	1.000	1.000	0.996	0.987	0.868	0.659	0.996	0.989	0.990	0.970	0.862	0.643
		5000	1.000	1.000	1.000	1.000	0.916	0.788	1.000	1.000	0.936	0.808	0.862	0.639
	60	250	1.000	1.000	0.991	0.976	0.876	0.685	0.999	0.997	0.954	0.865	0.859	0.660
		1000	1.000	1.000	1.000	0.999	0.891	0.715	1.000	1.000	0.936	0.807	0.902	0.742
		5000	1.000	1.000	1.000	1.000	0.964	0.903	1.000	1.000	1.000	1.000	0.914	0.778
5	20	250	0.885	0.536	0.755	0.218	0.676	0.174	0.825	0.329	0.728	0.185	0.634	0.087
		1000	0.936	0.688	0.884	0.559	0.650	0.118	0.871	0.466	0.776	0.287	0.696	0.161
		5000	0.953	0.772	0.839	0.430	0.737	0.253	0.874	0.565	0.751	0.219	0.629	0.082
	40	250	0.925	0.640	0.809	0.336	0.724	0.219	0.810	0.321	0.804	0.324	0.655	0.109
		1000	0.951	0.758	0.871	0.502	0.760	0.273	0.957	0.814	0.817	0.341	0.752	0.220
		5000	0.983	0.913	0.910	0.645	0.736	0.271	0.911	0.628	0.874	0.463	0.804	0.313
	60	250	0.971	0.856	0.902	0.594	0.711	0.205	0.946	0.740	0.890	0.524	0.741	0.239
		1000	0.990	0.957	0.927	0.687	0.808	0.380	0.961	0.809	0.869	0.442	0.742	0.233
		5000	1.000	1.000	0.959	0.822	0.825	0.414	0.969	0.848	0.956	0.800	0.811	0.362
7	20	250	0.829	0.242	0.764	0.138	0.634	0.035	0.800	0.164	0.721	0.091	0.577	0.022
		1000	0.835	0.269	0.684	0.048	0.636	0.037	0.784	0.137	0.694	0.059	0.658	0.061
		5000	0.892	0.461	0.742	0.105	0.677	0.051	0.823	0.214	0.722	0.100	0.612	0.029
	40	250	0.858	0.318	0.794	0.199	0.648	0.043	0.794	0.175	0.742	0.097	0.632	0.040
		1000	0.896	0.440	0.844	0.265	0.701	0.075	0.784	0.130	0.824	0.222	0.711	0.094
		5000	0.913	0.512	0.834	0.262	0.719	0.110	0.890	0.414	0.812	0.195	0.667	0.055
	60	250	0.852	0.317	0.814	0.235	0.677	0.078	0.859	0.312	0.790	0.167	0.662	0.070
		1000	0.927	0.577	0.864	0.353	0.720	0.098	0.923	0.537	0.815	0.228	0.724	0.100
		5000	0.969	0.823	0.899	0.453	0.743	0.127	0.925	0.546	0.847	0.260	0.747	0.112

Table 8. Agreement rates between Q_{new} and $\hat{Q}_{\text{new}}$ under the R-RUM model with 10% misspecified Q.

			Reduced RUM				Compensatory RUM			
			$\rho = .3$		$\rho = .6$		$\rho = .3$		$\rho = .6$	
K	*J*	*I*	CAR	VAR	CAR	VAR	CAR	VAR	CAR	VAR
3	20	250	0.925	0.806	0.871	0.662	0.698	0.341	0.651	0.281
		1000	0.919	0.764	0.843	0.613	0.722	0.408	0.704	0.351
		5000	0.941	0.837	0.892	0.680	0.740	0.447	0.736	0.424
	40	250	0.958	0.877	0.925	0.785	0.705	0.353	0.677	0.295
		1000	0.991	0.972	0.960	0.888	0.717	0.397	0.721	0.404
		5000	1.000	1.000	0.965	0.894	0.731	0.439	0.724	0.426
	60	250	0.968	0.910	0.911	0.744	0.713	0.374	0.680	0.311
		1000	0.997	0.990	0.986	0.958	0.722	0.408	0.728	0.420
		5000	0.999	0.998	0.993	0.980	0.738	0.448	0.738	0.447
5	20	250	0.767	0.293	0.721	0.182	0.671	0.166	0.631	0.135
		1000	0.842	0.410	0.736	0.211	0.702	0.234	0.672	0.194
		5000	0.853	0.492	0.769	0.226	0.715	0.293	0.712	0.260
	40	250	0.853	0.440	0.786	0.358	0.689	0.191	0.641	0.121
		1000	0.885	0.543	0.834	0.426	0.712	0.230	0.683	0.192
		5000	0.951	0.771	0.824	0.380	0.714	0.279	0.717	0.270
	60	250	0.906	0.616	0.813	0.327	0.647	0.135	0.631	0.111
		1000	0.901	0.637	0.878	0.475	0.720	0.247	0.666	0.165
		5000	0.976	0.889	0.933	0.698	0.712	0.284	0.710	0.249
7	20	250	0.708	0.091	0.705	0.079	0.654	0.074	0.620	0.045
		1000	0.777	0.156	0.712	0.093	0.682	0.113	0.661	0.077
		5000	0.764	0.125	0.715	0.075	0.705	0.190	0.694	0.152
	40	250	0.782	0.178	0.763	0.149	0.644	0.060	0.606	0.038
		1000	0.817	0.239	0.770	0.143	0.682	0.110	0.664	0.103
		5000	0.850	0.315	0.832	0.227	0.692	0.179	0.703	0.146
	60	250	0.791	0.199	0.743	0.126	0.646	0.075	0.618	0.044
		1000	0.861	0.325	0.800	0.206	0.681	0.109	0.646	0.070
		5000	0.898	0.463	0.866	0.331	0.709	0.178	0.688	0.136

Table 9. *Q* for the fraction subtraction data.

Item	*K* = 8							
1	0	0	0	1	0	1	1	0
2	0	0	0	1	0	0	1	0
3	0	0	0	1	0	0	1	0
4	0	1	1	0	1	0	1	0
5	0	1	0	1	0	0	1	1
6	0	0	0	0	0	0	1	0
7	1	1	0	0	0	0	1	0
8	0	0	0	0	0	0	1	0
9	0	1	0	0	0	0	0	0
10	0	1	0	0	1	0	1	1
11	0	1	0	0	1	0	1	0
12	0	0	0	0	0	0	1	1
13	0	1	0	1	1	0	1	0
14	0	1	0	0	0	0	1	0
15	1	0	0	0	0	0	1	0
16	0	1	0	0	0	0	1	0
17	0	1	0	0	1	0	1	0
18	0	1	0	0	1	1	1	0
19	1	1	1	0	1	0	1	0
20	0	1	1	0	1	0	1	0

from the assessment tests, the *q*-vector of the excluded item was estimated. The real data analysis used the CAR measure introduced in the simulation study to summarize how the estimated **Q** entries agree with published values of **Q** determined through expert opinion. As shown in Table 10, perfect agreement rates were obtained for 9 items (CAR = 1.000), and differed by only entry for 8 items (CAR = 0.875) out of the 20 items. The overall CAR value for the exam was 0.913. Though there is no true *Q*-matrix to refer to in this study, we can see that this data-driven method agrees nearly 91 percent of the time with the values chosen by researchers.

Nonparametric *Q* refinement method

The purpose of the section is to propose an algorithm to refine the number of attributes in *Q*-matrix by using the method proposed in this article. As shown in Table 8, 8 skills are believed to be involved in solving the fraction subtraction test. DeCarlo (2011) fitted several extended DINA models to the data and noted the misspecification of **Q** from large estimates of class sizes. He indicated that the third skill (3: simplify before subtracting) should be eliminated from the *Q*-matrix because it is unnecessary. As an example, we validate his findings as follows.

Table 10. Agreement rates between estimated q-vector and q-vector for each item.

Item	Agreement	Item	Agreement
1	0.875	11	1.000
2	1.000	12	0.750
3	1.000	13	0.875
4	0.875	14	1.000
5	0.750	15	0.875
6	1.000	16	1.000
7	1.000	17	1.000
8	1.000	18	0.875
9	0.875	19	0.875
10	0.750	20	0.875

Step 1. For a given $J \times K$ *Q*-matrix, let Q^{υ}, $\upsilon = 1, 2, \ldots, 2^K = V$ denote its subsets of $Q_{\max}$ obtained by selecting columns. Then define $Q_{\max}$ as $\{Q^1, Q^2, \ldots, Q^V\}$ for the 2^K combinations of K. To illustrate, given the *Q*-matrix with $K = 8$ for the fraction subtraction data, $2^8 = 256$ matrices were generated where the first 8 matrices were $J \times 1$ matrices, the next 28 matrices contained two columns of *Q*-matrix, etc.

Step 2. Given each $Q^{\upsilon} \subset Q_{\max}$, examinees' class membership is obtained by using the reverse technique of the proposed method to construct ideal response patterns. Then estimate the RSS_{υ}, $\upsilon = 1, \ldots, V$ as follows:

$$\mathrm{RSS}_{\upsilon} = \sum_{i=1}^{I} \sum_{j=1}^{J} \mid Y_{ij} - \eta_{ij} \mid \tag{11}$$

Step 3. Classify Q^{υ}, $\upsilon = 1, 2, \ldots, V$ into K groups by their number of required attributes. For the *Q*-matrix with $K = 8$ for the fraction subtraction data, eight groups were formed, e.g., all Q^{υ} in the first group with $K' = 1$ had one column, all Q^{υ} in the second group with $K' = 2$ had two columns, and as forth.

Step 4. In each group, find the Q^{υ} having the smallest RSS. Then $Q_{\max_{\min}} = \{Q_{\min}^{K'=1}, Q_{\min}^{K'=2}, \ldots, Q_{\min}^{K'=K}\}$.

Step 5. Identify the Q_{correct} with the correct number of attributes from $Q_{\max_{\min}}$ by comparing the distance between $Q_{\min}^{K'} \subset Q_{\max_{\min}}$. Since an overspecified *Q* barely causes an increase in RSS (Lim & Drasgow, under 2nd revision), the Q_{correct} with an accurate number of attributes is the RSS $Q_{\min}$ having the smaller number of attributes.

Since the fraction subtraction data *Q* was hypothesized to measure 8 different attributes, from Steps 1 to 4, we identified eight subsets of $Q_{\min}^{K'}$, $K' = 1, \ldots, 8$ as shown in Figure 1. The subset $Q_{\min}^{K'=1}$ measures Attribute 7; $Q_{\min}^{K'=2}$ Attributes 5 and 7; $Q_{\min}^{K'=3}$ Attributes 1, 5, and 7; $Q_{\min}^{K'=4}$ Attributes 1, 4, 5, and 7; $Q_{\min}^{K'=5}$ Attributes 1, 2, 4, 5, and 7; $Q_{\min}^{K'=6}$ Attributes 1, 2, 4, 5, 7, and 8; $Q_{\min}^{K'=7}$ Attributes 1, 2, 4, 5, 6, 7, and 8; $Q_{\min}^{K'=8}$ Attributes 1, 2, 3, 4, 5, 6, 7, 8.

Like the results of DeCarlo (2011), the RSS was stable when the third skill (3: simplify before subtracting) was eliminated ($QQ_{\min 7}$). When the sixth skill (6: column borrow to subtract the second numerator from the first) was eliminated, the RSS increased very little. Thus, this analysis suggests that $Q_{\min}^{K'=6}$ fits the data well.

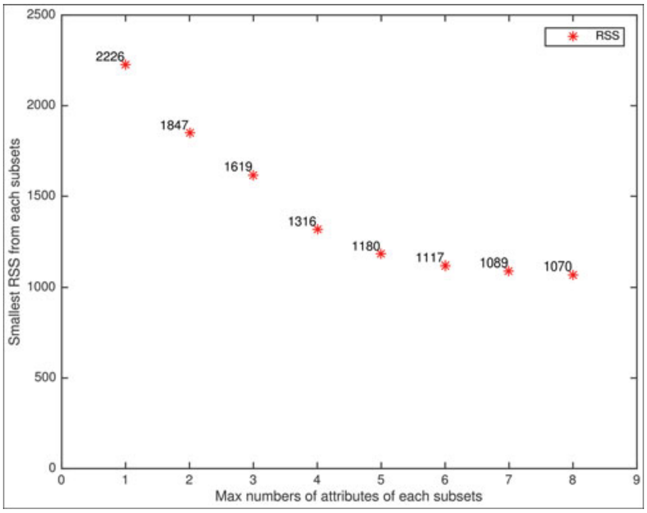

Figure 1. Refinement of number of attribute for fraction subtraction data.

Discussion

The significance of this research is that it offers an accurate and automated statistical method for calibrating Q matrices for new items, once Q has been established for a set of previously calibrated items. Most commonly, expert opinion is utilized to establish a Q-matrix, perhaps followed by extensive statistical validation. However, once a large item bank with a known Q-matrix has been established, the new nonparametric technique may be used to construct q-vectors for additional items, provided they require the same set of attributes. The method requires no assumed parametric cognitive diagnosis model, and was shown to be consistent under a very general class of cognitive diagnosis models. However, the method can easily be adapted to parametric modeling settings, and may be more efficient if the assumed model is correct.

A second important use of the new method, as seen in the analysis of the Tatsuoka data, is that it may be used one item at a time, to validate Q matrices determined through expert opinion. Experts are expert, but not necessarily infallible. For the Tatsuoka data, the new analysis largely confirmed their judgments, but suggested a few changes in the Q-matrix. Such confirmatory studies are valuable in that they provide objective evidence for the Q-matrix.

In both the theory and in simulation, it was clear that a long assessment exam is desirable for fitting calibration items, so that estimated attribute patterns closely approximate their true values. Interestingly, in the real data analysis when only 19 items were included after leaving out the studied item, there was a very high agreement rate with q-vectors determined through expert opinion. However, unlike in simulation, we cannot know the true values, or even if the attributes were correctly identified and labeled. Nevertheless, the high agreement rate gives us confidence that cognitive diagnosis modeling is meaningful in this setting. The computational limitations of this approach are the difficulty of estimating q-patterns when K gets larger, for example, into the tens or twenty as well as when the noise level gets larger.

Whether diagnostic assessment is the primary goal of a testing program, or is meant to offer additional information in reporting for a finer grained evaluation, nonparametric techniques are simple, easy to implement, and theoretically supported. The purpose of this article is to offer an efficient method for building the Q-matrix as items are introduced into an item bank, without relying on further consultation with experts. Nevertheless, to the extent resources are available, some interplay between statistical methods and expert opinion will always be desirable.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported by a grant.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors. institutions is not intended and should not be inferred.

ORCID

Youn Seon Lim http://orcid.org/0000-0003-0225-1527

References

Barnes, T. (2003). The Q-matrix method of fault-tolerant teaching in knowledge assessment and data mining, North Carolina State University, USA. Retrieved from http://www.lib.ncsu.edu/resolver/1840.16/4612

Brewer, P. (1996). Methods for concept mapping in computer based education. Unpublished master thesis dissertation, North Carolina State University, Raleigh, NC.

Chiu, C. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, 37, 598–618. doi:10.1177/0146621613488436

Chiu, C., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response pattern.

Journal of Classification, 30, 225–250. doi:10.1007/s00357-013-9132-9

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

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, 36, 447–468. doi:10.1177/0146621612449069

de la Torre, J., & Chiu, C. Y. (2010, April). General empirical method of Q-Matrix validation. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

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. (2008). An empirically based method of q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, 45, 343–362. doi:10.1111/j.1745-3984.2008.00069.x

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

de la Torre, J., & Chiu, C. (2015). A general method of empirical Q-matrix validation. *Psychometrika*, 1–21. doi:10.1007/s11336-015-9467-8

Hartz, S.M., & Roussos, L.A. (2008). The fusion model for skills diagnosis: Blending theory with practice. *ETS Research Report RR-08-71*. Princeton, NJ: Educational Testing Service.

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. doi:10.1007/s11336-008-9089-5

Hubal, R. (1992). Retro-adaptive testing and evaluation system. Unpublished masters thesis, North Carolina State University, Raleigh, NC.

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

Im, S., & Corter, J. E. (2011). Statistical consequences of attribute misspecification in the rule space method. *Educational and Psychological Measurement*, 74, 712–731. doi:10.1177/0013164410384855

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

Lim, Y. S., & Drasgow, F. (2017). Conditional independence and dimensionality of cognitive diagnostic models: A test for model fit. *Journal of Classification*.

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

Liu, J., Xu, G, & Ying, Z. (2011). Theory of the self-learning Q-matrix. *Bernoulli*, 36, 548–564. doi:10.3150/12-BEJ430

Rupp & Templin, (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

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

Roussos, L.A., DiBello, L.V., Stout, W.F., Hartz, S.M., Henson, R.A., & Templin, J.L. (2007). The fusion model skills diagnosis system. In J.P. Leighton & M.J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). Cambridge, UK: Cambridge University Press.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12, 257–285. doi:10.1016/0364-0213(88)90023-7

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

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345–354. doi:10.1111/j.1745-3984.1983.tb00212.x

Templin, J. (2006). *CDM user's guide*. Unpublished manuscript, University of Kansas, Kansas, NE.

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

von Davier, M. (2005, September). A general diagnostic model applied to language testing data. *(Research report No. RR-05-16)*. Princeton, NJ: Educational Testing Service.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61, 287–301. doi:10.1348/000711007X193957

Wang, S., & Douglas, J. (2013). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika*, 80(1), 85–100. doi:10.1007/s11336-013-9372-y

Zheng, Y., & Chiu, C. (2015). NPCD: An R package for nonparametric methods for cognitive diagnosis. *R package version 1.0-9*. Retrieved from http://CRAN.R-project.org/package=NPCD

Appendix A
Asymptotic consistency of the $\mathbf{Q}_{\text{new}}$ estimator

Typically asymptotic theory concerns the behavior of statistics as the sample size approaches infinity. However, in order to obtain consistent estimates with the proposed method, it is required that the number of assessment items J and the sample size I increase together, because accurate estimation of attribute profiles is needed for consistent estimation of q-vectors for new items.

Listed below are assumptions that will be used in the subsequent development. Throughout this section, the set C denotes the collection of all possible $2^K - 1$ q-patterns for a fixed number of attributes K. For item j_{new} in the calibration set of items, an arbitrary element of C is denoted by $\boldsymbol{q}_{j_{\text{new}}c}$, and the true q-vector is denoted by $\boldsymbol{q}_{j_{\text{new}}}$. Similarly $\eta_{j_{\text{newc}}}$ indicated the ideal response for $\mathbf{q}_{j\text{newc}}$ and $\eta_{j_{\text{new}}}$ is the ideal response for the true q-vector $\boldsymbol{q}_{j_{\text{new}}}$.

A1. For each examinee i, the collection of assessment item responses $\{y_{i1}, y_{i2}, \ldots, y_{iJ}\}$ is conditionally independent given attribute vector α_i.

A2. For each assessment item j, $\{y_{1j}, y_{2j}, \ldots, y_{Ij}\}$ are statistically independent. The same holds for item j_{new} in the calibration set.

A3. For some number $\delta \in (0, .5)$, $P(y_{ij} = 1 \mid \eta_{ij} = 0) < .5 - \delta$, $P(y_i = 1 \mid \eta_{ij} = 1) > .5 + \delta$, for all examinees i and assessment items j. The same holds for all calibration items j_{new}.

A4. The population proportion for each attribute pattern α is greater than 0.

A5. $\forall \varepsilon > 0$, $Ie^{-2J\varepsilon^2} \to 0$ as $J \to \infty, I \to \infty$.

A6. Define $A_{m,m'} = \{j \mid \eta_{mj} \neq \eta_{m'j}\}$, where m and m' index different attribute patterns among the 2^K possible patterns. $\text{Card}(A_{m,m'}) > \lambda J$ for some $\lambda > 0$ for sufficiently large J.

Assumption A3 reflects the essence of a cognitive diagnosis model. More specially, in a noncompensatory attribute relationship, those who are missing some required attribute are more likely to answer incorrectly than correctly, and those who are masters of all required attributes should answer correctly. In a compensatory attribute relationship, those who master none of attributes are more likely to answer incorrectly than correctly, and those who master one or more of required attributes should answer correctly. Bounding these probabilities away from 0.5 is required for the items to be uniformly informative. Assumption A5 controls the rate at which the assessment test must grow in order for attribute profile estimation to be accurate simultaneously for all examinees, and assumption A6 is essentially a condition on the Q-matrix of the assessment items that is needed to consistently estimate attribute profiles that are subsequently used for q-vector estimation.

Next, two propositions that are used in the proof of the theorem on consistency are stated and proven. Proposition 1 below concerns the expected Hamming distance between observed and ideal responses for a calibration item, and how this is minimized at the true q-vector.

Proposition 1. *Under assumptions A1–A4, and assuming α_i is known for each examinee i, if $\boldsymbol{q}_{j_{new}c} \neq \boldsymbol{q}_{j_{new}}$ then*

$$\lim_{I\to\infty}\left[\sum_{i=1}^{I} E \mid y_{ij_{new}} - \eta_{ij_{new}c} \mid - \sum_{i=1}^{I} E \mid y_{ij_{new}} - \eta_{ij_{new}} \mid\right] = \infty$$

Proof. We need to show that for a randomly selected examinee $E(|y_{j_{\text{new}}} - \eta_{j_{\text{new}}c}|) > E(|y_{j_{\text{new}}} - \eta_{j_{\text{new}}}|)$. When $\boldsymbol{q}_{j_{\text{new}}c} \neq \boldsymbol{q}_{j_{\text{new}}}$, there must be a value for α that results in a different ideal response, $\eta_{j_{\text{new}}c} \neq \eta_{j_{\text{new}}}$. We define $A_c = \{\alpha \mid \eta_{j_{\text{new}}c} \neq \eta_{j_{\text{new}}}\}$ to indicate examinees having $\eta_{j_{\text{new}}c} \neq \eta_{j_{\text{new}}}$. Only attribute profiles in A_c contribute to the inequality, because for α in the complement of A_c, $E(|y_{j_{\text{new}}} - \eta_{j_{\text{new}}c}|) - E(|y_{j_{\text{new}}} - \eta_{j_{\text{new}}}|) = 0$. It follows that

$$\begin{aligned} &E(|y_{j_{\text{new}}} - \eta_{j_{\text{new}}c}|) - E(|y_{j_{\text{new}}} - \eta_{j_{\text{new}}}|) \\ &\quad = P(A_c)[E(|y_{j_{\text{new}}} - \eta_{j_{\text{new}}c} \mid\mid \alpha \in A_c) \\ &\qquad - E(|y_{j_{\text{new}}} - \eta_{j_{\text{new}}} \mid\mid \alpha \in A_c)] \\ &\quad \geq P(A_c)\delta > 0 \end{aligned}$$

by expanding the expectations and applying A3 and A4. Then by summing over the I examinees,

$$\begin{aligned} &\sum_{i=1}^{I} E \mid y_{ij_{\text{new}}} - \eta_{ij_{\text{new}}c} \mid - \sum_{i=1}^{I} E \mid y_{ij_{\text{new}}} - \eta_{ij_{\text{new}}} \mid \\ &\quad = I[E \mid y_{j_{\text{new}}} - \eta_{j_{\text{new}}c} \mid - E \mid y_{j_{\text{new}}} - \eta_{j_{\text{new}}} \mid] \\ &\quad \geq IP(A_c)\delta \end{aligned}$$

which tends to infinity as I grows to infinity. □

Because the estimator relies on minimizing observed distances from ideal responses, we need to see how close the estimated distances are to their expected values. Let $d_i(\boldsymbol{q}_{j_{\text{new}}c})$ be $\mid y_{j_{\text{new}}} - \eta_{j_{\text{new}}c} \mid$ in subsequent proofs.

Proposition 2. *Under A1–A4, and A6, assuming α_i is known for each examinee, for any arbitrarily positive small number ϵ*

$$P\left[\left|\frac{1}{I}\sum_{i=1}^{I}\left(d_i(\boldsymbol{q}_{j_{new}c}) - E[d_i(\boldsymbol{q}_{j_{new}c})]\right)\right| \geq \epsilon\right] \leq 2e^{-2I\epsilon^2}$$

Proof. Under A1 and A2, for any calibration item j_{new}, $d_1(\boldsymbol{q}_{j_{\text{new}}c}), d_2(\boldsymbol{q}_{j_{\text{new}}c}), \ldots, d_I(\boldsymbol{q}_{j_{\text{new}}c})$ are independent random variables, such that $0 \leq d_i(\boldsymbol{q}_{j_{\text{new}}c}) \leq 1$. Then it follows from a result of Hoeffding's (1963) concerning deviations of means of bounded random variables from their expected values that

$$\begin{aligned} &P\left(\left|\frac{1}{I}\sum_{i=1}^{I}\left(d_i(\boldsymbol{q}_{j_{\text{new}}c}) - E[d_i(\boldsymbol{q}_{j_{\text{new}}c})]\right)\right| \geq \epsilon\right) \\ &\quad = P\left(\left|d_1(\boldsymbol{q}_{j_{\text{new}}c}) - E[d_1(\boldsymbol{q}_{j_{\text{new}}c})]\right| \geq I\epsilon\right) \\ &\quad \leq 2e^{-\frac{2(I\epsilon)^2}{I}} \\ &\quad = 2e^{-2I\epsilon^2} \end{aligned}$$

□

Next, we consider a theorem of Wang and Douglas (2013), that provides a probability inequality for the estimated attribute pattern differing from the observed attribute pattern. This is then utilized along with Propositions 1 and 2 to prove the theorem on consistency of the q-vector.

Theorem 1. *Let $\epsilon > 0$. Under A1–A6, for a large enough sample of examinees I and number of assessment items J*

$$P(\cap_{i=1}^{I}\{\hat{\alpha}_i = \alpha_i\}) \geq 1 - 2^{K+1} I e^{-2J\epsilon^2}$$

Theorem 1 is essentially a restatement of Theorem 2 of Wang and Douglas (2013), and its proof follows directly from the method of the proof for that theorem. Next we state the main theorem.

Theorem 2. *Under assumptions A1–A6, for a calibration item j_{new}*

$$\lim_{I,J \to \infty} P(\hat{\boldsymbol{q}}_{j_{new}} = \boldsymbol{q}_{j_{new}}) = 1$$

Proof. Observe that

$$\begin{aligned}
1 - P\left(\hat{\boldsymbol{q}}_{j_{\text{new}}} = \boldsymbol{q}_{j_{\text{new}}}\right) &= P\left(\hat{\boldsymbol{q}}_{j_{\text{new}}} \neq \boldsymbol{q}_{j_{\text{new}}}\right) \\
&= P\left(\hat{\boldsymbol{q}}_{j_{\text{new}}} \neq \boldsymbol{q}_{j_{\text{new}}} \mid \cap_{i=1}^{I}\{\hat{\alpha}_i = \alpha_i\}\right) \\
&\quad \times P\left(\cap_{i=1}^{I}\{\hat{\alpha}_i = \alpha_i\}\right) \\
&\quad + P\left(\hat{\boldsymbol{q}}_{j_{\text{new}}} \neq \boldsymbol{q}_{j_{\text{new}}} \mid \cup_{i=1}^{I}\{\hat{\alpha}_i \neq \alpha_i\}\right) \\
&\quad \times P\left(\cup_{i=1}^{I}\{\hat{\alpha}_i \neq \alpha_i\}\right) \\
&\leq P\left(\hat{\boldsymbol{q}}_{j_{\text{new}}} \neq \boldsymbol{q}_{j_{\text{new}}} \mid \cap_{i=1}^{I}\{\hat{\alpha}_i = \alpha_i\}\right) \\
&\quad \times P\left(\cap_{i=1}^{I}\{\hat{\alpha}_i = \alpha_i\}\right) + P\left(\cup_{i=1}^{I}\{\hat{\alpha}_i \neq \alpha_i\}\right)
\end{aligned}$$

Thus, by applying Proposition 2 and Theorem 1,

$$\begin{aligned}
1 - P\left(\hat{\boldsymbol{q}}_{j_{\text{new}}} = \boldsymbol{q}_{j_{\text{new}}}\right) &\leq P\left(\hat{\boldsymbol{q}}_{j_{\text{new}}} \neq \boldsymbol{q}_{j_{\text{new}}} \mid \cap_{i=1}^{I}\{\hat{\alpha}_i = \alpha_i\}\right) \\
&\times \left(1 - 2^{K+1} I e^{-2J\epsilon^2}\right) + 2^{K+1} I e^{-2J\epsilon^2}
\end{aligned}$$

With a constant value 2^{K+1}, the probability of the second term in the right side converges to 0 provided the sample size and assessment test length have the relationship $Ie^{-2J\epsilon^2} \to 0$ as $J, I \to \infty$. Thus, we only need to show that $P(\hat{\boldsymbol{q}}_{j_{\text{new}}} \neq \boldsymbol{q}_{j_{\text{new}}} \mid \cap_{i=1}^{I}\{\hat{\alpha}_i = \alpha_i\})$ converges to 0.

This follows readily from Proposition 2 which is under the assumption of a known α_i for each examinee.

$$P\left(\hat{\boldsymbol{q}}_{j_{\text{new}}} \neq \boldsymbol{q}_{j_{\text{new}}}\right) \leq \sum_{\boldsymbol{q}_{j_{\text{new}}c} \neq q_{j_{\text{new}}}} P[d(\boldsymbol{q}_{j_{\text{new}}}) > d(\boldsymbol{q}_{j_{\text{new}}c})].$$

The theorem follows by showing each term in the sum on the right converges to 0. Let $\boldsymbol{q}_{j_{\text{new}}c}$ be any q-vector not equal to $q_{j_{\text{new}}}$. By using the Proposition 1,

$$\begin{aligned}
&P(d(\boldsymbol{q}_{j_{\text{new}}}) > d(\boldsymbol{q}_{j_{\text{new}}c})) \\
&\quad = P[d(\boldsymbol{q}_{j_{\text{new}}}) - d(\boldsymbol{q}_{j_{\text{new}}c}) > 0] \\
&\quad = P[d(\boldsymbol{q}_{j_{\text{new}}}) - E[d(\boldsymbol{q}_{j_{\text{new}}})] + E[d(\boldsymbol{q}_{j_{\text{new}}})] \\
&\qquad - E[d(\boldsymbol{q}_{j_{\text{new}}c})] + E[d(\boldsymbol{q}_{j_{\text{new}}c})] - d(\boldsymbol{q}_{j_{\text{new}}c}) > 0] \\
&\quad \leq P[d(\boldsymbol{q}_{j_{\text{new}}}) - E[d(\boldsymbol{q}_{j_{\text{new}}})] \\
&\qquad - P(A_c)\delta - d(\boldsymbol{q}_{j_{\text{new}}c}) + E[d(\boldsymbol{q}_{j_{\text{new}}c}) > 0] \\
&\quad = P\Big[d(\boldsymbol{q}_{j_{\text{new}}}) - E[d(\boldsymbol{q}_{j_{\text{new}}})] + E[d(\boldsymbol{q}_{j_{\text{new}}c})] \\
&\qquad - d(\boldsymbol{q}_{j_{\text{new}}c}) > P(A_c)\delta\Big] \\
&\quad \leq P\left[d(\boldsymbol{q}_{j_{\text{new}}}) - E[d(\boldsymbol{q}_{j_{\text{new}}})] > \frac{P(A_c)\delta}{2}\right] \\
&\qquad + P\left[d(\boldsymbol{q}_{j_{\text{new}}c}) - E[d(\boldsymbol{q}_{j_{\text{new}}c})] > \frac{P(A_c)\delta}{2}\right] \\
&\quad \leq 4e^{-2I\left(\frac{P(A_c)\delta}{2}\right)^2}
\end{aligned}$$

Because $P(A_c)\delta$ is constant, the probability converges to 0 as I goes to infinity. □