the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2019)

www.wileyonlinelibrary.com

Data-driven Q-matrix validation using a residual-based statistic in cognitive diagnostic assessment

Xiaofeng Yu[1,2] and Ying Cheng[1]*

[1]Department of Psychology, University of Notre Dame, Notre Dame, Indiana, USA
[2]Jiangxi Normal University, Nanchang, Jiangxi, China

In a cognitive diagnostic assessment (CDA), attributes refer to fine-grained knowledge points or skills. The **Q**-matrix is a central component of CDA, which specifies the relationship between items and attributes. Oftentimes, attributes and **Q**-matrix are defined by subject-matter experts, and assumed to be appropriate without any misspecifications. However, this assumption does not always hold in real applications. To address this concern, this paper proposes a residual-based statistic for validating the **Q**-matrix. Its performance is evaluated in a simulation study and compared against that of an existing method proposed in Liu, Xu and Ying (2012, *Applied Psychological Measurement*, 36, 548). Simulation results indicate that the proposed method leads to a higher recovery rate of the **Q**-matrix and is computationally more efficient. The advantage in computational efficiency is particularly pronounced when the number of attributes measured by the test reaches five or more. Results also suggest that the two methods have different tendencies in estimating the attribute vector for each item. In cases where the methods fail to recover the correct **Q**-matrix, the method in Liu et al. (2012, Applied Psychological Measurement, 36, 548) tends to overestimate the number of attributes measured by the items, whereas our method does not show that bias.

1. Introduction

Due to the increasing popularity of and demand for formative assessment, cognitive diagnostic assessment (CDA) has seen rapid development in recent years (Leighton & Gierl, 2007; Rupp, Templin, & Henson, 2010; Tatsuoka, 2009). For a CDA, cognitive diagnostic models (CDMs), sometimes referred to as diagnostic classification models, play the critical role of specifying the statistical relationship between students' item responses and their underlying mastery status of skills or attributes. In the past several decades, the surge of interest in CDA has led to the development of many CDMs. By 2008, more than 60 CDMs had already been publicly released (DiBello, Roussos, & Stout, 2007; Fu & Li, 2007; Rupp & Templin, 2008b).

Most CDMs require a central element, the **Q**-matrix. The **Q**-matrix is a $J \times K$ binary matrix that specifies the attributes measured by each item, where J represents the number of items and K the total number of attributes measured by the CDA. The **Q**-matrix is usually determined by subject-matter experts, so there can be uncertainty, misspecifications, or

*Correspondence should be addressed to Ying Cheng, Department of Psychology, University of Notre Dame, 390 Corbett Hall, Notre Dame, IN 46556, USA (email: ycheng4@nd.edu).

This paper was peer reviewed and accepted under the editorship of Matthias von Davier.

DOI:10.1111/bmsp.12191

disagreement regarding some of its elements (Barnes, 2010; DeCarlo, 2012; de la Torre, 2008). Misspecification in a **Q**-matrix means that some entries that should be 0s are specified as 1s, or vice versa. For example, assuming $K = 3$, the attribute vector of the jth item $\mathbf{q}_j = [1 \quad 1 \quad 0]$ might be misspecified as $[1 \quad 0 \quad 0]$. The misspecification can be very detrimental to a CDA. For example, it may lead to misclassification of students, biased parameter estimates, or even non-identifiability of CDMs (Im & Corter, 2011; Rupp & Templin, 2008a; de la Torre, 2008; Xu & Zhang, 2016).

Therefore, having an accurately specified **Q**-matrix is fundamental for CDAs (Im & Corter, 2011; Lee & Sawaki, 2009; McGlohen & Chang, 2008). Because of this, there is a growing literature on the estimation, validation, or refinement of the **Q**-matrix (Akbay, 2016; Baghaei & Hohensinn, 2017; Chen, 2017; Chen, Culpepper, Chen, & Douglas, 2018; Chiu, 2013; Chung, 2014; Close, 2012; de la Torre, 2008; Feng, 2013; Li, 2016; Lim & Drasgow, 2017; Liu, 2016; Liu, Xu, & Ying, 2012, 2013; Ma, 2014; Romero & Ordonez, 2014; Romero, Ordonez, Ponsoda, & Revuelta, 2014; Xiang, 2013). Among these studies, one line of research focuses on validation, which needs a pre-specified **Q**-matrix to start with. People rely on expert input or student performance data to validate the pre-specified **Q**-matrix (e.g., de la Torre, 2008; de la Torre & Chiu, 2016). Another line of research does not require a pre-specified matrix but searches over the entire space of the **Q**-matrix to obtain a solution (e.g., Chung, 2014). However, the search over the whole **Q**-matrix space, which has $(2^K-1)^J$ elements, is a non-trivial problem, especially for relatively large J and/or K. The methods in this line of research therefore can be computationally burdensome or even infeasible, and may suffer from low accuracy. For these reasons, this study focuses on validation of a pre-specified **Q**-matrix.

In this study we propose a method based on a weighted residual. The rationale is that the residual can serve as an indicator of misfit between the model and the data. By minimizing the residual, one can find the best "fitting" **Q**-matrix. Its properties are proved in ideal conditions. Its performance is then evaluated in a simulation study under a widely used CDM, the deterministic input, noisy output AND gate (DINA) model (see de la Torre, 2009; Junker & Sijtsma, 2001), which assumes a conjunctive relationship among attributes.

The rest of this paper is organized as follows. First, we introduce the relevant background, which includes some notation, the DINA model, and existing methods for **Q**-matrix validation. Then we introduce the proposed method based on a weighted residual. Its performance is then evaluated in a simulation study in comparison to the Liu *et al.* (2012) method. Discussions and implications of the results are given at the end.

2. Background

2.1. Notation and the DINA model

For the sake of convenience but without loss of generality, we first introduce the following terms and notation that will be used in this paper hereafter. Consider a J-item CDA that measures a total of K attributes, each item requiring a distinct subset of attributes. The attribute vector of the jth item is denoted by $\mathbf{q}_j$. The $\mathbf{q}_j$s stacked vertically form the item–attribute association matrix (**Q**-matrix), which is a binary $J \times K$ matrix. Let $\mathbf{Q}_0$ denote the initial **Q**-matrix that was specified in advance, which may contain misspecifications and be different from the true **Q**. Suppose m items in $\mathbf{Q}_0$ are misspecified, which means that for each of the m items at least one entry in the currently defined attribute vector differs from the corresponding attribute vector in **Q**. Suppose there are N test-takers, each possessing a distinct subset of relevant attributes. Denote the attribute mastery pattern (AMP) of the ith

respondent by $\boldsymbol{\alpha}_i$. Let $\mathbf{p}$ represent the probability distribution of AMPs. There are 2^K possible AMPs, and $\mathbf{p}$ is therefore a $2^K \times 1$ vector. We use an $N \times J$ binary matrix $\mathbf{X}$ to represent the response matrix, where X_{ij} is the response of the ith person to the jth item. $X_{ij} = 0$ represents an incorrect response, and $X_{ij} = 1$ otherwise. s_j and g_j are used to denote the true values of the slipping and guessing parameters for the jth item in the DINA model, and $\hat{s}_j$ and $\hat{g}_j$ are their provisional estimates given an estimated attribute vector $\hat{\mathbf{q}}_j$.

Many existing **Q**-matrix validation methods have been built upon the DINA model, such as de la Torre (2008), Liu *et al.* (2012), and Terzi and de la Torre (2018). As one of the most parsimonious models, each item under the DINA model has only two parameters, slipping and guessing, denoted by s_j and g_j, $j = 1, 2, \ldots, J$. Define an indicator variable $\boldsymbol{\eta}_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ summarizing whether the ith respondent possesses all the required attributes of the jth item. If he or she does, $\boldsymbol{\eta}_{ij} = 1$, otherwise $\boldsymbol{\eta}_{ij} = 0$. The probability of a correct response to the jth item by the ith respondent under the DINA model is defined as

$$P(X_{ij} = 1|\boldsymbol{\alpha}_i) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}. \tag{1}$$

Under the DINA model, those who possess all attributes (i.e., those with $\eta_{ij} = 1$) respond correctly with have probability 1–s_j. Those who do not (those with $\eta_{ij} = 0$) provide a correct response with probability g_j.

2.2. *Existing methods for Q-matrix validation*

The fact that the **Q**-matrix is a binary matrix greatly increases the difficulty of **Q**-matrix estimation. With no prior information, completely data-driven approaches are computationally intensive and may lead to relatively low success rates of true **Q**-matrix identification. Fortunately, in the field of educational and psychological assessment, one can rely on expert judgement or historical information to obtain an initial **Q**-matrix, $\mathbf{Q}_0$. Then $\mathbf{Q}_0$ can be refined through the validation process. There are two types of approaches that are commonly used for **Q**-matrix validation: nonparametric approaches and parametric approaches.

As an example, one nonparametric method was developed by Chiu (2013) and Chiu and Kohn (2015), based on the comparison of the residual sum of squares computed from the observed and expected item responses. Results indicate that the method can be applied to find the correct attribute vector under certain conditions. Based on the Hamming distance, Lim and Drasgow (2017) proposed a nonparametric method to classify each respondent into a latent class and validate the appropriateness of the attribute vector for each item. When the underlying model is unknown, the nonparametric method could be preferred. However, nonparametric methods may be less efficient than parametric methods when the underlying model fits the data (Wang *et al.*, 2018).

In this study we primarily focus on the parametric approaches. Among the parametric approaches there are two common types of methods. The first type are the sequential search methods, first proposed in de la Torre's (2008) δ method, and later adopted in the ς^2 method (de la Torre & Chiu, 2016), the γ method (Tu, Cai, & Dai, 2012), and the modified δ method (Terzi & de la Torre, 2018). De la Torre (2008) proposed a sequential search algorithm for the item discrimination index (defined as 1–$\hat{s}_j$–$\hat{g}_j$), in which the attribute vector space for each item is sequentially searched until convergence is reached. The algorithm starts by calculating $\hat{\delta}_j$ based on a single-attribute vector $\hat{\mathbf{q}}_j^{(1)}$. The $\hat{\mathbf{q}}_j^{(1)}$ that is larger

than a user-specified cut-off and results in the largest $\hat{\delta}_j^{(1)}$ among all possible attribute vectors is considered a required attribute. Then the method proceeds to the two-attribute vectors $\hat{\mathbf{q}}_j^{(2)}$, which includes $\hat{\mathbf{q}}_j^{(1)}$ that was identified in the last step as one of the two required attributes. The $\hat{\mathbf{q}}_j^{(2)}$ that results in the largest $\hat{\delta}_j^{(2)}$ and $\hat{\delta}_j^{(2)} > \hat{\delta}_j^{(1)}$ would be the required two-attribute vector. Otherwise, set $\hat{\mathbf{q}}_j^{(1)}$ as the solution and stop. The algorithm proceeds in this manner by adding sequentially more required attributes until the attribute vector stops changing between two steps. Such sequential search methods typically require a user-specified cut-off value. One disadvantage of this approach is that there is no unequivocal rule for determining the cut-off. In addition, sequential search may result in local optima.

The other type of parametric approaches is based on a search of the whole item–attribute vector space. Liu *et al.* (2012) adopted this approach, which needs to search the entire attribute vector space that has 2^K-1 elements for each item. Liu *et al.* (2012) defined a **T**-matrix, which represents the linear dependency between the attribute distribution and the response distribution. The **T**-matrix has 2^K columns, and each row vector of the **T**-matrix contains the probabilities of answering a single item or an item combination correctly by the respondents corresponding to each of the 2^K attribute profiles:

$$\boldsymbol{T} = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ J \\ 1 \cup 2 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 & \cdots & \boldsymbol{\alpha}_{2^K} \end{array} \\ \begin{bmatrix} P_{\boldsymbol{\alpha}_1,1} & P_{\boldsymbol{\alpha}_2,1} & \cdots & P_{\boldsymbol{\alpha}_{2^K},1} \\ P_{\boldsymbol{\alpha}_1,2} & P_{\boldsymbol{\alpha}_2,2} & \cdots & P_{\boldsymbol{\alpha}_{2^K},2} \\ \vdots & \vdots & \vdots & \vdots \\ P_{\boldsymbol{\alpha}_1,J} & P_{\boldsymbol{\alpha}_1,J} & \cdots & P_{\boldsymbol{\alpha}_{2^K},J} \\ P_{\boldsymbol{\alpha}_1,1\cup 2} & P_{\boldsymbol{\alpha}_2,1\cup 2} & \cdots & P_{\boldsymbol{\alpha}_{2^K},1\cup 2} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \end{array}, \tag{2}$$

where the column identifier refers to each of the 2^K AMPs, which can be arranged in any order. The row number refers to a single item or an item combination, that is, $1 \cup 2$ refers to the combination of the first and the second item. $P_{\boldsymbol{\alpha}_1,1}$ denotes the probability of the first item being answered correctly by respondents with AMP $\boldsymbol{\alpha}_1$, and so on. In theory, all the combinations of items should be included in the **T**-matrix, and this is the so-called saturated **T**-matrix (Liu *et al.*, 2012), which is a $(2^J-1) \times 2^K$ matrix.

Apparently the **T**-matrix is a function of the item parameters (s,g) and the **Q**-matrix; in other words, we can denote it by $\mathbf{T}_{s,g}(\mathbf{Q})$. Multiplying the saturated **T**-matrix by the AMP distribution **p**, $\mathbf{T}_{s,g}(\mathbf{Q})\mathbf{p}$ becomes a vector with 2^J-1 entries, representing the expected response probability distribution of each single item or each combination of items given a particular set of parameters (**Q**,s,g,**p**). On the other hand, the corresponding observed response probability distribution can be calculated from the response data, a vector denoted by **β**. With a large sample size and a set of correctly specified parameters (**Q**,s,g,**p**), $\mathbf{T}_{s,g}(\mathbf{Q})\mathbf{p}$ should be close to **β**. Therefore, an objective function can be defined as

$$S_{s,g,\mathbf{p}}(\mathbf{Q}) = |\mathbf{T}_{s,g}(\mathbf{Q})\mathbf{p} - \boldsymbol{\beta}|, \tag{3}$$

where $|\cdot|$ refers to the Euclidean distance. It is expected that $s_{s,g,\mathbf{p}}(\mathbf{Q}) \to 0$ as $N \to \infty$ when all the parameters are correctly specified. Results in Liu *et al.* (2012) indicate that **Q**-matrix validation based on the *S* statistic can achieve a high success rate without requiring any user-specified cut-off. To estimate the attribute vector of the *j*th item, the method fixes

the attribute vector of the remaining J–1 items, surveys each possible attribute vector of the jth item, and then calculates the S statistic. The attribute vector that leads to the smallest S will be chosen. A practical problem is that the saturated **T**-matrix has 2^J–1 rows and 2^K columns, which is computationally infeasible for large J and/or K. Liu *et al.* (2012) suggested only including the 1-way, 2-way, ..., (K + 1)-way combinations of items in ascending order. Even in that case, the number of rows can still be prohibitively large. For example, the **T**-matrix obtained by just including up to the six-way combinations of items has 60,459 rows for a five-attribute, 20-item test. Further exacerbating the computational complexity, the **T**-matrix needs to be updated with the change of attribute vector of each item.

In this paper, we propose a residual-based statistic that belongs to the second type of parametric methods, namely the whole-attribute-vector search. However, our approach does not rely on the **T**-matrix, whose number of rows grows exponentially with the number of items. In addition, our method does not require an update of the **T**-matrix every time the attribute vector of an item is updated. Hence, our method will be computationally much more efficient.

3. Method

3.1. Q-matrix validation based on a residual-based statistic

In this section, we provide the rationale for our proposed approach, using a residual-based statistic to measure the appropriateness of the attribute vectors in the **Q**-matrix. We will further show analytically that under the DINA model, when the AMPs are known, $N \to \infty$ and $s_j, g_j \in (0,0.5)$, one or both of the item parameter estimates will be biased. The bias is expected to inflate the residual statistic. Hence we can use the residual-based statistic to correct the attribute vector when it is misspecified.

Definition 1. $\hat{\mathbf{q}}_j$ *is regarded as correct only when each entry in* $\hat{\mathbf{q}}_j$ *matches the corresponding entry in* $\boldsymbol{q}_j$. $\hat{\mathbf{Q}}$ *is regarded as correct only when the previous condition holds for each item.*

Definition 2. *There are three possible misspecifications for* $\boldsymbol{q}_j$*, which are denoted by* M_1*,* M_2*, and* M_3*.*

M_1: If $\hat{\mathbf{q}}_j$ is misspecified only by lacking some of the required attributes.
M_2: If $\hat{\mathbf{q}}_j$ is misspecified only by specifying some of the non-required attributes as required.
M_3: If $\hat{\mathbf{q}}_j$ is misspecified by lacking some of the required attributes and adding some of the non-required attributes.

M_1 and M_2 are both special cases of M_3. Table 1 provides an example of each of the three types of misspecification. The shaded entries in each **Q**-matrix refer to the misspecified attributes. In Table 1, K = 3.

We next show how the item parameters are expected to change with these misspecifications in a **Q**-matrix. Let us assume that the true AMPs of the respondents are known. Consider a 20-item test that measures three attributes (i.e., J = 20, K = 3). Assume that all AMPs are equally likely, that is, the expected number of students with each AMP is $N/2^K$, or in this example $N/2^3$, where N is the total number of students. For simplicity, all the s_j and g_j are assumed to be 0.2. For $\mathbf{Q}_0^{M_1}$, based on the attribute vector for

Table 1. The true **Q**-matrix and three misspecifications

	Type	Q-matrix
True **Q**-matrix	$\mathbf{Q}$	$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$
Misspecified **Q**-matrix	$\mathbf{Q}_0^{M_1}$	$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$
	$\mathbf{Q}_0^{M_2}$	$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$
	$\mathbf{Q}_0^{M_3}$	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

the second item $\hat{\mathbf{q}}_2$, half of the respondents, whose $\boldsymbol{\alpha}_i$ are $\lfloor 1 \quad 0 \quad 0 \rfloor'$, $\lfloor 1 \quad 1 \quad 0 \rfloor'$, $\lfloor 1 \quad 0 \quad 1 \rfloor'$, or $\lfloor 1 \quad 1 \quad 1 \rfloor'$, should be classified into the mastery group (i.e., $\eta_{i2} = 1$), and the rest should be classified into the non-mastery group (i.e., $\eta_{i2} = 0$), whose AMPs are $\lfloor 0 \quad 0 \quad 0 \rfloor'$, $\lfloor 0 \quad 1 \quad 0 \rfloor'$, $\lfloor 0 \quad 0 \quad 1 \rfloor'$, or $\lfloor 0 \quad 1 \quad 1 \rfloor'$. Taking $\mathbf{Q}_0^{M_1}$ as an example, the maximum likelihood estimate for the slipping parameter of item 2 can be obtained by the proportion of test-takers in the mastery group giving incorrect answers to item 2 (de la Torre, 2009). Let the number of test-takers in the mastery group be N_{m}. As explained above, this encompasses all participants whose $\boldsymbol{\alpha}_i$ are $\lfloor 1 \quad 0 \quad 0 \rfloor'$, $\lfloor 1 \quad 1 \quad 0 \rfloor'$, $\lfloor 1 \quad 0 \quad 1 \rfloor'$, or $\lfloor 1 \quad 1 \quad 1 \rfloor'$, each having N_1, N_2, N_3, and N_4 members, respectively, and $N_{\mathrm{m}} = N_1 + N_2 + N_3 + N_4$. When the AMPs are known, N_1, N_2, N_3, N_4, and N_{m} are all known constants. The numbers of people answering item 2 incorrectly in these groups are denoted by W_1, W_2, W_3, and W_4, respectively. Hence

true master using sliping

fake master fundamentally is guessing

$$\hat{s}_2 = \frac{W_1 + W_2 + W_3 + W_4}{N_1 + N_2 + N_3 + N_4},$$

and

$$E[\hat{s}_2] = \frac{E[W_1 + W_2 + W_3 + W_4]}{N_1 + N_2 + N_3 + N_4} = \frac{E[W_1] + E[W_2] + E[W_3] + E[W_4]}{N_1 + N_2 + N_3 + N_4}.$$

Given the assumption that all AMPs are evenly distributed, $E[W_1] = E[W_2] = .2 \cdot N_{\mathrm{m}}/4$ and $E[W_3] = E[W_4] = .8 \cdot N_{\mathrm{m}}/4$. Therefore,

over estimate the true sliping: 0.2 v.s. 0.5

$$E[\hat{s}_2] = \frac{.2N_{\mathrm{m}}/4 + .2N_{\mathrm{m}}/4 + .8N_{\mathrm{m}}/4 + .8N_{\mathrm{m}}/4}{N_{\mathrm{m}}} = .5.$$

Following the same reasoning, we can get the following results for item 2: (1) for $\mathbf{Q}_0^{M_1}$, $E[\hat{s}_2] = .5, E[\hat{g}_2] = .2$; (2) for $\mathbf{Q}_0^{M_2}$, $E[\hat{s}_2] = .2, E[\hat{g}_2] = 2/7 \approx .286$; $\mathbf{Q}_0^{M_3}$, $E[\hat{s}_2] = .5, E[\hat{g}_2] = .3$.

Consistent with de la Torre (2008), this example shows that $\hat{s}_j$ has positive bias when M_1 occurs, $\hat{g}_j$ has positive bias when M_2 occurs, and both $\hat{s}_j$ and $\hat{g}_j$ have positive bias when

M_3 occurs. Theorem 1 formalizes the relationship between misspecification in the Q matrix and the bias in the resulting item parameter estimates. Based on these considerations, we propose a residual-based statistic, denoted by R_j, which purports to measure the appropriateness of the attribute vector of an item.

Definition 3. *Let $E(X_{ij}|\boldsymbol{\alpha}_i)$ refer to the expected score of the jth item for the ith respondent with AMP $\boldsymbol{\alpha}_i$, and $P(x_{ij}|\boldsymbol{\alpha}_i)$ is the probability of the respondent obtaining score x_{ij}, which is a binary variable with value 0 or 1. Then the appropriateness index R_j for the attribute vector of the jth item can be defined as*

$$R_j = \sum_{i=1}^{N} \log\left[\frac{x_{ij} - E\left(X_{ij}|\boldsymbol{\alpha}_i\right)}{P\left(x_{ij}|\boldsymbol{\alpha}_i\right)}\right]^2.$$

Note that $\frac{x_{ij}-E\left(X_{ij}|\boldsymbol{\alpha}_i\right)}{P\left(x_{ij}|\boldsymbol{\alpha}_i\right)}$ is similar to the standardized residual in the item response theory literature, especially in the context of Rasch models, which takes the form (Masters & Wright, 1997) $\frac{x_{ij}-E\left(X_{ij}|\boldsymbol{\theta}_i\right)}{\sqrt{\mathrm{Var}\left(X_{ij}|\boldsymbol{\theta}_i\right)}}$, where $\mathrm{Var}\left(X_{ij}|\boldsymbol{\theta}_i\right)$ is the variance of X_{ij} given θ_i. Given a dichotomous response, $\sqrt{\mathrm{Var}\left(X_{ij}|\boldsymbol{\theta}_i\right)} = \sqrt{P_{ij}(\boldsymbol{\theta})Q_{ij}(\boldsymbol{\theta})}$, where $P_{ij}(\boldsymbol{\theta}) = P\left(X_{ij}=1|\boldsymbol{\theta}_i\right) = E\left(X_{ij}|\boldsymbol{\theta}_i\right)$. The numerator is the difference between the observed response and the expected response (implied by the model), which captures model misfit. The ratio is a normalized difference. When summed over individuals, the standardized residual can serve as an index of item misfit. If we were to construct a similar standardized residual under the CDM, then it would take the form $\frac{x_{ij}-E\left(X_{ij}|\boldsymbol{\alpha}_i\right)}{\sqrt{\mathrm{Var}\left(X_{ij}|\boldsymbol{\alpha}_i\right)}}$. By using $P\left(x_{ij}|\boldsymbol{\alpha}_i\right)$ as the denominator, when $P\left(x_{ij}|\boldsymbol{\alpha}_i\right)$ is small, the difference between the observed and expected responses is magnified, compared to when $P\left(x_{ij}|\boldsymbol{\alpha}_i\right)$ is large. In doing so, an improbable response leads to a large residual. In this way R_j captures the model–data misfit at the item level. A similar weighted residual was proposed in Yu & Cheng (2019) under the IRT model to detect inattentive response behavior.

Under the DINA model, based on the attribute profile $\boldsymbol{\alpha}_i$, the attribute vector of the jth item $\mathbf{q}_j$ and the response X_{ij}, the ith respondent can be classified into one of the four groups G_1, G_2, G_3 and G_4. Let $N_{11}^j, N_{10}^j, N_{01}^j$ and N_{00}^j be the number of respondents in each of the four groups, respectively. The two numbers in the subscript refer to the values of η and the item response. For example, N_{11}^j indicates the number of respondents in the sample in group G_1, in which the respondents possess all the required attributes of item j and answer it correctly, that is, $\eta_{ij} = 1$ and $X_{ij} = 1$. Note that $N_{11}^j + N_{10}^j = N_{\mathrm{m}}$, and $N_{01}^j + N_{00}^j = N - N_{\mathrm{m}}$.

Under the DINA model, R_j can be expanded to give

$$\begin{aligned} R_j &= \sum_{i=1}^{N} \log\left[\eta_{ij}\left(\frac{s_j}{1-s_j}\right)^{x_{ij}}\left(\frac{1-s_j}{s_j}\right)^{1-x_{ij}} + \left(1-\eta_{ij}\right)\left(\frac{g_j}{1-g_j}\right)^{1-x_{ij}}\left(\frac{1-g_j}{g_j}\right)^{x_{ij}}\right]^2 \\ &= 2\left[N_{11}^j \log\left(\frac{s_j}{1-s_j}\right) + N_{10}^j \log\left(\frac{1-s_j}{s_j}\right) + N_{01}^j \log\left(\frac{1-g_j}{g_j}\right) + N_{00}^j \log\left(\frac{g_j}{1-g_j}\right)\right] \\ &= 2\left[\left(N_{11}^j - N_{10}^j\right)\log\left(\frac{s_j}{1-s_j}\right) + \left(N_{00}^j - N_{01}^j\right)\log\left(\frac{g_j}{1-g_j}\right)\right]. \end{aligned} \tag{5}$$

Definition 4. *The appropriateness index of an operational* **Q***-matrix can be determined as*

$$R = \sum_{j=1}^{J} R_j, \tag{6}$$

or, under the DINA model,

$$R = 2\left[\sum_{j=1}^{J}\left(N_{11}^{j} - N_{10}^{j}\right)\log\left(\frac{s_j}{1-s_j}\right) + \sum_{j=1}^{J}\left(N_{00}^{j} - N_{01}^{j}\right)\log\left(\frac{g_j}{1-g_j}\right)\right]. \tag{7}$$

By summing R_j over all items, R aims to capture the overall model–data misfit. It is also easy to see that the computation of R increases linearly with the test length J, while computational complexity increases exponentially with J for the S method. Note that when the AMPs (i.e., the $\boldsymbol{\alpha}$) and **Q**-matrix are known, N_{m} is a fixed number, and $N_{11}^{j} \sim B(N_{\mathrm{m}}, 1\text{–}s_j)$, $N_{10}^{j} \sim B(N_{\mathrm{m}}, s_j)$, $N_{01}^{j} \sim B(N - N_{\mathrm{m}}, g_j)$, and $N_{00}^{j} \sim B(N - N_{\mathrm{m}}, 1\text{–}g_j)$. If s_j and $g_j \in (0, .5)$, it follows that $E[N_{11}^{j}] > E[N_{10}^{j}]$, and $E[N_{00}^{j}] > E[N_{01}^{j}]$. Given estimates of $\boldsymbol{\alpha}$ and $\mathbf{q}_j$ and the response data from a sample, one can calculate $N_{11}^{j}, N_{10}^{j}, N_{01}^{j}$ and N_{00}^{j}. Then given estimates of s_j and g_j, one can compute $\hat{R}_j$ by plugging these estimates into equation (5). By summing up all the $\hat{R}_j$, one can obtain $\hat{R}$.

Through an illustrative example, we show on page 6 that **Q**-matrix misspecification can lead to positive bias in item parameter estimates. In Appendix B we provide the formal proof of Theorem 1 that this holds true under certain assumptions. Our rationale is that the misspecified **Q**-matrix will lead to (positively) biased item parameter estimates, which in turn will result in compromised model–data fit, as captured by R_j and R. Hence we believe we can approach the true **Q**-matrix by minimizing R_j and R.

3.2. Proposed Q-matrix validation process

Starting with an initial **Q**-matrix $\mathbf{Q_0}$, which is often provided by subject-matter experts in a real test, we propose an iterative process to validate the **Q**-matrix. Denote by $\hat{\mathbf{Q}}^{(t-1)}$ the provisional **Q**-matrix in the $(t-1)$th iteration. Then the tth iteration proceeds with the following steps.

Step 1. For the jth item in $\hat{\mathbf{Q}}^{(t-1)}$, replace its attribute vector $\hat{\boldsymbol{q}}^{(t-1)}$ with $\hat{\boldsymbol{q}}_j^{(t)} = \arg\min_{\mathbf{q}_j^* \in \tilde{\mathbf{q}}} \hat{R}_j\left(\hat{s}_j\left(\mathbf{q}_j^*\right), \hat{g}_j\left(\mathbf{q}_j^*\right), \hat{\alpha}\left(\hat{\mathbf{Q}}^{(t-1)*}\right)\right)$, where $\mathbf{q}_j^*$ is any attribute vector in the item vector space $\tilde{\boldsymbol{q}}$, which contains $2^K - 1$ possible attribute vectors. Note that an attribute vector containing all 0s is not considered a valid attribute vector because that item then is completely irrelevant to the assessment. $\hat{\boldsymbol{\alpha}}\left(\mathbf{Q}^{(t-1)*}\right)$ is the estimated AMPs based on **X** and $\hat{\mathbf{Q}}^{(t-1)*}$. Replace the attribute vector of the jth item by $\mathbf{q}_j^*$ in $\mathbf{Q}^{(t-1)}$, then $\hat{\mathbf{Q}}^{(t-1)}$ becomes $\hat{\mathbf{Q}}^{(t-1)*}$.

Step 2. Repeat Step 1 for all J items. The resulting **Q**-matrix is $\hat{\mathbf{Q}}^{(t)}$.

Step 3. Repeat Steps 1 and 2 until $\hat{\mathbf{Q}}^{(t-1)} = \hat{\mathbf{Q}}^{(t)}$.

Conceptually, the difference between Liu *et al.*'s (2012) method and our proposed method lies in Step 1: how $\hat{\boldsymbol{q}}_j^{(t-1)}$ is updated. In Step 1 of each iteration of the R method

(our proposed method) and the *S* method (Liu *et al.*, 2012), one computes $2^K - 1$ instances of R_j or S_j over the $2^K - 1$ possible elements in the whole-attribute-vector space $\tilde{\mathbf{q}}$, and finds the attribute vector that minimizes the statistic. This is repeated for all J items to update the entire **Q**-matrix. Computationally, however, the *S* method needs to update the large **T**-matrix (which contains 2^J–1 rows if saturated) in each evaluation. It is therefore much more computationally intensive than the *R* method.

In order to evaluate the performance of the proposed statistic with the estimated AMPs and realistic sample sizes, three simulation studies are conducted. The preliminary simulation study explores the relationship between attribute number, test length, and respondents' classification accuracy, and is a preliminary study for the second study. The primary simulation I evaluates the performance of the *R* method and compares it against that of the *S* method, as both methods use the whole vector search algorithm, and neither requires an arbitrarily pre-specified cut-off. The primary simulation II assesses the performance of the two methods for unevenly distributed population; everything else was kept the same as in the second study except for the distribution of the AMPs.

4. Simulation study

The AMPs, if not estimated accurately, will make it difficult to recover the item parameters and the **Q**-matrix. Hence, the preliminary study examines the relationship between the number of attributes, test length, and classification accuracy of respondents. This study examines what test length is appropriate for the number of attributes being measured in the test. If the AMP and item parameter estimation are poor given a correctly specified **Q**-matrix at a given test length, poorer performances are expected when the **Q**-matrix is misspecified. The appropriate test length refers to a test length that affords high estimation accuracy of AMPs and item parameters, both critical to the identification of **Q**-matrix misspecification. Hence, in the preliminary simulation, we assume that the **Q**-matrices are correctly specified, which will allow us to find the minimum test length for a specific number of attributes.

4.1. Preliminary simulation: relationships between number of attributes, test length and AMP estimation

In the preliminary simulation we consider four levels for attribute number (3, 4, 5, and 6); two levels for the test lengths (20 items for 3-, 4-, 5-, and 6-attribute tests, and 30 items for 5- and 6-attribute tests); and four levels for the sample size (800, 1,000, 2,000, and 4,000). For the 20-item test, the same **Q**-matrices as in Liu *et al.* (2012) are used. For the 30-item test with five attributes, the same **Q**-matrix as in de la Torre and Chiu (2016) is adopted. Taking the identifiability of the CDM into account (Xu & Zhang, 2016), we develop two **Q**-matrices specifically for the 20-item and 30-item test with six attributes, which have similar structure to those in Liu *et al.* (2012) and de la Torre and Chiu (2016). There are in total six **Q**-matrices, all presented in the Appendix A. Item parameters are drawn from U (0.05, 0.25), and the AMPs are randomly sampled from the 2^K possible attribute patterns. AMPs are estimated using maximum likelihood (Cheng, 2009; Huebner & Wang, 2011).

Based on the DINA model, the probability of success on item J for respondents possessing all the required attributes is $1–s_j$, and for respondents lacking at least one of the required attributes it is g_j. To simulate the score (1 or 0) for each respondent, a random number u from the uniform distribution on the interval [0, 1] is drawn; when $u < 1–s_j$, the response to the item is set to 1, and 0 otherwise. There are a total of 4 (numbers of

attributes) × 1 (20-item test) × 4 (sample sizes) + 2 (number of attributes) × 1 (30-item test) × 4 (sample sizes) = 24 conditions. For each condition, 100 replications are conducted. For each simulation condition, the following evaluation criteria are applied.

Root mean squared error (RMSE). The RMSE is used to evaluate the estimation accuracy of the item parameters (Chen, Liu, Xu, & Ying, 2015; Chen, Xin, Wang, & Chang, 2012), which are defined as

$$g_{\text{RMSE}} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(g_j - \hat{g}_j\right)^2}, \tag{8}$$

$$s_{\text{RMSE}} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(s_j - \hat{s}_j\right)^2} \tag{9}$$

The RMSE criterion reflects the average magnitude of the bias between the true item parameters and their associated estimates. A smaller RMSE suggests higher estimation accuracy. Average RMSE values across 100 replications will be reported.

Pattern correct classification rate (PCCR). This index quantifies the estimation accuracy of the respondents' AMPs (Chen *et al.*, 2012):

$$\text{PCCR} = \frac{\sum_{i=1}^{N} \text{I}(\boldsymbol{\alpha}_i == \hat{\boldsymbol{\alpha}}_i)}{N}, \tag{10}$$

where $\text{I}(\boldsymbol{\alpha}_i == \hat{\boldsymbol{\alpha}}_i)$ equals 1 if the estimate profile $\hat{\boldsymbol{\alpha}}_i$ matches its true value $\boldsymbol{\alpha}_i$, and 0 otherwise. Average PCCR values across replications will be reported. A higher PCCR is desirable and indicates better estimation of the AMPs.

Attribute correct classification rate (ACCR). Different from the PCCR, which quantifies the accuracy at the attribute profile level, the ACCR quantifies the average estimation accuracy at the attribute level (Chen *et al.*, 2012).

$$\text{ACCR} = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} \text{I}(\alpha_{ik} == \hat{\alpha}_{ik})}{NK}, \tag{11}$$

where $\text{I}(\alpha_{ik} == \hat{\alpha}_{ik})$ is an indicator function which equals 1 if the estimate attribute $\hat{\alpha}_{ik}$ equates to its true value α_{ik}, and 0 otherwise. The average ACCR across replications will be reported. A higher ACCR is desirable.

Table 2 presents the parameter estimation accuracy for both item and respondents. According to Table 2, for the three-attribute test, 20 items are sufficient to obtain a high PCCR, in mid to upper .90s. For the four-attribute tests, the PCCR ranged between the upper .70s and the mid .80s. With the increase in sample size, the estimation accuracy for the item parameters shows visible improvement.

Overall, Table 2 indicates that when the **Q**-matrices are correctly specified: (1) when the sample size is 1,000 or larger, 20 items seem to be sufficient for a five-attribute test to achieve a PCCR of .80 or above; (2) a 20-item test seems too short for a six-attribute test; and (3) at least 30 items are needed for a six-attribute test to achieve a PCCR close to .80. These findings will inform the design of the main simulation study.

Table 2. Classification accuracy and item parameter accuracy

Test length	Attributes	Sample	PCCR	ACCR	s_{RMSE}	g_{RMSE}
20	3	800	.914	.968	.022	.018
		1,000	.937	.977	.020	.015
		2,000	.939	.978	.015	.011
		4,000	.939	.978	.011	.007
	4	800	.779	.933	.030	.017
		1,000	.792	.937	.027	.016
		2,000	.812	.945	.020	.012
		4,000	.843	.954	.013	.008
	5	800	.773	.946	.029	.019
		1,000	.802	.954	.023	.016
		2,000	.809	.955	.017	.011
		4,000	.826	.959	.012	.008
	6	800	.568	.904	.037	.022
		1,000	.580	.907	.035	.022
		2,000	.606	.914	.022	.014
		4,000	.627	.920	.017	.010
30	5	800	.803	.950	.028	.016
		1,000	.830	.958	.025	.015
		2,000	.834	.959	.018	.010
		4,000	.849	.963	.013	.007
	6	800	.747	.947	.028	.018
		1,000	.772	.952	.024	.016
		2,000	.777	.954	.017	.010
		4,000	.779	.955	.012	.008

4.2. Primary simulation I: Evaluating the performance of the proposed validation method under evenly distributed AMPs

According to the results of the preliminary simulation, even when the **Q**-matrix is correctly specified, 20 items are required in order to achieve adequate estimation accuracy of item parameters and respondent AMPs for the three-attribute and four-attribute test, and 30 items are needed for the five-attribute and six-attribute test. Hence these combinations of test length and number of attributes will be used in primary simulation I.

As explained above, for the three-attribute and four-attribute test, matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ will be used as the true matrices. In order to evaluate the performances for the two methods under different misspecification severities, 3–10 misspecified items are considered. Because the wrong attribute vector for a misspecified item cannot be the zero vector or the true vector, we generate misspecified attribute vectors by randomly sampling from the remaining 2^K–2 possible vectors. In total, there are 2 (number of attributes: 3 or 4) × 1 (20-item test) × 4 (sample sizes: 800, 1,000, 2,000, or 4,000) × 8 (number of misspecified items: 3, 4, . . ., 9, 10) + 2 (number of attributes: 5 or 6) ×1 (30-item test) × 4 (sample sizes as above) × 8 (number of misspecified items as above) = 128 conditions. For each condition, 100 replications are conducted. Based on the data set and the initial **Q**-matrix with some misspecifications, one of the two methods, the *S* method or the *R* method, is used to validate the **Q**-matrix.

For the sake of fair comparison, we apply both *R* and *S* methods to the same data set in each replication, and then calculate the corresponding criteria for the two methods, respectively. In order to provide a comprehensive comparison between the two methods, seven evaluation indices are considered. The first five indices are indicators of the accuracy of **Q**-matrix validation process, while the last two indices are indicators of computational efficiency.

Q-*matrix recovery rate (QRR)*. The QRR refers to the rate of successfully recovering the true **Q**-matrix in all replications. A high QRR indicates that the associated validation method has a high probability of recovering the true **Q**-matrix. It is defined as

$$\text{QRR} = \frac{\sum_{r=1}^{Rep} \text{I}(\mathbf{Q} == \hat{\mathbf{Q}})}{Rep}, \tag{12}$$

where $\text{I}(\mathbf{Q} == \hat{\mathbf{Q}})$ equals 1 if each attribute vector in the estimated **Q**-matrix, $\hat{\mathbf{Q}}$, matches its true value in **Q**, and 0 otherwise. *Rep* refers to the number of the replication for each condition. *Rep* = 100 in our study. Note that QRR $\in$ [0,1].

Number of recovered items (NRI). In contrast to the QRR, which is defined at the **Q**-matrix level, the NRI is defined at the item level, and refers to the number of items whose attribute vectors are correctly recovered:

$$\text{NRI} = \frac{\sum_{r=1}^{Rep} \sum_{j=1}^{J} \text{I}(\mathbf{q}_j == \hat{\mathbf{q}}_j)}{Rep}, \tag{13}$$

where $\text{I}(\mathbf{q}_j == \hat{\mathbf{q}}_j)$ equals 1 if the estimated attribute vector for the *j*th item, $\hat{\mathbf{q}}_j$, matches its true value $\mathbf{q}_j$, and 0 otherwise. If the estimated **Q**-matrix $\hat{\mathbf{Q}}$ cannot match **Q**, a method that leads to a higher NRI is considered better. Note that NRI $\in [0, J]$: for the 20-item test, the range of *NRI* is 0–20; and for the 30-item test, the range is 0–30.

Number of recovered attributes (NRA). The NRA represents the number of recovered attributes, which is defined as the attribute level. It is computed for each replication after validation is finished, and is given by

$$\text{NRA} = \frac{\sum_{r=1}^{Rep} \sum_{j=1}^{J} \sum_{k=1}^{K} \text{I}(q_{jk} == \hat{q}_{jk})}{Rep}. \tag{14}$$

Note that NRA $\in [0, JK]$: for example, for the 20-item, three-attribute test, the range of *NRA* is therefore 0–60. The following two indices consider the tendency to attribute recovery. Overestimation of an attribute refers to the case where its estimated value is 1, but its true value is 0, and underestimation vice versa.

Number of overestimated attributes (NOA). The NOA reflects the number of attributes that are overestimated (i.e., 0 in an attribute vector being misspecified as 1). As shown in Appendix B, an item with overestimated attributes tends to lead to upward-biased guessing parameter estimates under the DINA model. It is therefore important to examine the NOA, which is given by

$$\text{NOA} = \frac{\sum_{r=1}^{Rep} \sum_{j=1}^{J} \sum_{k=1}^{K} \text{I}(\hat{q}_{jk} > q_{jk})}{Rep}, \tag{15}$$

where $\text{I}(\hat{q}_{jk} > q_{jk})$ equals 1 if $\hat{q}_{jk} = 1$ and $q_{jk} = 0$.

Number of underestimated attributes (NUA). The NUA reflects the number of attributes that are underestimated (i.e., 1 in an attribute vector being misspecified as 0). An item with underestimated attributes tends to have a higher slipping parameter under the DINA model. Hence it is also important to report the NUA, which is given by

$$\text{NUA} = \frac{\sum_{r=1}^{Rep}\sum_{j=1}^{J}\sum_{k=1}^{K} \text{I}(\hat{q}_{jk} < q_{jk})}{Rep}, \tag{16}$$

where $\text{I}(\hat{q}_{jk} > q_{jk})$ equals 1 if $\hat{q}_{jk} = 0$ and $q_{jk} = 1$. Note the relationship NRA + NOA+NUA = JK.

Average number of iterations (ANI). The R and S methods are both iterative. A smaller number of iterations is generally preferable, as it suggests faster convergence of the validation process. Hence the average number of iterations is evaluated:

$$\text{ANI} = \frac{\sum_{r=1}^{Rep} \text{IT}_r}{Rep}, \tag{17}$$

where IT_r refers to the number of iterations for the rth replication.

Average running time (ART). Similar to the ANI, the running times of both methods for validating the operational **Q**-matrix are recorded. The ART is also an index of computational efficiency, given by

$$\text{ART} = \frac{\sum_{r=1}^{Rep} t_r}{Rep}, \tag{18}$$

where t_r refers to the total running time for the rth replication, and ART refers to the average running time required for each condition.

Results for **Q***-matrix recovery*. Table 3 presents the estimation accuracy of $\mathbf{Q}_1$ under different conditions. The range of values for each criterion is provided in the second row of Tables 3–6, and numbers in bold denote better performance in the associated criterion for the corresponding method. At the matrix level, the proposed R method outperforms the S method in terms of QRR. For example, the QRRs of the R method for $\mathbf{Q}_1$ with 10 misspecified items under sample size 800, 1,000, 2,000, and 4,000 are .87, .87, .90, and 0.92, respectively; they are respectively 15, 16, 21, and 22 percentage points better than the corresponding QRRs of the S method. This suggests that the proposed R method leads to a higher recovery rate than the S method, at the entire matrix level. Considering recovery at the item and attribute level, the performance for the R method is also better than that of the S method. The R method leads to higher NRI values than the S method, suggesting that more item attribute vectors are successfully recovered by the former. Using the R method, fewer than two items are estimated inaccurately in the worst case, whereas using S method up to four items can still have misspecified attribute vectors in the end. The R method also results in higher NRA, that is, a higher number of attributes that are successfully recovered. For the NRA, about 3 attributes are misspecified by the R method and six attributes by the S method in the worst case. The S method leads to visibly higher NOA than NUA, suggesting that it tends to overestimate the number of required attributes. Overall Table 3 indicates that the R method leads to higher recovery rate at the matrix, item, and attribute levels than the S method, given a 20-item, three-attribute test.

Table 3. **Q**-matrix estimation accuracy for three attributes, 20-item test ($\mathbf{Q}_1$)

		QRR		NRI		NRA		NOA		NUA	
		[0,1]		[0,20]		[0,60]		[0,60]		[0,60]	
N		*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*
800	3	**1.00**	.83	**20.00**	19.40	**60.00**	59.39	**0.00**	0.52	**0.00**	0.09
	4	**1.00**	.77	**20.00**	19.21	**60.00**	59.01	**0.00**	0.65	**0.00**	0.34
	5	**.99**	.83	**19.99**	19.34	**59.99**	59.26	**0.01**	0.55	**0.00**	0.19
	6	**.99**	.70	**19.98**	18.46	**59.98**	58.01	**0.01**	1.35	**0.01**	0.64
	7	**1.00**	.74	**20.00**	18.55	**60.00**	58.14	**0.00**	1.31	**0.00**	0.55
	8	**.93**	.67	**19.32**	17.52	**58.81**	56.78	**0.63**	2.32	**0.56**	0.90
	9	**.98**	.57	**19.83**	16.83	**59.75**	55.64	**0.13**	2.75	**0.12**	1.61
	10	**.87**	.62	**18.40**	16.41	**57.32**	54.51	**1.49**	3.56	**1.19**	1.93
1,000	3	**1.00**	.87	**20.00**	19.60	**60.00**	59.59	**0.00**	0.30	**0.00**	0.11
	4	**1.00**	.91	**20.00**	19.79	**60.00**	59.78	**0.00**	0.17	**0.00**	0.05
	5	**1.00**	.82	**20.00**	19.35	**60.00**	59.31	**0.00**	0.50	**0.00**	0.19
	6	**1.00**	.88	**20.00**	19.25	**60.00**	59.12	**0.00**	0.65	**0.00**	0.23
	7	**1.00**	.76	**20.00**	18.53	**60.00**	57.93	**0.00**	1.36	**0.00**	0.71
	8	**.95**	.80	**19.40**	18.61	**59.09**	57.90	**0.49**	1.32	**0.42**	0.78
	9	**.98**	.67	**19.83**	17.36	**59.69**	56.29	**0.17**	2.62	**0.14**	1.09
	10	**.87**	.71	**18.33**	16.87	**56.93**	54.53	**1.67**	3.11	**1.40**	2.36
2,000	3	**1.00**	.99	**20.00**	19.95	**60.00**	59.95	**0.00**	0.04	**0.00**	0.01
	4	**1.00**	.96	**20.00**	19.80	**60.00**	59.76	**0.00**	0.16	**0.00**	0.08
	5	**1.00**	.95	**20.00**	19.64	**60.00**	59.49	**0.00**	0.31	**0.00**	0.20
	6	**1.00**	.91	**20.00**	19.55	**60.00**	59.54	**0.00**	0.34	**0.00**	0.12
	7	**1.00**	.93	**20.00**	19.20	**60.00**	58.65	**0.00**	0.81	**0.00**	0.54
	8	**.98**	.90	**19.76**	18.88	**59.52**	58.16	**0.24**	1.08	**0.24**	0.76
	9	**1.00**	.81	**20.00**	18.26	**60.00**	57.38	**0.00**	1.67	**0.00**	0.95
	10	**.90**	.69	**18.86**	16.30	**57.93**	53.62	**1.07**	3.72	**1.00**	2.66
4,000	3	**1.00**	1.00	**20.00**	20.00	**60.00**	60.00	**0.00**	0.00	**0.00**	0.00
	4	**1.00**	.99	**20.00**	19.95	**60.00**	59.95	**0.00**	0.04	**0.00**	0.01
	5	**1.00**	.96	**20.00**	19.64	**60.00**	59.40	**0.00**	0.34	**0.00**	0.26
	6	**1.00**	.95	**20.00**	19.53	**60.00**	59.35	**0.00**	0.42	**0.00**	0.23
	7	**1.00**	.93	**20.00**	19.20	**60.00**	58.65	**0.00**	0.75	**0.00**	0.60
	8	**.98**	.89	**19.77**	18.70	**59.61**	57.65	**0.21**	1.31	**0.18**	1.04
	9	**1.00**	.89	**20.00**	18.79	**60.00**	57.89	**0.00**	1.13	**0.00**	0.98
	10	**.92**	.70	**19.08**	16.43	**58.45**	53.41	**0.80**	3.64	**0.75**	2.95

Note. Numbers in bold indicate better performance in the associated criterion for the corresponding method. For the criteria QRR, NRI, and NRA, a higher value denotes better performance, while for NOA and NUA, a lower value denotes better performance. The second row shows the range for each index.

The comparison between the *R* and *S* methods is of key interest to us. Meanwhile, there are some other noticeable trends as well. First, in most cases, the QRR criterion increases with sample size. But this rule does not always hold; for example, when the number of the misspecified items reaches 10, some exceptions occur. At a sample size of 1,000 the QRR attained by the *S* method is 71%, which is higher than the 69% achieved for a sample size of 2000. This type of exception indicates that different combinations of the large number of misspecified items can have differential effects on **Q**-matrix estimation. Second, it is expected that higher numbers of misspecified items result in lower recovery rate.

Table 4. **Q**-matrix estimation accuracy for four attributes, 20-item test ($\mathbf{Q}_2$)

		QRR		NRI		NRA		NOA		NUA	
		[0,1]		[0,20]		[0,80]		[0,80]		[0,80]	
N		*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*
800	3	**.91**	.40	**19.77**	18.06	**79.61**	77.83	**0.24**	1.84	**0.15**	0.33
	4	**.87**	.31	**19.60**	17.22	**79.47**	76.62	**0.19**	2.66	**0.34**	0.72
	5	**.84**	.29	**19.49**	16.89	**79.28**	75.74	**0.35**	3.31	**0.37**	0.95
	6	**.8**	.41	**19.02**	16.89	**78.56**	75.67	**0.78**	3.43	**0.66**	0.90
	7	**.8**	.30	**18.83**	14.82	**77.91**	72.40	**1.16**	5.56	**0.93**	2.04
	8	**.72**	.30	**18.04**	14.24	**76.62**	70.97	**1.76**	6.24	**1.62**	2.79
	9	**.68**	.29	**16.92**	13.85	**74.45**	70.10	**3.00**	6.87	**2.55**	3.03
	10	**.55**	.23	**15.17**	11.52	**71.24**	65.04	**4.58**	9.80	**4.18**	5.16
1,000	3	**.97**	.64	**19.97**	18.92	**79.97**	78.59	**0.01**	1.13	**0.02**	0.28
	4	**.98**	.49	**19.95**	18.23	**79.90**	77.77	**0.04**	1.72	**0.06**	0.51
	5	**.90**	.49	**19.59**	17.95	**79.47**	77.43	**0.25**	1.95	**0.28**	0.62
	6	**.94**	.51	**19.74**	17.72	**79.60**	76.96	**0.25**	2.31	**0.15**	0.73
	7	**.90**	.41	**19.56**	15.83	**79.21**	73.66	**0.45**	4.41	**0.34**	1.93
	8	**.85**	.35	**18.54**	15.17	**77.36**	72.55	**1.34**	5.35	**1.30**	2.10
	9	**.84**	.31	**18.48**	13.36	**77.40**	68.94	**1.37**	7.42	**1.23**	3.64
	10	**.64**	.27	**16.04**	12.51	**72.57**	67.09	**3.96**	8.48	**3.47**	4.43
2,000	3	**.99**	.79	**19.99**	19.20	**79.99**	78.99	**0.00**	0.75	**0.01**	0.26
	4	**1.00**	.8	**20.00**	19.08	**80.00**	78.76	**0.00**	0.96	**0.00**	0.28
	5	**1.00**	.87	**20.00**	19.56	**80.00**	79.41	**0.00**	0.48	**0.00**	0.11
	6	**1.00**	.76	**20.00**	18.21	**80.00**	77.16	**0.00**	1.82	**0.00**	1.02
	7	**.99**	.58	**19.92**	16.80	**79.91**	74.85	**0.04**	3.37	**0.05**	1.78
	8	**.91**	.54	**19.14**	16.29	**78.46**	74.00	**0.87**	3.78	**0.67**	2.22
	9	**.89**	.44	**18.80**	14.49	**77.89**	70.34	**1.10**	6.22	**1.01**	3.44
	10	**.81**	.45	**17.79**	13.93	**75.81**	69.21	**2.11**	6.71	**2.08**	4.08
4,000	3	**1.00**	.92	**20.00**	19.58	**80.00**	79.42	**0.00**	0.47	**0.00**	0.11
	4	**1.00**	.97	**20.00**	19.81	**80.00**	79.64	**0.00**	0.22	**0.00**	0.14
	5	**1.00**	.92	**20.00**	19.40	**80.00**	78.98	**0.00**	0.65	**0.00**	0.37
	6	**1.00**	.92	**20.00**	19.45	**80.00**	79.15	**0.00**	0.53	**0.00**	0.32
	7	**.99**	.73	**19.92**	17.17	**79.90**	75.20	**0.06**	3.00	**0.04**	1.80
	8	**.95**	.77	**19.44**	17.79	**79.02**	75.97	**0.53**	2.33	**0.45**	1.70
	9	**.86**	.56	**18.59**	14.75	**77.34**	70.30	**1.31**	5.85	**1.35**	3.85
	10	**.81**	.54	**17.76**	14.54	**75.85**	69.97	**2.12**	6.16	**2.03**	3.87

Note. Numbers in bold indicate better performance in the associated criterion for the corresponding method.

However, this is not always the case, further indicating that different combinations of misspecified items could have a complex effect on **Q**-matrix recovery. This is understandable, because certain misspecifications may cause identifiability issues for the AMPs.

Table 4 presents the estimation accuracy of $\mathbf{Q}_2$. In terms of our main research question, the same pattern as in Table 3 is observed – the *R* method leads to a higher recovery rate of the true **Q**-matrix at the matrix, item, and attribute levels. In addition, compared to Table 3, the QRR, NRI, and NRA are lower under the same sample size and number of misspecified items. This is consistent with our expectations, as it gets harder to recover the true **Q**-matrix as the number of attributes increases. In addition, the advantage of the *R*

Table 5. **Q**-matrix estimation accuracy for five attributes, 30-item test ($\mathbf{Q}_5$)

		QRR		*NRI*		*NRA*		*NOA*		*NUA*	
		[0,1]		[0,30]		[0,150]		[0,150]		[0,150]	
N		*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*
800	3	**.91**	.53	**29.85**	28.92	**149.80**	148.00	**0.15**	1.71	**0.05**	0.29
	4	**.87**	.43	**29.78**	28.81	**149.76**	147.77	**0.18**	1.95	**0.06**	0.28
	5	**.76**	.51	**29.56**	29.05	**149.42**	148.21	**0.37**	1.53	**0.21**	0.26
	6	**.80**	.55	**29.68**	28.88	**149.46**	147.92	**0.38**	1.72	**0.16**	0.36
	7	**.78**	.42	**29.63**	28.91	**149.46**	147.70	**0.44**	1.96	**0.10**	0.34
	8	**.88**	.47	**29.84**	28.90	**149.80**	147.09	**0.20**	2.25	**0.00**	0.66
	9	**.70**	.46	**29.34**	29.00	**149.15**	147.18	**0.67**	2.38	0.18	0.44
	10	**.80**	.58	**29.45**	28.86	**149.24**	146.27	**0.54**	3.12	**0.22**	0.61
1,000	3	**.92**	.64	**29.90**	29.29	**149.87**	148.96	**0.12**	0.93	**0.01**	0.11
	4	**.95**	.70	**29.95**	29.39	**149.95**	148.41	**0.05**	1.40	**0.00**	0.19
	5	**.90**	.67	**29.82**	29.33	**149.77**	148.97	**0.15**	0.89	**0.08**	0.14
	6	**.89**	.62	**29.84**	29.29	**149.81**	148.58	**0.17**	1.16	**0.02**	0.26
	7	**.91**	.60	**29.87**	29.06	**149.83**	148.11	**0.15**	1.50	**0.02**	0.39
	8	**.91**	.65	**29.88**	29.40	**149.87**	148.55	**0.11**	1.21	**0.02**	0.24
	9	**.90**	.63	**29.75**	28.99	**149.69**	148.02	**0.21**	1.65	**0.10**	0.33
	10	**.92**	.58	**29.86**	29.22	**149.80**	146.82	**0.16**	2.37	**0.04**	0.81
2,000	3	**1.00**	.88	**30.00**	29.79	**150.00**	149.69	**0.00**	0.30	**0.00**	0.01
	4	**1.00**	.86	**30.00**	29.77	**150.00**	149.75	**0.00**	0.24	**0.00**	0.01
	5	**1.00**	.94	**30.00**	29.94	**150.00**	149.77	**0.00**	0.21	**0.00**	0.02
	6	**1.00**	.94	**30.00**	29.94	**150.00**	149.62	**0.00**	0.31	**0.00**	0.07
	7	**1.00**	.87	**30.00**	29.74	**150.00**	149.25	**0.00**	0.63	**0.00**	0.12
	8	**1.00**	.94	**30.00**	29.93	**150.00**	149.52	**0.00**	0.46	**0.00**	0.02
	9	**1.00**	.84	**30.00**	29.75	**150.00**	149.14	**0.00**	0.64	**0.00**	0.22
	10	**1.00**	.85	**30.00**	29.71	**150.00**	148.23	**0.00**	1.36	**0.00**	0.41
4,000	3	**1.00**	.98	**30.00**	29.97	**150.00**	149.94	**0.00**	0.06	**0.00**	0.00
	4	**1.00**	.99	**30.00**	29.99	**150.00**	149.95	**0.00**	0.05	**0.00**	0.00
	5	**1.00**	.99	**30.00**	29.99	**150.00**	149.95	**0.00**	0.05	**0.00**	0.00
	6	**1.00**	1.00	**30.00**	30.00	**150.00**	149.67	**0.00**	0.17	**0.00**	0.16
	7	**1.00**	.99	**30.00**	29.99	**150.00**	149.97	**0.00**	0.03	**0.00**	0.00
	8	**1.00**	1.00	**30.00**	30.00	**150.00**	149.94	**0.00**	0.03	**0.00**	0.03
	9	**1.00**	1.00	**30.00**	30.00	**150.00**	149.97	**0.00**	0.03	**0.00**	0.00
	10	**1.00**	.97	**30.00**	29.97	**150.00**	149.27	**0.00**	0.49	**0.00**	0.24

Note. Numbers in bold indicate better performance in the associated criterion for the corresponding method.

method over the *S* method is more pronounced in Table 4. Recall that in Table 3, the QRRs of the *R* method with 10 misspecified items and sample sizes of 800, 1,000, 2,000, and 4,000 are 15, 16, 21, and 22 percentage points higher than the corresponding QRRs of the *S* method. With the same number of misspecified items and sample sizes, Table 4 shows that the *R* method yields QRRs that are 22, 37, 36, and 27 percentage points higher than the corresponding QRRs of the *S* method. This suggests that the *R* method can be more advantageous as the number of attributes increases.

Tables 5 and 6 present the results for the 30-item tests, with true matrices $\mathbf{Q}_5$ (five attributes) and $\mathbf{Q}_6$ (six attributes), respectively. Regarding our key question, the *R* method

Table 6. **Q**-matrix estimation accuracy for six attributes, 30-item test ($\mathbf{Q}_6$)

		QRR		*NRI*		*NRA*		*NOA*		*NUA*	
		[0,1]		[0,30]		[0,180]		[0,180]		[0,180]	
N		*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*
800	3	**.78**	.46	**29.56**	28.10	**179.30**	178.70	**0.63**	1.11	**0.07**	0.19
	4	**.81**	.39	**29.74**	27.91	**179.46**	178.70	**0.48**	1.17	**0.06**	0.13
	5	**.81**	.44	**29.70**	28.38	**179.57**	178.88	**0.40**	0.98	**0.03**	0.14
	6	**.73**	.38	**29.39**	28.21	**178.77**	178.52	**1.05**	1.19	**0.18**	0.29
	7	**.75**	.41	**29.29**	27.99	**178.86**	178.65	**1.03**	1.11	**0.11**	0.24
	8	**.80**	.46	**29.05**	27.58	**178.89**	178.14	**1.03**	1.50	**0.08**	0.36
	9	**.68**	.38	**28.10**	27.52	**178.83**	176.24	**1.07**	2.85	**0.10**	0.91
	10	**.63**	.29	**27.50**	26.79	**178.52**	175.30	**1.23**	3.66	**0.25**	1.04
1,000	3	**.87**	.59	**29.84**	29.04	**179.66**	179.21	**0.31**	0.70	**0.03**	0.09
	4	**.96**	.53	**29.91**	28.55	**179.85**	179.27	**0.13**	0.67	**0.02**	0.06
	5	**.91**	.56	**29.77**	29.03	**179.64**	179.29	**0.29**	0.68	**0.07**	0.03
	6	**.89**	.51	**29.77**	28.65	**179.69**	179.15	**0.29**	0.79	**0.02**	0.06
	7	**.88**	.53	**29.60**	28.31	**179.09**	179.03	**0.55**	0.86	**0.36**	0.11
	8	**.86**	.53	**29.17**	28.69	**179.43**	178.16	**0.54**	1.46	**0.03**	0.38
	9	**.82**	.50	**29.35**	28.22	**179.00**	178.46	**0.93**	1.15	**0.07**	0.39
	10	**.77**	.52	**28.14**	27.32	**179.15**	176.44	**0.75**	2.56	**0.10**	1.00
2,000	3	**1.00**	.87	**30.00**	29.69	**180.00**	179.79	**0.00**	0.21	**0.00**	0.00
	4	**1.00**	.88	**30.00**	29.77	**180.00**	179.77	**0.00**	0.22	**0.00**	0.01
	5	**1.00**	.87	**30.00**	29.77	**180.00**	179.94	**0.00**	0.06	**0.00**	0.00
	6	**.98**	.86	**29.88**	29.66	**179.94**	179.76	**0.06**	0.20	**0.00**	0.04
	7	**.98**	.76	**29.78**	29.32	**179.70**	179.50	**0.30**	0.30	**0.00**	0.20
	8	**1.00**	.78	**30.00**	29.53	**180.00**	179.93	**0.00**	0.07	**0.00**	0.00
	9	**.98**	.74	**29.84**	29.21	**179.74**	179.70	**0.26**	0.24	**0.00**	0.06
	10	**.93**	.78	**29.33**	28.58	**179.71**	178.62	**0.29**	0.99	**0.00**	0.39
4,000	3	**1.00**	.99	**30.00**	29.94	**180.00**	179.97	**0.00**	0.03	**0.00**	0.00
	4	**1.00**	.99	**30.00**	29.95	**180.00**	179.99	**0.00**	0.01	**0.00**	0.00
	5	**1.00**	.96	**30.00**	29.95	**180.00**	179.99	**0.00**	0.01	**0.00**	0.00
	6	**1.00**	.98	**30.00**	29.83	**180.00**	180.00	**0.00**	0.00	**0.00**	0.00
	7	**.99**	.98	**29.97**	29.81	**179.99**	179.56	**0.01**	0.32	**0.00**	0.12
	8	**1.00**	.99	**30.00**	29.94	**180.00**	180.00	**0.00**	0.00	**0.00**	0.00
	9	**1.00**	.97	**30.00**	29.97	**180.00**	180.00	**0.00**	0.00	**0.00**	0.00
	10	**1.00**	.95	**30.00**	29.47	**180.00**	179.47	**0.00**	0.53	**0.00**	0.00

Note. Numbers in bold indicate better performance in the associated criterion for the corresponding method.

still outperforms the *S* method in Q-matrix recovery at the matrix, item, and attribute levels. There are also some other interesting trends. With a 30-item test, the recovery rates converge to perfect recovery at a sample size of 2,000 or above. This did not happen with the 20-item test, even when the number of attributes was small. Further, at a sample size of 1,000, the *R* method can already achieve a QRR of .9 or above. For the *S* method, it takes a sample size of 2,000 to achieve the same QRR level. At smaller sample sizes, the advantage of the *R* method over the *S* method is more pronounced, gradually diminishing as the sample size increases. The trends regarding NRI, NRA, NOA, and NUA are very similar to those observed with 20-item tests.

Table 7. Q-matrix estimation efficiency

		3				4				5				6			
		ANI		ART		ANI		ART		ANI		ART		ANI		ART	
N		*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*
800	3	**2.00**	2.54	**155.13**	326.15	**2.05**	3.00	**535.86**	3,781.03	**2.01**	2.55	**1,357.30**	80,892.36	**2.07**	2.46	**3,995.75**	228,126.40
	4	**2.00**	2.63	**180.89**	415.30	**2.17**	3.22	**540.29**	4,098.32	**2.04**	2.6	**1,337.53**	79,304.92	**2.05**	2.54	**3,884.82**	233,240.40
	5	**2.00**	2.93	**166.39**	511.40	**2.18**	3.34	**542.99**	4,000.43	**2.09**	2.81	**1,600.59**	93,614.03	**2.11**	2.69	**6,977.03**	487,793.27
	6	**2.01**	3.4	**157.22**	543.40	**2.23**	3.90	**600.26**	4,925.72	**2.07**	2.95	**1,475.44**	89,256.52	**2.07**	2.75	**7,468.79**	477,779.38
	7	**2.02**	3.57	**183.94**	580.31	**2.36**	4.60	**946.69**	7,375.86	**2.06**	3.13	**1,324.89**	91,019.27	**2.28**	2.92	**9,407.30**	505,694.61
	8	**2.12**	3.53	**196.21**	548.73	**2.34**	4.76	**838.92**	6,627.67	**2.05**	3.18	**1,363.23**	91,303.71	**2.22**	3.00	**11,932.63**	698,688.14
	9	**2.14**	3.83	**189.07**	644.01	**2.60**	4.84	**992.74**	6,155.02	**2.21**	3.31	**1,607.96**	94,727.09	**2.51**	3.06	**18,195.46**	738,946.86
	10	**2.34**	3.76	**239.85**	648.19	**2.93**	4.94	**1,696.12**	8,067.53	**2.19**	3.54	**1,603.01**	101,592.34	**2.48**	3.14	**18,065.29**	728,729.51
1,000	3	**2.00**	2.41	**192.90**	345.29	**2.04**	2.85	**570.83**	3,886.05	**2.01**	2.36	**1,466.39**	65,796.50	**2.06**	2.26	**7,502.69**	377,201.00
	4	**2.00**	2.63	**176.61**	398.02	**2.02**	3.13	**409.00**	3,175.71	**2.03**	2.52	**1,501.12**	70,126.25	**2.02**	2.44	**8,132.37**	458,117.02
	5	**2.00**	3.05	**191.04**	523.95	**2.08**	3.68	**546.94**	4,722.26	**2.03**	2.53	**1,627.94**	78,087.27	**2.07**	2.44	**11,046.23**	340,031.16
	6	**2.00**	3.14	**220.65**	646.08	**2.11**	3.92	**478.48**	4,246.79	**2.02**	2.81	**1,515.72**	78,163.45	**2.07**	2.53	**8,464.73**	409,394.42
	7	**2.00**	3.42	**186.93**	619.69	**2.24**	4.42	**474.59**	4,638.83	**2.03**	2.89	**1,608.96**	85,140.48	**2.07**	2.69	**8,688.97**	462,350.82
	8	**2.15**	3.53	**305.03**	863.16	**2.40**	4.46	**612.92**	4,175.98	**2.02**	2.98	**1,732.70**	95,208.96	**2.19**	2.84	**11,744.51**	621,596.49
	9	**2.14**	3.65	**273.39**	893.33	**2.49**	4.77	**798.63**	5,146.23	**2.05**	3.2	**1,956.38**	110,580.81	**2.30**	2.96	**15,009.47**	629,376.48
	10	**2.31**	3.68	**327.50**	824.54	**2.63**	5.20	**1,012.23**	5,640.97	**2.13**	3.34	**1,819.46**	99,845.39	**2.38**	3.23	**23,286.48**	734,467.38
2,000	3	**2.00**	2.28	**365.66**	568.54	**2.02**	2.71	**752.81**	3,042.69	**2.00**	2.07	**3,154.06**	70,375.79	**2.00**	2.07	**14,128.61**	371,909.18
	4	**2.00**	2.6	**359.36**	682.04	**2.00**	3.00	**1,015.42**	4,631.35	**2.00**	2.21	**2,908.72**	68,373.82	**2.00**	2.17	**15,791.23**	426,750.80
	5	**2.00**	2.78	**359.87**	723.58	**2.01**	3.23	**965.89**	4,840.78	**2.00**	2.23	**2,968.97**	70,547.89	**2.00**	2.24	**8,888.93**	229,417.35
	6	**2.01**	3.03	**341.83**	804.80	**2.02**	3.47	**1,300.53**	7,052.13	**2.00**	2.39	**3,074.33**	76,812.58	**2.00**	2.30	**17,043.63**	454,348.81
	7	**2.00**	3.23	**347.87**	851.02	**2.08**	4.28	**1,447.79**	9,136.36	**2.00**	2.51	**3,086.85**	81,112.69	**2.04**	2.57	**21,202.57**	497,631.07
	8	**2.08**	3.48	**403.80**	1,049.60	**2.11**	4.22	**1,727.57**	9,855.98	**2.00**	2.82	**3,381.36**	98,470.62	**2.00**	2.64	**13,664.83**	438,992.88
	9	**2.06**	3.76	**401.20**	1,158.89	**2.18**	4.32	**1,465.14**	8,146.38	**2.01**	3.02	**3,230.91**	93,257.26	**2.01**	2.85	**18,864.84**	582,386.30
	10	**2.22**	3.72	**507.41**	1,279.01	**2.46**	4.78	**2,297.41**	10,735.07	**2.00**	3.16	**2,977.52**	89,691.08	**2.06**	3.05	**22,298.43**	588,388.82

Continued

Table 7. (*Continued*)

		3				4				5				6			
		ANI		ART		ANI		ART		ANI		ART		ANI		ART	
N		*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*
4,000	3	**2.00**	2.25	**669.50**	874.49	**2.00**	2.53	**3,019.00**	7,767.42	**2.00**	2.06	**5,711.99**	65,685.79	**2.00**	2.02	**28,981.75**	375,757.58
	4	**2.00**	2.56	**636.80**	990.19	**2.00**	2.80	**2,736.33**	7,890.05	**2.00**	2.04	**6,472.74**	74,092.00	**2.00**	2.04	**30,234.60**	389,949.16
	5	**2.00**	2.79	**703.66**	1,236.56	**2.00**	3.01	**2,366.18**	7,310.50	**2.00**	2.07	**6,676.14**	77,693.98	**2.00**	2.05	**27,773.84**	357,789.51
	6	**2.00**	3.1	**657.30**	1,355.21	**2.04**	3.43	**2,485.29**	8,779.46	**2.00**	2.22	**6,568.00**	81,726.51	**2.00**	2.28	**34,078.71**	502,262.94
	7	**2.00**	3.23	**710.70**	1,562.43	**2.09**	3.96	**2,443.08**	10,419.75	**2.00**	2.43	**6,246.15**	84,939.22	**2.01**	2.37	**31,225.19**	377,166.30
	8	**2.07**	3.33	**824.91**	1,778.29	**2.11**	4.29	**2,464.49**	10,244.44	**2.00**	2.6	**6,042.51**	82,494.12	**2.00**	2.52	**20,761.23**	305,869.73
	9	**2.06**	3.64	**818.13**	2,063.05	**2.19**	4.44	**2,871.67**	11,875.28	**2.01**	2.76	**6,481.76**	86,933.12	**2.00**	2.63	**20,489.54**	280,631.67
	10	**2.18**	3.84	**816.44**	1,960.71	**2.41**	4.84	**4,059.31**	16,171.49	**2.00**	3.07	**7,183.22**	112,090.43	**2.00**	2.60	**19,506.58**	271,895.75

Note. ANI = average number of iterations; ART = average running time (in seconds).
[a]Numbers in bold denote greater **Q**-matrix estimation efficiency.

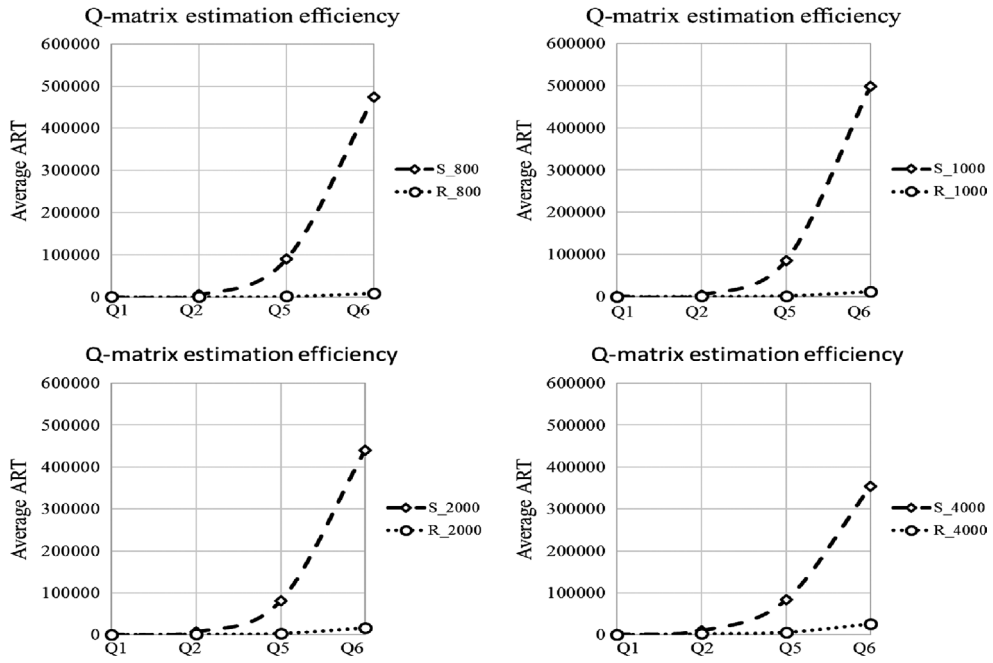

Figure 1. Average running time for **Q**-matrix estimation under different numbers of misspecified items for a fixed sample size.

As mentioned earlier in this paper, a key motivation for developing the *R* method is to improve computational efficiency. Table 7 presents the results pertaining to the computational efficiency of the two methods. As we can see, the *R* method has a clear advantage in computational efficiency. All the ANIs for the *R* method are smaller than 3, indicating that the validation process can be finished within up to three iterations. In contrast, for the *S* method, it might take up to five iterations to converge. According to the ART, huge differences are observed between the two methods in the computation time they require. All simulations were programmed in Matlab 2017b, and were run on parallel grids at the Center for Research Computing (CRC) at University of Notre Dame. Each replication for the two methods was run on the same machine, so the ART still can provide a perspective for computational efficiency. As we can see, the *S* method takes at least twice as long as the *R* method for the estimation of the $\mathbf{Q}_1$, which involves three attributes. The ARTs of the *S* method rise sharply as the number of attributes rises. Take the sample size of 800 as an example. Compared to $\mathbf{Q}_1$, the *S* method requires 10 times as long an ART to estimate $\mathbf{Q}_2$, about 200 times to estimate $\mathbf{Q}_5$, and at least 700 times to estimate $\mathbf{Q}_6$. The rise in the ART of the *R* method is much more moderate. Again taking the sample size 800 as an example, relative to $\mathbf{Q}_1$, the *R* method requires no more than 8 times as long an ART to estimate $\mathbf{Q}_2$, no more than 10 times to estimate $\mathbf{Q}_5$, and no more than 100 times to estimate $\mathbf{Q}_6$. More specifically, for sample size 800, in order to estimate the $\mathbf{Q}_6$, the average ART (under different misspecified items) for the *R* and *S* methods is about 2.6 and 131.8 hrs, respectively. Figure 1 illustrates the average ARTs, which show a stark contrast in the efficiency between the two methods.

In sum, the simulation study shows that when the AMPs are unknown, the *R* method improves upon the *S* method in terms of both **Q**-matrix recovery and computational efficiency.

4.3. Primary simulation II: comparison of the two methods based on unevenly distributed AMPs

The goal of primary simulation II is to assess the performance of the two methods for unevenly distributed population. For the data simulation, everything else was kept the same as in primary simulation I except the distribution of the AMPs. In this study the AMPs of the population of interest are simulated from the Bernoulli distribution, $\alpha_{ik} \sim$ Bernoulli (0.3), where the probability of mastering an attribute is .3. The distribution of the number of attributes mastered by an examinee therefore follows Binomial(K, .3). To take a five-attribute CDA as an example, the probability of an examinee possessing 0, 1, 2, 3, 4, and 5

Table 8. **Q**-matrix estimation accuracy for three attributes, 20-item test ($\mathbf{Q}_1$)

		QRR		NRI		NRA		NOA		NUA	
		[0,1]		[0,20]		[0,60]		[0,60]		[0,60]	
N		R	S	R	S	R	S	R	S	R	S
800	3	**.97**	.81	**19.70**	18.85	**59.97**	59.71	**0.00**	0.20	**0.03**	0.09
	4	**.98**	.81	**19.68**	18.68	**59.88**	59.71	**0.07**	0.21	**0.05**	0.08
	5	**.95**	.78	**19.62**	18.59	**59.66**	59.65	**0.18**	0.23	**0.16**	0.12
	6	**.89**	.69	**19.51**	18.56	**58.67**	58.68	**0.71**	0.82	**0.62**	0.50
	7	**.77**	.69	**18.59**	17.83	**57.87**	57.04	**1.39**	1.69	**0.74**	1.27
	8	**.62**	.56	**18.51**	17.74	**55.93**	53.89	**2.47**	3.31	**1.60**	2.80
	9	**.46**	.41	**17.53**	16.62	**52.26**	49.71	**4.49**	5.41	**3.25**	4.88
	10	**.29**	.28	**16.66**	15.59	**49.35**	47.00	**6.21**	6.65	**4.44**	6.35
1,000	3	**.99**	.85	**19.85**	19.13	**59.99**	59.83	**0.00**	0.14	**0.01**	0.03
	4	**.99**	.84	**19.73**	19.11	**59.99**	59.33	**0.00**	0.44	**0.01**	0.23
	5	**.95**	.84	**19.67**	18.69	**59.44**	58.33	**0.37**	0.87	**0.19**	0.80
	6	**.89**	.84	**19.54**	18.62	**59.67**	59.21	**0.20**	0.40	**0.13**	0.39
	7	**.76**	.74	**18.67**	18.36	**57.55**	56.51	**1.56**	1.90	**0.89**	1.59
	8	**.71**	.55	**18.51**	17.72	**56.68**	54.22	**1.93**	3.09	**1.39**	2.69
	9	**.55**	.36	**17.54**	16.67	**53.77**	49.58	**3.60**	5.63	**2.63**	4.79
	10	**.47**	.31	**16.58**	15.61	**50.98**	47.40	**5.25**	6.28	**3.77**	6.32
2,000	3	**1.00**	.98	**20.00**	19.72	**60.00**	59.78	**0.00**	0.12	**0.00**	0.10
	4	**1.00**	.96	**20.00**	19.68	**60.00**	59.65	**0.00**	0.20	**0.00**	0.15
	5	**.99**	.96	**19.67**	19.64	**59.99**	59.46	**0.01**	0.25	**0.00**	0.29
	6	**.96**	.86	**19.67**	19.51	**59.73**	58.73	**0.21**	0.73	**0.06**	0.54
	7	**.93**	.83	**19.55**	18.53	**59.18**	57.23	**0.49**	1.51	**0.33**	1.26
	8	**.87**	.67	**18.71**	17.66	**57.15**	55.05	**1.74**	2.38	**1.11**	2.57
	9	**.85**	.47	**17.65**	16.66	**54.62**	51.27	**3.10**	4.43	**2.28**	4.30
	10	**.76**	.40	**16.64**	15.53	**52.42**	47.15	**4.33**	6.33	**3.25**	6.52
4,000	3	**1.00**	1.00	**20.00**	20.00	**60.00**	60.00	**0.00**	0.00	**0.00**	0.00
	4	**1.00**	.98	**20.00**	19.72	**60.00**	59.78	**0.00**	0.12	**0.00**	0.10
	5	**1.00**	.98	**20.00**	19.72	**60.00**	59.72	**0.00**	0.15	**0.00**	0.13
	6	**.95**	.91	**19.74**	19.53	**59.15**	59.04	**0.50**	0.52	**0.35**	0.44
	7	**.95**	.87	**19.57**	18.61	**59.19**	57.88	**0.48**	1.08	**0.33**	1.04
	8	**.91**	.73	**18.86**	17.69	**56.91**	54.42	**1.80**	2.81	**1.29**	2.77
	9	**.89**	.69	**17.76**	16.67	**54.72**	51.06	**3.07**	4.62	**2.21**	4.32
	10	**.82**	.54	**17.51**	15.64	**52.95**	47.09	**3.98**	6.39	**3.07**	6.52

Note. Numbers in bold indicate better performance in the associated criterion for the corresponding method.

Table 9. **Q**-matrix estimation accuracy for four attributes, 20-item test ($\mathbf{Q}_2$)

		QRR		NRI		NRA		NOA		NUA	
		[0,1]		[0,20]		[0,80]		[0,80]		[0,80]	
N		*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*
800	3	**.47**	.38	**18.75**	17.67	**78.78**	70.07	**0.48**	5.51	**0.74**	4.42
	4	**.42**	.31	**18.67**	16.58	**78.10**	63.56	**0.95**	10.09	**0.95**	6.35
	5	**.38**	.29	**18.62**	15.64	**76.98**	65.79	**1.57**	8.20	**1.45**	6.01
	6	**.32**	.24	**18.59**	15.59	**74.29**	66.23	**3.22**	7.93	**2.49**	5.84
	7	**.30**	.22	**17.66**	14.65	**70.52**	61.23	**5.25**	11.00	**4.23**	7.77
	8	**.29**	.18	**15.75**	14.52	**65.56**	58.13	**8.33**	13.23	**6.11**	8.64
	9	**.26**	.17	**15.53**	13.71	**62.63**	56.21	**9.55**	13.58	**7.82**	10.21
	10	**.25**	.15	**14.72**	13.59	**61.16**	54.75	**10.08**	13.84	**8.76**	11.41
1,000	3	**.59**	.53	**18.73**	17.57	**79.37**	70.16	**0.17**	5.29	**0.46**	4.55
	4	**.56**	.48	**18.73**	16.68	**78.26**	68.17	**0.87**	7.07	**0.87**	4.76
	5	**.55**	.46	**18.67**	15.55	**77.72**	65.92	**1.13**	8.41	**1.15**	5.67
	6	**.52**	.43	**18.64**	15.53	**74.27**	63.98	**3.39**	9.28	**2.34**	6.74
	7	**.50**	.41	**16.64**	14.61	**68.07**	59.84	**6.78**	11.98	**5.15**	8.18
	8	**.49**	.31	**16.53**	14.57	**67.54**	58.08	**6.91**	12.82	**5.55**	9.10
	9	**.44**	.28	**15.59**	13.53	**62.93**	56.81	**9.47**	13.29	**7.60**	9.90
	10	**.37**	.25	**14.69**	13.52	**61.09**	56.15	**10.30**	12.98	**8.61**	10.87
2,000	3	**.98**	.71	**18.70**	18.75	**79.81**	76.66	**0.08**	1.67	**0.11**	1.67
	4	**.92**	.70	**18.63**	18.56	**79.76**	74.77	**0.14**	3.01	**0.10**	2.22
	5	**.89**	.67	**18.57**	17.61	**77.84**	71.97	**1.25**	4.63	**0.91**	3.40
	6	**.87**	.66	**18.53**	16.57	**75.77**	69.62	**2.69**	6.07	**1.54**	4.31
	7	**.85**	.58	**17.55**	15.52	**72.40**	64.09	**4.41**	9.46	**3.19**	6.45
	8	**.83**	.41	**16.62**	14.70	**69.25**	59.70	**6.35**	11.92	**4.40**	8.38
	9	**.82**	.40	**15.70**	14.67	**64.44**	58.16	**8.67**	12.36	**6.89**	9.48
	10	**.72**	.37	**15.62**	14.54	**62.78**	58.52	**9.30**	11.57	**7.92**	9.91
4,000	3	**.99**	.91	**18.67**	18.71	**79.95**	78.58	**0.03**	0.73	**0.02**	0.69
	4	**.95**	.90	**18.64**	18.57	**79.66**	77.38	**0.24**	1.53	**0.10**	1.09
	5	**.94**	.89	**18.63**	18.50	**79.08**	75.20	**0.63**	2.79	**0.29**	2.01
	6	**.91**	.89	**18.60**	17.50	**76.11**	70.25	**2.16**	5.51	**1.73**	4.24
	7	**.90**	.71	**17.53**	16.51	**73.49**	66.32	**3.72**	8.00	**2.79**	5.68
	8	**.88**	.72	**16.64**	15.70	**69.87**	62.42	**6.41**	10.55	**4.72**	7.03
	9	**.81**	.54	**15.75**	14.69	**65.23**	59.44	**8.92**	11.41	**6.85**	9.15
	10	**.80**	.51	**15.68**	14.68	**64.60**	58.62	**8.73**	11.66	**6.67**	9.72

Note. Numbers in bold indicate better performance in the associated criterion for the corresponding method.

attributes is $\left[(1-0.3)^5, C_5^1 \times (1-0.3)^4 \times 0.3^1, C_5^2 \times (1-0.3)^3 \times 0.3^2, C_5^3 \times (1-0.3)^2 \times 0.3^3, C_5^4 \times (1-0.3)^1 \times 0.3^4, 0.3^5\right]$, respectively. The AMPs are not uniformly distributed in the population, as the pattern of mastering all attributes is much smaller than that of mastering none of the attributes. In contrast, in primary simulation study I, the AMPs are evenly sampled from all 2^K possible patterns.

Tables 8–11 summarize the results in terms of **Q**-matrix estimation accuracy. First, results indicate that in the presence of unevenly distributed AMPs, both the *R* method and the *S* method show declined performance, especially when the number of misspecified items is large and the sample size is relatively small. Second, for $\mathbf{Q}_1$, $\mathbf{Q}_2$, $\mathbf{Q}_5$, and $\mathbf{Q}_6$, the *R*

Table 10. **Q**-matrix estimation accuracy for five attributes, 30-item test ($\boldsymbol{Q}_5$)

		QRR		NRI		NRA		NOA		NUA	
		[0,1]		[0,30]		[0,150]		[0,150]		[0,150]	
N		*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*
800	3	**.51**	.34	**28.67**	27.75	**148.28**	140.08	**0.99**	7.01	**0.73**	2.91
	4	**.48**	.26	**28.65**	27.67	**146.41**	139.13	**2.21**	7.71	**1.38**	3.16
	5	**.37**	.20	**28.64**	26.75	**148.21**	134.87	**1.02**	10.21	**0.77**	4.92
	6	**.32**	.17	**28.62**	26.52	**147.40**	133.98	**1.42**	11.09	**1.18**	4.93
	7	**.29**	.15	**28.62**	25.68	**147.62**	130.08	**1.35**	13.82	**1.03**	6.10
	8	**.27**	.11	**28.61**	25.52	**145.79**	128.24	**2.60**	14.45	**1.61**	7.31
	9	**.25**	.10	**28.56**	24.57	**143.42**	126.19	**3.90**	15.60	**2.68**	8.21
	10	**.24**	.10	**27.65**	23.58	**139.00**	121.55	**6.96**	18.32	**4.04**	10.13
1,000	3	**.56**	.47	**28.75**	28.52	**149.02**	142.63	**0.46**	5.03	**0.52**	2.34
	4	**.52**	.44	**28.74**	27.68	**148.92**	142.09	**0.55**	5.54	**0.53**	2.37
	5	**.50**	.42	**28.61**	27.65	**148.01**	140.45	**1.10**	6.78	**0.89**	2.77
	6	**.49**	.36	**28.57**	27.64	**148.33**	138.29	**0.84**	8.14	**0.83**	3.57
	7	**.45**	.27	**28.57**	26.70	**146.00**	133.98	**2.65**	10.91	**1.35**	5.11
	8	**.42**	.16	**28.53**	25.57	**145.72**	129.70	**2.68**	13.38	**1.60**	6.92
	9	**.33**	.14	**28.52**	24.66	**144.83**	124.25	**5.12**	16.28	**3.05**	9.47
	10	**.30**	.10	**27.68**	24.62	**143.59**	125.86	**3.73**	15.99	**2.68**	8.15
2,000	3	**.98**	.58	**28.74**	28.70	**149.82**	148.12	**0.09**	1.29	**0.09**	0.59
	4	**.89**	.57	**28.73**	28.64	**148.85**	146.32	**0.72**	2.54	**0.43**	1.14
	5	**.85**	.53	**28.71**	28.61	**149.52**	147.82	**0.30**	1.59	**0.18**	0.59
	6	**.84**	.51	**28.70**	28.60	**149.71**	148.70	**0.16**	1.01	**0.13**	0.29
	7	**.83**	.46	**28.64**	28.56	**148.30**	145.00	**1.13**	3.50	**0.57**	1.50
	8	**.81**	.43	**28.63**	28.51	**147.96**	144.38	**1.29**	3.61	**0.75**	2.01
	9	**.77**	.37	**28.56**	27.73	**146.75**	137.68	**2.25**	8.14	**1.00**	4.18
	10	**.67**	.34	**28.51**	27.62	**144.33**	137.81	**3.65**	7.90	**2.02**	4.29
4,000	3	**.99**	.95	**28.75**	28.74	**149.99**	149.94	**0.00**	0.06	**0.01**	0.00
	4	**.99**	.94	**28.74**	28.65	**149.99**	149.90	**0.01**	0.09	**0.00**	0.01
	5	**.96**	.88	**28.65**	28.61	**149.77**	149.59	**0.18**	0.30	**0.05**	0.11
	6	**.98**	.88	**28.65**	28.60	**149.92**	148.95	**0.06**	0.74	**0.02**	0.31
	7	**.92**	.88	**28.61**	28.60	**149.09**	148.59	**0.61**	0.97	**0.30**	0.44
	8	**.89**	.80	**28.60**	28.53	**148.21**	147.14	**1.11**	1.85	**0.68**	1.01
	9	**.85**	.79	**28.57**	28.53	**147.80**	146.75	**1.18**	2.11	**1.02**	1.14
	10	**.80**	.74	**28.55**	28.50	**146.94**	145.40	**1.31**	2.78	**1.75**	1.82

Note. Numbers in bold indicate better performance in the associated criterion for the corresponding method.

method still outperforms the *S* method. Third, the *R* method shows a greater advantage over the *S* method when the sample size is 2,000. When the sample size reaches 4,000, both methods exhibit comparable performance for evenly and unevenly distributed examinees. Results also indicate that unevenly distributed AMPs result in more iterations and longer computation time. There are several reasons why these results were observed. Under unevenly distributed AMPs, there are likely a small number of examinees mastering many (or very few) attributes. This can make it challenging to estimate well the item parameters and the AMPs, which negatively affects the **Q**-matrix validation. Note that

Table 11. **Q**-matrix estimation accuracy for six attributes, 30-item test ($\boldsymbol{Q}_6$)

		QRR		NRI		NRA		NOA		NUA	
		[0,1]		[0,30]		[0,180]		[0,180]		[0,180]	
N		*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*
800	3	**.37**	.31	**28.72**	25.71	**178.91**	155.05	**0.64**	19.58	**0.45**	5.37
	4	**.35**	.23	**28.62**	25.65	**178.57**	153.12	**1.05**	21.26	**0.38**	5.62
	5	**.25**	.18	**28.61**	25.58	**177.26**	154.17	**1.95**	20.15	**0.79**	5.68
	6	**.18**	.14	**28.57**	24.70	**176.25**	148.37	**2.84**	24.97	**0.91**	6.66
	7	**.16**	.10	**28.56**	24.51	**174.55**	147.28	**4.05**	25.65	**1.40**	7.07
	8	**.17**	.08	**28.54**	23.61	**172.91**	143.18	**5.28**	28.77	**1.81**	8.05
	9	**.12**	.06	**28.51**	22.68	**172.17**	140.83	**5.89**	29.99	**1.94**	9.18
	10	**.12**	.04	**27.69**	22.61	**170.45**	138.53	**7.41**	31.72	**2.14**	9.75
1,000	3	**.52**	.44	**28.75**	27.63	**179.22**	166.52	**0.48**	10.81	**0.30**	2.67
	4	**.47**	.43	**28.74**	26.64	**178.72**	162.80	**0.94**	13.49	**0.34**	3.71
	5	**.47**	.42	**28.73**	26.62	**178.61**	163.04	**1.07**	13.42	**0.32**	3.54
	6	**.48**	.33	**28.68**	26.61	**178.29**	160.55	**1.30**	15.45	**0.40**	4.00
	7	**.43**	.21	**28.65**	25.63	**176.95**	154.40	**2.26**	19.86	**0.79**	5.75
	8	**.35**	.13	**28.64**	24.73	**176.39**	147.46	**2.79**	25.07	**0.82**	7.47
	9	**.31**	.10	**28.60**	23.57	**174.81**	142.49	**3.96**	28.54	**1.24**	8.97
	10	**.28**	.05	**28.55**	23.50	**173.25**	141.14	**5.22**	29.56	**1.53**	9.30
2,000	3	**.96**	.53	**28.74**	28.71	**179.96**	177.31	**0.02**	2.13	**0.02**	0.57
	4	**.90**	.56	**28.73**	28.69	**179.87**	177.07	**0.05**	2.40	**0.07**	0.52
	5	**.86**	.44	**28.70**	28.60	**179.47**	174.97	**0.37**	3.92	**0.15**	1.12
	6	**.82**	.41	**28.68**	27.51	**178.95**	169.07	**0.82**	8.68	**0.23**	2.25
	7	**.82**	.23	**28.66**	26.55	**178.20**	164.20	**1.41**	12.27	**0.39**	3.53
	8	**.81**	.15	**28.54**	25.72	**177.65**	155.20	**1.70**	19.03	**0.65**	5.76
	9	**.65**	.11	**28.52**	25.53	**176.26**	153.09	**2.71**	20.64	**1.03**	6.27
	10	**.63**	.06	**28.52**	24.56	**175.19**	148.99	**3.30**	23.55	**1.51**	7.46
4,000	3	**.98**	.91	**28.74**	28.75	**179.94**	179.57	**0.05**	0.38	**0.01**	0.05
	4	**.98**	.91	**28.72**	28.72	**179.95**	178.52	**0.04**	1.21	**0.01**	0.27
	5	**.97**	.89	**28.71**	28.61	**179.64**	178.82	**0.28**	0.98	**0.08**	0.20
	6	**.96**	.79	**28.70**	28.61	**179.68**	176.55	**0.22**	2.60	**0.10**	0.85
	7	**.89**	.68	**28.67**	28.57	**178.55**	172.93	**1.09**	5.60	**0.36**	1.47
	8	**.87**	.45	**28.64**	27.59	**179.53**	165.72	**0.31**	10.94	**0.16**	3.34
	9	**.82**	.38	**28.61**	26.51	**177.09**	161.38	**1.84**	14.19	**1.07**	4.43
	10	**.76**	.29	**28.55**	25.57	**175.66**	157.70	**3.05**	17.05	**1.29**	5.25

Note. Numbers in bold indicate better performance in the associated criterion for the corresponding method.

estimation of the AMP is required to calculate the *S* statistic, but not for the *R* statistic – that may be why the *S* statistic suffers more from unevenly distributed AMPs.

Table 12 presents the results in terms of **Q**-matrix estimation efficiency. Again the *R* method showed a clear advantage in computational efficiency. As in primary simulation study I, all the ANIs of the *R* method are smaller than those of the *S* method. For the *S* method, the ANIs increase substantially when the number of misspecifications is large. The ANIs of the *S* method can reach up to 6.81 (compared to 5 when the AMPs are evenly distributed). This suggests that the *S* method needs more iterations to converge under the unevenly distributed samples.

Table 12. Q-matrix estimation efficiency

		3				4				5				6			
		ANI		ART		ANI		ART		ANI		ART		ANI		ART	
N		*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*
800	3	**2.05**	3.10	**350.52**	773.56	**2.34**	3.64	**2,151.72**	9,030.30	**2.53**	3.49	**5,307.76**	116,207.54	**3.27**	4.49	**21,008.59**	359,181.45
	4	**2.04**	3.16	**370.77**	802.45	**2.27**	3.91	**2,157.24**	8,941.23	**2.81**	3.34	**5,435.61**	121,833.36	**3.55**	4.75	**26,461.76**	366,213.68
	5	**2.14**	3.41	**368.57**	826.86	**2.41**	4.21	**2,112.28**	9,018.82	**2.62**	2.46	**6,295.68**	115,606.59	**3.73**	4.33	**21,618.01**	578,771.32
	6	**2.30**	3.79	**416.32**	899.74	**2.44**	4.73	**1,833.35**	9,763.80	**3.04**	3.01	**8,368.08**	119,239.12	**3.54**	4.90	**30,534.66**	594,637.40
	7	**2.37**	4.16	**435.46**	955.08	**2.46**	4.97	**2,218.98**	9,916.41	**3.27**	3.01	**9,242.56**	119,044.08	**3.58**	4.64	**29,753.70**	696,138.65
	8	**2.50**	4.51	**465.32**	1053.97	**3.61**	5.21	**2,277.71**	10,660.02,	**3.17**	3.61	**8,975.99**	154,946.98	**3.91**	6.68	**31,863.65**	705,589.64
	9	**2.70**	4.92	**521.90**	1102.18	**3.53**	5.50	**2,120.09**	14,357.74	**3.59**	4.24	**10,359.93**	167,362.52	**4.09**	6.44	**30,455.89**	811,225.36
	10	**2.95**	4.97	**611.92**	1149.37	**3.65**	5.68	**2,833.01**	15,535.56	**3.75**	4.46	**11,550.36**	180,119.35	**4.40**	6.53	**33,360.66**	820,296.13
1,000	3	**2.08**	3.15	**403.77**	823.16	**2.31**	3.57	**1,775.88**	9,412.95	**2.98**	3.25	**4,891.04**	114,494.01	**3.25**	4.48	**21,242.03**	357,936.84
	4	**2.04**	3.37	**397.16**	880.42	**2.54**	3.84	**2,262.75**	8,774.75	**2.96**	3.43	**5,254.41**	115,633.44	**3.43**	4.11	**29,711.49**	364,507.73
	5	**2.16**	3.77	**440.87**	968.15	**2.44**	4.24	**1,811.88**	8,859.41	**3.12**	3.87	**6,349.95**	118,923.95	**3.75**	4.64	**26,029.97**	576,911.56
	6	**2.23**	3.07	**445.17**	817.70	**2.65**	4.63	**2,109.48**	9,357.89	**3.21**	3.82	**8,498.51**	118,508.69	**3.83**	4.63	**316,30.73**	581,992.93
	7	**2.41**	3.99	**521.41**	1,020.77	**2.57**	5.16	**2,203.47**	11,117.89	**3.25**	3.26	**9,471.23**	121,863.08	**3.29**	4.76	**378,15.19**	693,567.93
	8	**2.34**	4.33	**482.71**	1,134.09	**3.47**	5.37	**1,996.91**	12,147.59	**3.39**	3.72	**10,595.41**	145,219.88	**3.83**	6.36	**39,652.40**	709,726.54
	9	**2.45**	4.69	**552.60**	1,207.79	**3.58**	5.55	**2,190.90**	13,709.32	**3.59**	4.12	**10,541.80**	158,249.72	**4.02**	6.25	**42,789.09**	825,694.09
	10	**2.74**	4.97	**635.26**	1,270.40	**3.59**	5.60	**2,107.45**	18,744.25	**3.66**	4.99	**12,802.01**	227,126.70	**4.06**	6.45	**47,881.34**	846,655.31
2,000	3	**2.00**	3.08	**584.42**	904.33	**2.50**	3.54	**1,436.67**	8,742.24	**2.17**	3.25	**4,963.71**	116,644.93	**3.03**	4.32	**21,301.60**	370,165.86
	4	**2.05**	3.17	**588.36**	929.63	**2.56**	3.67	**1,947.36**	9,361.91	**2.40**	3.60	**6,046.38**	120,252.63	**3.28**	4.43	**25,668.46**	380,363.13
	5	**2.00**	3.32	**578.35**	965.86	**2.97**	3.99	**2,485.79**	9,718.31	**2.25**	3.37	**7,382.50**	117,990.65	**3.66**	4.64	**32,263.66**	594,165.39
	6	**2.08**	3.63	**628.21**	1,512.71	**2.87**	4.53	**3,146.87**	9,921.61	**2.13**	3.18	**7,677.41**	116,020.76	**3.84**	4.13	**41,072.34**	592,529.86
	7	**2.17**	3.97	**649.32**	1,678.76	**2.25**	5.02	**3,853.89**	1,861.92	**2.59**	3.13	**8,464.75**	124,728.45	**3.65**	4.59	**59,461.70**	719,565.78
	8	**2.28**	4.01	**775.76**	2,110.00	**3.71**	5.16	**5,653.44**	13,945.80	**2.60**	3.22	**8,413.06**	186,031.70	**3.63**	6.72	**71,306.69**	717,680.47
	9	**2.49**	4.47	**954.21**	2,153.72	**3.22**	5.23	**5,233.80**	14,305.37	**2.88**	3.48	**1,3024.80**	188,294.37	**3.96**	6.81	**73,242.31**	834,880.60
	10	**2.64**	4.77	**1062.70**	2,259.26	**3.33**	5.80	**5,922.50**	18,383.97	**3.68**	3.89	**1,3347.79**	232,889.40	**4.13**	6.31	**91,687.15**	837,493.65

Continued

Table 12. (*Continued*)

		3				4				5				6			
		ANI		ART		ANI		ART		ANI		ART		ANI		ART	
N		*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*	*R*	*S*
4,000	3	**2.00**	3.06	**716.14**	1,330.90	**2.25**	3.36	**3,571.59**	11,326.46	**2.01**	3.05	**8,349.50**	119,329.74	**3.04**	4.05	**34,901.05**	377,044.04
	4	**2.00**	3.11	**676.16**	1,507.97	**2.20**	3.56	**4,328.44**	13,766.77	**2.00**	3.12	**8,269.85**	119,926.05	**3.13**	4.15	**33,122.05**	381,967.77
	5	**2.02**	3.27	**777.39**	1,662.21	**2.40**	3.84	**4,997.11**	14,076.02	**2.01**	3.25	**8,676.83**	121,924.69	**3.11**	4.60	**42,846.30**	605,777.18
	6	**2.08**	3.56	**833.20**	1,821.11	**2.77**	4.45	**7,367.20**	19,285.08	**2.12**	3.35	**10,129.70**	123,128.51	**3.25**	4.92	**47,699.79**	620,854.90
	7	**2.12**	3.84	**1,068.81**	2,231.98	**2.12**	4.92	**8,146.91**	21,318.94	**2.13**	3.62	**10,795.97**	126,569.06	**3.65**	4.17	**61,261.05**	735,240.90
	8	**2.26**	4.07	**1,412.12**	2,505.32	**3.15**	5.41	**9,738.62**	25,884.73	**3.25**	3.92	**13,118.71**	161,179.08	**3.98**	6.38	**97,001.18**	742,123.64
	9	**2.43**	4.55	**1,309.59**	2,688.84	**3.12**	5.45	**9,976.29**	26,144.49	**3.36**	4.06	**15,310.11**	193,778.96	**4.42**	6.69	**10,5235.07**	863,854.55
	10	**2.51**	4.66	**1,213.08**	2,855.26	**3.40**	5.51	**11,065.23**	28,200.51	**3.46**	4.56	**15,977.81**	280,385.42	**4.42**	6.05	**11,8327.90**	881,933.92

Note. ANI = average number of iterations; ART = average running time (in seconds).
Numbers in bold denote greater **Q**-matrix estimation efficiency.

In summary, when the unknown AMPs are not uniformly distributed, both the *R* method and the *S* method performed less well than when the AMPs are uniformly distributed. However, the *R* method still leads to better performance in terms of **Q** matrix estimation accuracy and efficiency than the *S* method.

5. Real data analysis

The Examination for the Certificate of Proficiency in English (ECPE) is a test administered by the English Language Institute of the University of Michigan (https://michiganassessment.org/test-takers/tests/ecpe/), which is designed to measure advanced English skills in examinees whose primary language is not English (Templin & Hoffman, 2013). The data set analysed here consists of 2,922 examinees from a single year's administration, which can be found in the R package *CDM* (https://cran.r-project.org/web/packages/CDM/). Based on the ECPE data, Templin and Hoffman (2013) demonstrated how to use Mplus to fit cognitive diagnostic models. Like Templin and Hoffman (2013), we also consider the 28 multiple-choice questions which measure three skills: (A1) morphosyntactic rules, (A2) cohesive rules, and (A3) lexical rules. The original **Q**-matrix is provided in Table 13.

Taking the original **Q**-matrix and the item response data set as input, we applied the *R* method to estimate the **Q**-matrix. The **Q**-matrix suggested by the *R*-method is provided in Table 14. As we can see, the *R*-method suggested 9 items and a total of 10 attributes be revised. The estimation process took five iterations and 2,576.53 s. Results also indicated that the distribution of the estimated AMPs is non-uniform: the probability of the eight latent classes [0 0 0], [1 0 0], [0 1 0], [0 0 1], [1 1 0], [1 0 1, [0 1 1], and [1 1 1] is .20, .00, .14, .01, .03, .00, .13, and .48, respectively. Note that only six of the eight possible AMPs were represented in this sample. None of the test takers had AMP [1 0 0] or [1 0 1].

For the data set, the main disagreement lies in the definition of the second attribute, for which seven revisions were suggested by the *R*-method, and all of them from 0 to 1. This indicates that although the attribute "cohesive rules" was measured by some items, it might have been ignored by content experts for other items. For the third attribute, three revisions (two $1 \rightarrow 0$, one $0 \rightarrow 1$) were suggested. No revision was suggested for the first attribute. Because details of the 28 items were not available to us, we could not make further comments about these revisions. That said, this example shows that how the *R*

Table 13. The original Q-matrix of the ECPE data

Item	A1	A2	A3	Item	A1	A2	A3
1	1	1	0	15	0	0	1
2	0	1	0	16	1	0	1
3	1	0	1	17	0	1	1
4	0	0	1	18	0	0	1
5	0	0	1	19	0	0	1
6	0	0	1	20	1	0	1
7	1	0	1	21	1	0	1
8	0	1	0	22	0	0	1
9	0	0	1	23	0	1	0
10	1	0	0	24	0	1	0
11	1	0	1	25	1	0	0
12	1	0	1	26	0	0	1
13	1	0	0	27	1	0	0
14	1	0	0	28	0	0	1

Table 14. The **Q**-matrix suggested by the *R* method

Item	A1	A2	A3	Item	A1	A2	A3
1	1	1	0	15	0	0	1
2	0	1	0	16	1	0	1
3	1	***1***	1	17	0	1	1
4	0	***1***	1	***18***	0	***1***	1
5	0	0	1	19	0	0	1
6	0	***1***	1	20	1	0	1
7	1	0	***0***	***21***	1	***1***	1
8	0	1	0	22	0	0	1
9	0	0	1	23	0	1	0
10	1	0	0	24	0	1	0
11	1	0	***0***	25	1	0	0
12	1	0	1	***26***	0	***1***	1
13	1	0	0	27	1	0	0
14	1	***1***	***1***	28	0	0	1

Note. The entries in bold italics refer to the changes suggested by the *R* method.

method can be used to provided valuable information to the testing program for close inspection of its **Q**-matrix.

6. Discussion

The **Q**-matrix plays an important role in CDA because its misspecification can have serious detrimental effects on the classification of respondents and item parameter estimation (Im & Corter, 2011; Rupp & Templin, 2008a). Researchers have emphasized that a correctly specified **Q**-matrix is one of the most crucial pieces of a CDA (Lim & Drasgow, 2017). Based on the DINA model, this paper proposes a residual-based statistic to validate the **Q**-matrix. We show analytically the properties of the *R* statistic under ideal conditions, and evaluate its performance under realistic conditions through simulation studies. Results show that the *R* method leads to higher accuracy in **Q**-matrix recovery and higher computational efficiency than the *S* method. Therefore, we believe we offer a very promising new method for **Q**-matrix validation.

However, our study is limited in a number of ways. First, the *R* method is developed and evaluated under the DINA model. The properties of the weighted residual are yet to be demonstrated under other CDMs, and the performance of *R* method is still an open question. Parallel residual-based methods following the same logic could potentially be developed for other CDMs, for example, the G-DINA model (Ma & de la Torre, 2019; de la Torre, 2011). Second, it is worthwhile to consider **Q**-matrix estimation with hierarchical attributes (Leighton, Gierl, & Hunka, 2004), which can be important in real applications. Third, in addition to the methods that fall squarely into either the sequential or whole-vector search categories, there exist other methods for **Q**-matrix validation. For example, through a Bayesian extension of the DINA model, DeCarlo (2012) showed that **Q**-matrix uncertainty could be recognized and explored. Based on a model–data fit perspective, Chen (2017) used a combination of fit measures to validate the **Q**-matrix, and showed that the misspecified items can be detected and adjusted sequentially. By means of the EM algorithm (de la Torre, 2009), Wang *et al.* (2018) considered **Q**-matrix validation by taking the candidate vector corresponding to the maximum likelihood as the item attribute

vector, and their results exhibit varying degrees of effectiveness in **Q**-matrix validation under different conditions. It is important in the future to consider a comprehensive comparison of these methods under various scenarios and models, and possibly identify the best method for different conditions.

Acknowledgement

This work is supported in part by NSF CAREER grant DRL-1350787 awarded to the corresponding author.

References

Akbay, L. (2016). *Identification, estimation, and Q-matrix validation of hierarchically structured attributes in cognitive diagnosis* (Doctoral dissertation). State University of New Jersey. https://doi.org/10.7282/T3RR21JV

Baghaei, P., & Hohensinn, C. (2017). A method of Q-matrix validation for the linear logistic test model. *Frontiers in Psychology*, *8*, 1–7. https://doi.org/10.3389/fpsyg.2017.00897

Barnes, T. (2010). Novel derivation and application of skill matrices: The Q-matrix method. In C. Romero, S. Ventura, M. Pechenizkiy & R. S. J. d. Baker (Eds.), *Handbook on educational data mining* (pp. 159–172). Boca Raton, FL: CRC Press. https://doi.org/10.1201/b10274

Chen, J. S. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement*, *41*, 277–293. https://doi.org/10.1177/0146621616686021

Chen, P., Xin, T., Wang, C., & Chang, H. H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, *77*(2), 201–222. https://doi.org/10.1007/S11336-012-9255-7

Chen, Y. G., Culpepper, S. A., Chen, Y. G., & Douglas, J. (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika*, *83*(1), 89–108. https://doi.org/10.1007/s11336-017-9579-4

Chen, Y. X., Liu, J. C., Xu, G. J., & Ying, Z. L. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of American Statistical Association*, *110*, 850–866. https://doi.org/10.1080/01621459.2014.934827

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*(4), 619–632. https://doi.org/10.1007/s11336-009-9123-2

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

Chiu, C. Y., & Kohn, H. F. (2015). Consistency of cluster analysis for cognitive diagnosis: The DINO model and the DINA model revisited. *Applied Psychological Measurement*, *39*, 465–479. https://doi.org/10.1177/0146621615577087

Chung, M. T. (2014). *Estimating the Q-matrix for cognitive diangnosis models in a Bayesian framework* (Doctoral dissertation). Columbia University. https://doi.org/10.7916/D857195B

Close, C. N. (2012). *An exploratory technique for finding the Q-matrix for the DINA model in cognitive diagnostic assessment: Combining theory with data* (Doctoral dissertation). University of Minnesota. Retrieved from http://hdl.handle.net/11299/121595

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*(4), 343–362. https://doi.org/10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130. https://doi.org/10.3102/1076998607309474

de la Torre, J. (2011). The generalized dina model framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/S11336-011-9207-7

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*(2), 253–273. https://doi.org/10.1007/s11336-015-9467-8

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

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

Feng, Y. L. (2013). *Estimation and Q-matrix validation for diagnostic classification models* (Doctoral dissertation). University of South Carolina.

Fu, J. B., & Li, Y. M. (2007). *Cognitively diagnostic psychometric models: An integrative review*. Chicago, IL: Paper presented at the National Council on Measurement in Education.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, *71*(2), 407–419. https://doi.org/10.1177/0013164410388832

Im, S., & Corter, J. E. (2011). Statistical consequences of attribute misspecification in the rule space method. *Educational and Psychological Measurement*, *71*, 712–731. https://doi.org/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*(3), 258–272. https://doi.org/10.1177/01466210122032064

Lee, Y. W., & Sawaki, Y. (2009). Cognitive diagnosis and Q-matrices in language assessment. *Language Assessment Quarterly*, *6*, 169–171. https://doi.org/10.1080/15434300903059598

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511611186

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*(3), 205–237. https://doi.org/10.1111/j.1745-3984.2004.tb01163.x

Li, H. C. (2016). *Estimation of Q-matrix for DINA model using the constrained generalized DINA framework* (Doctoral dissertation). Columbia University. https://doi.org/10.7916/D85B097W

Lim, Y. S., & Drasgow, F. (2017). Nonparametric calibration of item-by-attribute matrix in cognitive diagnosis. *Multivariate Behavioral Research*, *52*(5), 562–575. https://doi.org/10.1080/00273171.2017.1341829

Liu, J. C. (2016). On the consistency of Q-matrix estimation: A commentary. *Psychometrika*, *82*, 523–527. https://doi.org/10.1007/s11336-015-9487-4

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

Liu, J. C., Xu, G. J., & Ying, Z. L. (2013). Theory of the self-learning Q-matrix. *Bernoulli*, *19*(5A), 1790–1817. https://doi.org/10.3150/12-BEJ430

Ma, L. (2014). *Validation of the item-attribute matrix in TIMSS–Mathematics using multiple regression and the LSDM* (Doctoral dissertation). University of Denver.

Ma, W., & de la Torre, J. (2019). Digital module 05: Diagnostic measurement – The G-DINA framework. *Educational Measurement: Issues and Practice*, *38*, 114–115.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden (Ed.), *Handbook of modern item response theory* (pp. 101–121). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-2691-6_6

McGlohen, M., & Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*(3), 808–821. https://doi.org/10.3758/BRM.40.3.808

Romero, S. J., & Ordonez, X. G. (2014). Validation of the cognitive structure of an arithmetic test with the least squres distance model (LSDM). *Universitas Psychologica*, *13*(1), 333–344.

Romero, S. J., Ordonez, X. G., Ponsoda, V., & Revuelta, J. (2014). Detection of Q-matrix misspecification using two criteria for validation of cognitive structures under the least squares

distance model. *Psicologica: International Journal of Methodology and Experimental Psychology*, *35*(1), 149–169.

Rupp, A. A., & Templin, J. L. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*(1), 78–96. https://doi.org/10.1177/0013164407301545

Rupp, A. A., & Templin, J. L., (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, *6*, 219–262. https://doi.org/10.1080/15366360802490866

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

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York, NY: Routledge. https://doi.org/10.4324/9780203883372

Templin, J. L., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(2), 37–50. https://doi.org/10.1111/emip.12010

Terzi, R., & de la Torre, J. (2018). An iterative method for empirically-based Q-matrix validation. *International Journal of Assessment Tools in Education*, *5*(2), 248–262. https://doi.org/10.21449/ijate.407193

Tu, D. B., Cai, Y., & Dai, H. Q. (2012). A new method of Q-matrix validation based on DINA model. *Acta Psychological Sinica*, *44*(4), 558–568.

Wang, W. Y., Song, L. H., Ding, S. L., Meng, Y. R., Cao, C. X., & Jie, Y. J. (2018). An EM-based method for Q-matrix validation. *Applied Psychological Measurement*, *42*, 446–459. https://doi.org/10.1177/0146621617752991

Xiang, R. (2013). *Nonlinear penalized estimation of true Q-Matrix in cognitive diagnostic models* (Doctoral dissertation). Columbia University. https://doi.org/10.7916/D8J96DKZ

Xu, G. J., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika*, *81*, 625–649. https://doi.org/10.1007/s11336-015-9471-z

Yu, X., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods*, *24*(5), 658–674. https://doi.org/10.1037/met0000212.

Received 26 September 2018; revised version received 15 July 2019

Appendix A:

The six true Q-matrices

$$\mathbf{Q}_1 = \begin{bmatrix} 1&0&0\\ 0&1&0\\ 0&0&1\\ 1&0&0\\ 0&1&0\\ 0&0&1\\ 1&0&0\\ 0&1&0\\ 0&0&1\\ 1&1&0\\ 1&0&1\\ 1&0&1\\ 0&1&1\\ 1&0&1\\ 0&1&1\\ 1&1&0\\ 0&1&1\\ 1&0&1\\ 0&1&1\\ 1&1&1 \end{bmatrix}, \mathbf{Q}_2 = \begin{bmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1\\ 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&1\\ 1&1&0&0\\ 1&0&1&0\\ 1&0&0&1\\ 0&1&1&0\\ 0&1&0&1\\ 0&0&1&1\\ 1&1&1&0\\ 1&1&0&1\\ 1&0&1&1\\ 0&1&1&1\\ 1&1&1&1 \end{bmatrix}, \mathbf{Q}_3 = \begin{bmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1\\ 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1\\ 1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1\\ 0&1&1&0&0\\ 0&1&0&1&0\\ 0&1&0&0&1\\ 0&0&1&1&0\\ 0&0&1&0&1\\ 0&0&0&1&1 \end{bmatrix}, \mathbf{Q}_4 = \begin{bmatrix} 1&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&1&0&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&1\\ 1&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&1&0&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&1\\ 1&1&0&0&0&1\\ 1&0&1&0&1&0\\ 1&0&0&1&0&1\\ 1&1&0&0&1&0\\ 0&0&1&0&1&1\\ 0&0&1&1&0&1\\ 0&1&1&1&0&0\\ 0&1&0&1&1&0 \end{bmatrix}, \mathbf{Q}_5 = \begin{bmatrix} 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1\\ 1&0&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1\\ 1&1&0&0&0\\ 1&0&1&0&0\\ 1&0&0&1&0\\ 1&0&0&0&1\\ 0&1&1&0&0\\ 0&1&0&1&0\\ 0&1&0&0&1\\ 0&0&1&1&0\\ 0&0&1&0&1\\ 0&0&0&1&1 \end{bmatrix}, \mathbf{Q}_6 = \begin{bmatrix} 1&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&1&0&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&1\\ 1&0&0&0&0&0\\ 0&1&0&0&0&0\\ 0&0&1&0&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&1\\ 1&1&0&0&0&0\\ 1&0&1&0&0&0\\ 1&0&0&1&0&0\\ 1&0&0&0&1&0\\ 1&0&0&0&0&1\\ 0&1&1&0&0&0\\ 0&1&0&1&0&0\\ 0&1&0&0&1&0\\ 0&1&0&0&0&1\\ 0&0&1&1&0&0\\ 0&0&1&0&1&0\\ 0&0&1&0&0&1\\ 0&0&0&1&1&0\\ 0&0&0&1&0&1\\ 0&0&0&0&1&1\\ 1&0&1&0&1&0\\ 0&1&0&1&0&1\\ 1&0&0&0&0&1 \end{bmatrix}$$

Appendix B:

Proof of Theorem 1

Theorem 1. *Suppose the following assumptions hold: (a) $N \to \infty$; (b) AMPs are known; (c) $s_j, g_j \in (0, 0.5)$. If M_1, M_2, or M_3 occurs on the jth item, then one or both of the item parameter estimates are positively biased.*

Proof. Under the DINA model, according to the values of the η and the response to the jth item, each respondent is classified into one of the four groups,G_1, G_2, G_3 and G_4, in which there are N_{11}^j, N_{10}^j, N_{01}^j, and N_{00}^j respondents, respectively. We have $N_{11}^j + N_{10}^j + N_{01}^j + N_{00}^j = N$. Consistent with the main text, the two numbers in the subscript refer to the values of the η and the item response. For example, N_{11}^j indicates the number of respondents in group G_1, in which the respondents possess all the required attributes of item j and answer it correctly, that is, $\eta_{ij} = 1$ and $X_{ij} = 1$.

As shown on page 8, given the assumptions that the AMPs and the true **Q**-matrix are known, these two inequalities, $E[N_{11}^j] > E\left[N_{10}^j\right]$, $E[N_{00}^j] > E[N_{01}^j]$ hold.

The maximum likelihood estimates of the item parameters are

$$\hat{s}_j = \frac{N_{10}^j}{N_{11}^j + N_{10}^j}, \tag{19}$$

$$\hat{g}_j = \frac{N_{01}^j}{N_{01}^j + N_{00}^j}. \tag{20}$$

When the attribute vector of item j is defined without misspecification, $E\left[\hat{s}_j\right] = s_j$ and $E\left[\hat{g}_j\right] = g_j$. However, if the attribute vector is misspecified, there can be systematic bias in $\hat{s}_j$ and/or $\hat{g}_j$, depending on the type of misspecification.

Case 1. When M_1 occurs, some of the respondents in G_3 and G_4 will be misclassified into G_1 and G_2, respectively. Denote the total number of people who move from G_3 and G_4 to G_1 and G_2 by ΔN. This is the number of non-masters, given the misspecified attribute vector, who are now considered masters. ΔX_{01}^j and ΔX_{00}^j are two random variables that represent the number of people who will move into G_1 and G_2, respectively, and their realizations ΔN_{01}^j and ΔN_{00}^j in the sample (where $\Delta N_{00}^j + \Delta N_{01}^j = \Delta N$). It follows that $\Delta X_{01}^j \sim B(\Delta N, g_j)$. Therefore, g_j can be estimated via maximum likelihood as

$$\hat{g}_{j\Delta} = \frac{\Delta N_{01}^j}{\Delta N_{01}^j + \Delta N_{00}^j},$$

and $E\left[\hat{g}_{j\Delta}\right] = g_j$. On the other hand, when the **Q**-matrix was correctly specified, $\hat{g}_j = \frac{N_{01}^j}{N_{01}^j + N_{00}^j}$ and $E\left[\hat{g}_j\right] = g_j$.

Given the misspecification, the numbers of respondents in the four groups in the sample are updated as $N_{11}^{j*} = N_{11}^j + \Delta N_{01}^j$, $N_{10}^{j*} = N_{10}^j + \Delta N_{00}^j$, $N_{01}^{j*} = N_{01}^j - \Delta N_{01}^j$, and $N_{00}^{j*} = N_{00}^j - \Delta N_{00}^j$. Given the updated numbers, the new maximum likelihood estimate of g_j under M_1 becomes

$$\hat{g}_{j*} = \frac{N_{01}^{j*}}{N_{01}^{j*} + N_{00}^{j*}} = \frac{N_{01}^{j} - \Delta N_{01}^{j}}{N_{01}^{j} + N_{00}^{j} - \Delta N_{01}^{j} - \Delta N_{00}^{j}}$$
$$= \frac{(N_{01}^{j} + N_{00}^{j})\hat{g}_j - \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)\hat{g}_{j\Delta}}{(N_{01}^{j} + N_{00}^{j}) - \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)}.$$

Therefore:

$$E\left[\hat{g}_{j*}\right] = \frac{(N_{01}^{j} + N_{00}^{j})E\left(\hat{g}_j\right) - \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)E\left(\hat{g}_{j\Delta}\right)}{(N_{01}^{j} + N_{00}^{j}) - \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)}$$
$$= \frac{(N_{01}^{j} + N_{00}^{j})g_j - \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)g_j}{(N_{01}^{j} + N_{00}^{j}) - \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)} = g_j.$$

This result suggests that when M_1 occurs, the estimate of g_j remains unbiased. However, the estimate of s_j under M_1 becomes:

$$\hat{s}_{j*} = \frac{N_{10}^{j*}}{N_{11}^{j*} + N_{10}^{j*}} = \frac{N_{10}^{j} + \Delta N_{00}^{j}}{N_{11}^{j} + N_{10}^{j} + \Delta N_{01}^{j} + \Delta N_{00}^{j}}$$
$$= \frac{(N_{11}^{j} + N_{10}^{j})\hat{s}_j + \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)\left(1 - \hat{g}_{j\Delta}\right)}{(N_{11}^{j} + N_{10}^{j}) + \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)}. \tag{21}$$

Hence

$$E\left[\hat{s}_{j*}\right] = \frac{(N_{11}^{j} + N_{10}^{j})E\left[\hat{s}_j\right] + \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)E\left[\left(1 - \hat{g}_{j\Delta}\right)\right]}{(N_{11}^{j} + N_{10}^{j}) + \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)}$$
$$= \frac{(N_{11}^{j} + N_{10}^{j})s_j + \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)\left(1 - g_j\right)}{(N_{11}^{j} + N_{10}^{j}) + \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)}.$$

Given that $s_j, g_j \in (0, 0.5)$, we have $\left(1 - g_j\right) > 0.5 > s_j$. Therefore,

$$E\left[\hat{s}_{j*}\right] > \frac{(N_{11}^{j} + N_{10}^{j})s_j + \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)s_j}{(N_{11}^{j} + N_{10}^{j}) + \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)} = s_j.$$

This result indicates that under M_1, the estimate of s_j is positively biased. The magnitude of bias is a function of $1 - g_j$–s_j.

Case 2. When M_2 occurs, some of the respondents in G_1 and G_2 will be classified into G_3 and G_4, respectively. Denote the number of such respondents by ΔN_{11}^{j} and ΔN_{10}^{j}. The numbers of respondents in the four groups become $N_{11}^{j*} = N_{11}^{j} - \Delta N_{11}^{j}$, $N_{10}^{j*} = N_{10}^{j} - \Delta N_{10}^{j}$, $N_{01}^{j*} = N_{01}^{j} + \Delta N_{11}^{j}$, and $N_{00}^{j*} = N_{00}^{j} + \Delta N_{10}^{j}$. The item parameters under M_2 can be estimated as follows:

$$\hat{s}_{j*} = \frac{N_{10}^{j*}}{N_{11}^{j*} + N_{10}^{j*}} = \frac{N_{10}^{j} - \Delta N_{10}^{j}}{N_{11}^{j} + N_{10}^{j} - \Delta N_{11}^{j} - \Delta N_{10}^{j}}, \tag{23}$$

$$\hat{g}_{j*} = \frac{N_{01}^{j*}}{N_{01}^{j*} + N_{00}^{j*}} = \frac{N_{01}^{j} + \Delta N_{11}^{j}}{N_{01}^{j*} + N_{00}^{j*} + \Delta N_{11}^{j} + \Delta N_{10}^{j}}. \tag{24}$$

Following the same strategy as above, we can show that under M_2, $E\left[\hat{s}_{j*}\right] = s_j$. On the other hand, given that $1 - s_j > 0.5 > g_j$, $E[\hat{g}_{j*}] > g_j$.

Case 3. When M_3 occurs, some of the respondents in G_1 and G_2 will be classified into G_3 and G_4, respectively; denote the number of such respondents by ΔN_{11}^{j} and ΔN_{10}^{j}. Meanwhile, some of the respondents in G_3 and G_4 will be classified into G_1 and G_2, respectively; denote their number by ΔN_{01}^{j} and ΔN_{00}^{j}. The numbers of respondents in the four groups become $N_{11}^{j*} = N_{11}^{j} - \Delta N_{11}^{j} + \Delta N_{01}^{j}$, $N_{10}^{j*} = N_{10}^{j} - \Delta N_{10}^{j} + \Delta N_{00}^{j}$, $N_{01}^{j*} = N_{01}^{j} + \Delta N_{11}^{j} - \Delta N_{01}^{j}$, and $N_{00}^{j*} = N_{00}^{j} + \Delta N_{10}^{j} - \Delta N_{00}^{j}$. The item parameter estimates can be updated as

$$\begin{aligned} \hat{s}_{j*} &= \frac{N_{10}^{j*}}{N_{11}^{j*} + N_{10}^{j*}} = \frac{N_{10}^{j} - \Delta N_{10}^{j} + \Delta N_{00}^{j}}{(N_{11}^{j} + N_{10}^{j}) - \left(\Delta N_{11}^{j} + \Delta N_{10}^{j}\right) + \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)} \\ &= \frac{(N_{11}^{j} + N_{10}^{j})\hat{s}_j - \left(\Delta N_{11}^{j} + \Delta N_{10}^{j}\right)\hat{s}_{j\Delta} + \Delta N_{00}^{j}(1 - \hat{g}_{j\Delta})}{(N_{11}^{j} + N_{10}^{j}) - \left(\Delta N_{11}^{j} + \Delta N_{10}^{j}\right) + \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)}, \end{aligned} \tag{25}$$

$$\begin{aligned} \hat{g}_{j*} &= \frac{N_{01}^{j*}}{N_{01}^{j*} + N_{00}^{j*}} = \frac{N_{01}^{j} + \Delta N_{11}^{j} - \Delta N_{01}^{j}}{(N_{01}^{j} + N_{00}^{j}) + \left(\Delta N_{11}^{j} + \Delta N_{10}^{j}\right) - \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)} \\ &= \frac{(N_{01}^{j} + N_{00}^{j})\hat{g}_j + \left(\Delta N_{11}^{j} + \Delta N_{10}^{j}\right)(1 - \hat{s}_{j\Delta}) - \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)\hat{g}_{j\Delta}}{(N_{01}^{j} + N_{00}^{j}) + \left(\Delta N_{11}^{j} + \Delta N_{10}^{j}\right) - \left(\Delta N_{01}^{j} + \Delta N_{00}^{j}\right)}. \end{aligned} \tag{26}$$

Given that $(1 - g_j) > 0.5 > s_j$, from equation (25) we get $E[\hat{s}_{j*}] > s_j$. On the other hand, given that $1 - s_j > 0.5 > g_j$, from equation (26) we get $E[\hat{g}_{j*}] > g_j$. ■