Online Item Calibration for **Q**-matrix in CD-CAT

Yunxiao Chen, Jingchen Liu, and Zhiliang Ying

November 8, 2013

Abstract

Item replenishment is important to maintaining a large scale item bank. In this paper we consider calibrating new items based on pre-calibrated operational items under the DINA model, the specification of which includes the so-called **Q**-matrix, as well as the slipping and guessing parameters. Making use of the maximum likelihood and Bayesian estimators for the latent knowledge states, we propose two methods for the calibration. These methods are applicable to both traditional paper-pencil based tests, for which the selection of operational items is prefixed, and computerized adaptive tests, for which the selection of operational items is sequential and random. Extensive simulations are done to assess and to compare the performance of these approaches. Extensions to other diagnostic classification models are also discussed.

Keywords: Online calibration, Computerized adaptive testing, Diagnostic classification models

1 Introduction

Diagnostic classification models (DCM) are an important statistical tool in cognitive diagnosis that can be used in a number of disciplines, including educational assessment and clinical psychology Rupp and Templin (2008b). A key component of many DCMs is the so-called **Q**-matrix Tatsuoka (1983), which specifies the item-attribute relationships of a diagnostic test. Various DCMs have been built around the **Q**-matrix. One simple and widely studied example is the DINA model (deterministic inputs, noisy-and-gate; see Haertel (1989); Junker and Sijtsma (2001)) that is the main focus of this paper. Other important models and developments can be found in DiBello et al. (1995); Junker and Sijtsma (2001); Hartz (2002); Tatsuoka (2002); Leighton et al. (2004); von Davier (2005); Templin and Henson (2006); Chiu et al. (2009); Tatsuoka (2009); Rupp and Templin (2010).

Computerized adaptive testing (CAT) is a testing mode in which the item selection is sequential and individually tailored to each examinee. In particular, subsequent items are selected based on the examinee's responses to prior items. CAT was originally proposed by Lord (1971) for item response theory (IRT) models, for which items are tailored for each examinee to 'best fit' his/her ability level θ, so that more capable examinees avoid receiving items that are too simple and less capable examinees avoid receiving items that are too difficult. Such individualized testing schemes perform better than traditional exams with a prefixed selection of items because the optimal selection of testing items is subject dependent. It also leads to greater efficiency and precision than that can be achieved in traditional tests Wainer et al. (1990); van der Linden and Glas (2000).

For CAT under IRT settings, items are typically chosen to maximize the Fisher information (MFI) Lord (1980); Thissen and Mislevy (2000) or to minimize the expected posterior variance (MEPV) van der Linden (1998); Owen (1975). For CAT under diagnostic classification models, recent developments include Xu et al. (2003); Cheng (2009); Liu et al. (2013).

An important task in maintaining a large scale item bank for CAT is item replenishment. As an item becomes exposed to more and more examinees, it needs to be replaced by new ones, for which the item-specific parameters need to be calibrated according to existing items in the bank. In CAT, online calibration is commonly employed to calibrate new items Stocking (1988); Wainer and Mislevy (1990). That is, to estimate the item-specific parameters, new items are assigned to examinees during their tests together with the existing items in the bank (also known as the operational items). In the literature, several online calibration methods have been developed for item-response-theory-based computerized adaptive tests for which the examinees' latent traits are characterized by a unidimensional θ. A short list of methods includes Stocking's Method A and Method B Stocking (1988), marginal maximum likelihood estimation with one EM iteration (OEM) Wainer and Mislevy (1990), marginal maximum likelihood estimation with multiple EM iterations (MEM) Ban et al. (2001, 2002), the BILOG/Prior method Ban et al. (2001), and marginal Bayesian estimation (MBE) with MCMC approach Segall (2003).

In the context of cognitive diagnosis, three online calibration methods, namely CD-Method A, CD-OEM and CD-MEM, are proposed by Chen et al. (2012). These methods focus on the calibration of the slipping and guessing parameters, assuming the corresponding $\mathbf{Q}$-matrix entries are known. They are parallel to the calibration methods for IRT as described in the preceding paragraph. The CD-Method A is a natural extension of Stocking's Method A that plugs in the knowledge state estimates. The CD-OEM and CD-MEM methods are similar to the OEM and MEM methods developed for IRT models. When the $\mathbf{Q}$-matrix of the new items is unknown, the joint estimation algorithm (JEA) is proposed by Chen and Xin (2011), which depends entirely on

the examinees' responses to the operational and new items to jointly calibrate the **Q**-matrix and the slipping and guessing parameters.

In this paper, we further extend the work of Chen et al. (2012) by considering item calibration in terms of both the **Q**-matrix and the slipping and guessing parameters. The **Q**-matrix is a key component in the specification of DCM and, when correctly specified, allows for the accurate calibration of the other parameters. On the other hand, its misspecification could lead to serious problems in all aspects; for example, see Rupp and Templin (2008a) for the effects of misspecification of **Q**-matrix in DINA model. In this paper, we extend the analysis by considering the **Q**-matrix entries of the new items as additional item-specific parameters that are to be estimated simultaneously with the slipping and guessing parameters. Such a data-driven **Q**-matrix also serves as a validation of the subjective item-attribute relationship specified in the initial construction of new items. Our methods are different from the joint estimation algorithm Chen and Xin (2011) and a comparison is made in the simulation study.

The rest of this paper is organized as follows. In the next section, we first review existing online calibration methods of new items whose **Q**-matrix is completely specified as well as the joint estimation algorithm when the **Q**-matrix of new items is unknown and then propose new approaches that simultaneously calibrate both the **Q**-matrix and the slipping and guessing parameters. A simulation section is presented to compare the performance among various methods. In the last section, conclusions drawn from the simulation studies as well as further discussions are provided.

2 Calibration for cognitive diagnosis

2.1 Problem setting

Throughout this paper, we consider an item bank containing a sufficiently large number of operational items whose parameters have already been calibrated. There are m additional items whose parameters are to be calibrated. Both the operational items and the new items are associated with at most k attributes. The calibration procedure is carried out as follows. Each examinee responds to C operational items and D new items. In a traditional paper-pencil test, the operational items assigned to each examinee are identical. In a computerized adaptive test, the item selections are tailored to each examinee. The proposed calibration procedure does not particularly depend on the testing mode. Furthermore, for $j = 1, ..., m$, we let n_j be the total number of examinees responding to the new item j, $\mathbf{R}_j = (R_{1,j}, ..., R_{n_j,j})$ be the vector of responses to new item j, $\mathbf{r}_i = (r_{i,1}, ..., r_{i,C})$ be the response vector of examinee i to the C operational items.

The DINA model that is commonly used in educational assessment will be assumed. Under

the DINA model, the knowledge state is described by a k dimensional vector with zero-one entries. Specifically, examinee i's knowledge state is given by a vector $\boldsymbol{\alpha}_i = (\alpha_i^1, ..., \alpha_i^k)$, where α_i^l is either one or zero, indicating the presence or absence, respectively, of the lth skill. In this paper, the terms "knowledge state", "attribute profile", and "skill" are exchangeable and denoted by vector $\boldsymbol{\alpha}$. The DINA model assumes a conjunctive relationship among the skills. Consider an item and let $q = (q^1, ..., q^k)$ be the corresponding row vector in the $\mathbf{Q}$-matrix, where $q^l = 1$ indicates that the correct response of this item requires the presence of attribute l. Furthermore, the DINA model assumes that an examinee is capable of providing a correct answer to this item when the examinee possesses all the required skills. Thus, we define the *ideal response* of an examinee of attribute $\boldsymbol{\alpha}$ to an item of row vector q as

$$\xi(q, \boldsymbol{\alpha}) = \prod_{l=1}^{k} (\alpha^l)^{q^l} = \mathbf{1}(\alpha^l \geq q^l \text{ for all } l = 1, ..., k).$$

The response distribution is then defined as

$$p_{s,g}(q, \boldsymbol{\alpha}) \triangleq P(R = 1 | q, \boldsymbol{\alpha}) = \begin{cases} 1 - s, & \text{if } \xi(q, \boldsymbol{\alpha}) = 1; \\ g, & \text{if } \xi(q, \boldsymbol{\alpha}) = 0. \end{cases} \tag{1}$$

The parameter s is known as the slipping parameter, representing the probability of an incorrect response to the item for examinees who are capable of answering correctly, and g is known as the guessing parameter, representing the probability of a correct response for those who are not capable.

Suppose that an examinee's responses to a set of operational items $\mathbf{r} = (r_1, ..., r_C)$ have been collected. We use q_i, s_i, and g_i to denote the row vectors of the $\mathbf{Q}$-matrix, the slipping parameters, and the guessing parameters, respectively. In the setting of computerized adaptive testing, the selection of items is possibly random in that the specific choice of (q_j, s_j, g_j) typically depends on the examinee's previous responses $(r_1, ..., r_{j-1})$. Here, we make the assumption that the sequential selection rule of subsequent items only depends on the responses $(r_1, ..., r_{j-1})$ and does not depend on any other information of the knowledge state $\boldsymbol{\alpha}$. Therefore, the observation of item selections does not provide further information on the knowledge state. Based on this, we can write down the likelihood function of knowledge state

$$L(\boldsymbol{\alpha}; \mathbf{q}, \mathbf{s}, \mathbf{g}, \mathbf{r}) = \prod_{j=1}^{C} p_{s_j, g_j}(q_j, \boldsymbol{\alpha})^{r_j} [1 - p_{s_j, g_j}(q_j, \boldsymbol{\alpha})]^{1 - r_j} \tag{2}$$

where $\mathbf{q} = (q_1, ..., q_C)$, $\mathbf{s} = (s_1, ..., s_C)$, and $\mathbf{g} = (g_1, ..., g_C)$. Under the Bayesian framework, inferences about $\boldsymbol{\alpha}$ can be made based on its posterior distribution

$$\pi(\boldsymbol{\alpha}|\mathbf{q}, \mathbf{s}, \mathbf{g}, \mathbf{r}) \propto L(\boldsymbol{\alpha}; \mathbf{q}, \mathbf{s}, \mathbf{g}, \mathbf{r})\pi(\boldsymbol{\alpha})$$

where $\pi(\boldsymbol{\alpha})$ is the prior distribution and the symbol "$\propto$" reads as "is proportional to."

2.2 Existing methods for online calibration for CD-CAT with a known Q-matrix

We begin with a brief review of the three online calibration methods proposed in Chen et al. (2012). The purpose of these methods is to estimate the slipping and guessing parameters s and g when the corresponding **Q**-matrix is specified (known).

For a specific new item j, suppose that there are n_j examinees responding to the item. The first method, which is known as CD-Method A, considers the estimated the knowledge state $\hat{\boldsymbol{\alpha}}_i$ as the true, for $i = 1, ..., n_j$. Estimates of the slipping and guessing parameter are obtained via the maximum likelihood estimator that solves the following normal equations

$$\frac{\partial l_j}{\partial s_j} = 0, \qquad \frac{\partial l_j}{\partial g_j} = 0, \tag{3}$$

where $l_j(q_j, s_j, g_j) = \log(\prod_{i=1}^{n_j} p_{s_j,g_j}(q_j, \hat{\boldsymbol{\alpha}}_i)^{R_{i,j}}[1 - p_{s_j,g_j}(q_j, \hat{\boldsymbol{\alpha}}_i)]^{1-R_{i,j}})$ and q_j is the row vector of **Q**-matrix for the new item. The parameters s_j and g_j enter the likelihood through the probability $p_{s_j,g_j}(q_j, \hat{\boldsymbol{\alpha}}_j)$ defined as in (1).

The second method, which is known as the CD-OEM, considers the uncertainty contained in the estimates $\hat{\boldsymbol{\alpha}}_i$ by incorporating the entire posterior distribution and uses a single cycle of an EM-type algorithm to obtain the marginal maximum likelihood estimate. In particular, for a given new item j, the CD-OEM method first takes one E-step with respect to the posterior distribution of the knowledge states, given the responses to the operational items. Next, the M-step maximizes the logarithm of the expected likelihood.

The third method, CD-MEM, is an extension of the CD-OEM method. It increases the number of EM cycles until some convergence criterion is satisfied. Specifically, the first EM cycle of the CD-MEM method is identical to the CD-OEM method, and the new item parameter estimates obtained from the first EM cycle are regarded as the initial new item parameters of the second EM cycle. From the second EM cycle onwards, the CD-MEM method utilizes the responses from both the operational and new items to obtain the posterior distribution of the knowledge states for the E-step. The M-step is the same as that of the CD-OEM method, except that the likelihood

is marginalized with respect to the posterior distribution given responses to both the operational and the new items. One advantage of the CD-MEM method is that it fully utilizes the information from both the operational and the new items.

2.3 The joint estimation algorithm

When the **Q**-matrix is unknown, the joint estimation algorithm Chen and Xin (2011) estimates both the **Q**-matrix and the slipping and guessing parameters of the new items. The algorithm, as an extension of CD-Method A, treats the estimated knowledge state $\hat{\boldsymbol{\alpha}}_i$ as the true. In particular, the posterior mode is used to estimate the examinees' knowledge states based on their responses to the operational items. The algorithm calibrates one item at a time. For a specific item j, the joint estimation algorithm optimizes $l_j(q_j, s_j, g_j)$ with respect to q_j given (s_j, g_j) and optimizes $l_j(q_j, s_j, g_j)$ with respect to (s_j, g_j) given q_j iteratively until convergence is reached according to some criterion. The advantage of this algorithm is that it is easy to implement.

2.4 Online calibration of Q-matrix

In this subsection, we consider the new item calibration under the DINA model. To motivate our methods, we first consider a hypothetical situation in which the slipping and guessing parameters are known and the **Q**-matrix is the only unknown parameter in need of calibration. We then consider calibrating both the **Q**-matrix and the slipping and guessing parameters. For this, we first present an approach which calibrates one item at a time and then a second approach which deals with multiple items simultaneously. We discuss the advantages in efficiency of the latter over the former.

2.4.1 Calibration with known slipping and guessing parameters

Without loss of generality, we can always rearrange indices so that a new item j is assigned to examinees 1, ..., n_j. For examinee i, we use $\pi_i(\boldsymbol{\alpha}_i)$ to denote the posterior distribution of the knowledge state given his/her responses to the operational items. For a new item with **Q**-matrix row vector q_j, the posterior predictive distribution of a particular response pattern $R_{i,j}$ is

$$p_i(q_j, s_j, g_j) \triangleq P(R_{i,j} = 1 | q_j, s_j, g_j) = \sum_{\boldsymbol{\alpha}} \pi_i(\boldsymbol{\alpha}) p_{s_j, g_j}(q_j, \boldsymbol{\alpha}),\ i = 1, ..., n_j,$$

where $p_{s,g}$ is defined as in (1). Therefore, we can write down the likelihood function based on the responses of n_j examinees as

$$L_j(q_j, s_j, g_j) = \prod_{i=1}^{n_j} p_i(q_j, s_j, g_j)^{R_{i,j}} [1 - p_i(q_j, s_j, g_j)]^{1-R_{i,j}}. \tag{4}$$

Note that here both s_j and g_j are assumed to be known. An estimate of q_j can be obtained through the maximum likelihood estimator (MLE), that is

$$\hat{q}_j = \arg\max_{q_j} L_j(q_j, s_j, g_j).$$

For the computation of the above MLE, notice that there are $2^k - 1$ possible q_j's. We simply compute $L_j(q_j, s_j, g_j)$ for each possible q_j and choose the maximum. This is not much of a computational burden and can be carried out easily for k less than 10.

2.4.2 Calibration for a single item with unknown slipping and guessing parameters

We now proceed to the more realistic situation when s_j and g_j are also unknown and need to be calibrated along with q_j. As in the previous discussion, we still work with the likelihood function (4). The maximum likelihood estimator is then defined as

$$(\hat{q}_j, \hat{s}_j, \hat{g}_j) = \arg\max_{q_j, s_j, g_j} L_j(q_j, s_j, g_j). \tag{5}$$

Because the likelihood here is a function of both discrete q_j and continuous (s_j, g_j), its maximization is not easy to carry out numerically. Our approach is to break it down into two steps. In step 1, for each possible q_j value, we compute the maximized likelihood estimates with respect to s_j and g_j, that is,

$$(\hat{s}_j(q_j), \hat{g}_j(q_j)) = \arg\max_{s_j, g_j} L_j(q_j, s_j, g_j).$$

This step can be carried out by the EM algorithm that is an iterative algorithm. More precisely, the algorithm starts from an initial value (s_j^0, g_j^0). Let (s_j^t, g_j^t) be the parameter values at iteration t. The evolution from (s_j^t, g_j^t) to (s_j^{t+1}, g_j^{t+1}) consists of an E-step and an M-step. In the E-step, the posterior distribution of $\boldsymbol{\alpha}_i$ given a particular response $R_{i,j}$ to the new item is obtained by

$$\tilde{\pi}_i(\boldsymbol{\alpha}_i; s_j^t, g_j^t) \propto \pi_i(\boldsymbol{\alpha}_i) p_{s_j^t, g_j^t}(q_j, \boldsymbol{\alpha}_i)^{R_{i,j}} [1 - p_{s_j^t, g_j^t}(q_j, \boldsymbol{\alpha}_i)]^{1-R_{i,j}}.$$

Then the expected log-likelihood

$$\tilde{l}_j = \sum_{i=1}^{n_j} \sum_{\boldsymbol{\alpha}_i} \tilde{\pi}_i(\boldsymbol{\alpha}_i; s_j^t, g_j^t)[R_{i,j} \log p_{s_j,g_j}(q_j, \boldsymbol{\alpha}_i) + (1 - R_{i,j}) \log(1 - p_{s_j,g_j}(q_j, \boldsymbol{\alpha}_i))]$$

is computed. In the M-step, the parameters are updated by (s_j^{t+1}, g_j^{t+1}) maximizing $\tilde{l}_j$ with respect to (s_j, g_j). Equivalently, (s_j^{t+1}, g_j^{t+1}) solves the normal equations

$$\frac{\partial \tilde{l}_j}{\partial s_j} = 0 \qquad \frac{\partial \tilde{l}_j}{\partial g_j} = 0.$$

The algorithm iterates the E-step and the M-step until convergence, as signaled by some precision rule. Our simulation study shows that the convergence of the EM algorithm is very fast and it typically takes only a few steps.

In step 2, we then obtain $\hat{q}_j$ as the maximizer of the profile likelihood function, that is

$$\hat{q}_j = \arg\max_{q_j} L_j(q_j, \hat{s}_j(q_j), \hat{g}_j(q_j)).$$

Once $\hat{q}_j$ has been computed, the estimates of the slipping and the guessing parameter are then given as $\hat{s}_j(\hat{q}_j)$ and $\hat{g}_j(\hat{q}_j)$.

The above approach calibrates a single item at a time and it is a natural procedure when each examinee is given only a single new item. It is also applicable when multiple new items are assigned to an examinee for which we focus on a particular new item for its calibration and ignore all others. We call this method the single item estimation method (SIE). Under the setting of simulation study 1 in the following section, the calibration of 12 new items in one simulation using the SIE method takes approximately 3.3 seconds in R (version 2.13.1) on a 2.5 Ghz laptop running Windows 7 Professional.

Both JEA and SIE calibrate a single item at a time. However, unlike JEA, the SIE method takes the uncertainty of the knowledge state estimates into account. Instead of plugging in the estimates of the knowledge states, the posterior distributions are used in SIE to calculate the posterior predictive distribution of response patterns. In other words, more information from examinees' responses to the operational items is utilized in the SIE method. Therefore, SIE is expected to be more efficient than JEA in estimating the **Q**-matrix and the slipping and guessing parameters, especially when the estimates of examinees' knowledge states are not accurate and when the sample size is relatively large.

2.4.3 Calibration of multiple items

In this section, we further propose a calibration procedure to calibrate multiple items simultaneously. To start with, we would like to explain why simultaneous calibration could improve the efficiency of the calibration method described in the preceding section. For the calibration of new item-specific parameters, it is clear from (2) that ideally we would like to have examinees' knowledge states known. However, this is practically infeasible. Thus, we make use of the operational items to first get estimates of examinees' knowledge states and then, based on the estimated knowledge states as characterized by their posterior distributions, we proceed to calibrating the new items. Therefore, the more accurate our information about the knowledge states is, the better our calibration will be. The idea of simultaneous calibration is to borrow the information contained in the responses to new items so as to further improve the measurement of the unknown knowledge states. One issue with this idea is that using information from a new item whose parameters (especially **Q**) have not been adequately calibrated may have an adverse effect on the measurement of examinees' knowledge states. Therefore, it is necessary to select the new items with sufficient calibration accuracy (based on the data). In this connection, we introduce an item-specific statistic η_j to quantify the accuracy of the estimation of q_j. We call η_j the confidence index that represents our confidence in the fit of q_j. To start with, we obtain an estimate for each q_j separately via (5) and denote it by $\hat{q}_j$. The confidence index is defined as

$$\eta_j \triangleq \log(\max_{q_j,s_j,g_j} L_j(q_j,s_j,g_j)) - \log(\max_{q_j\neq\hat{q}_j,s_j,g_j} L_j(q_j,s_j,g_j)).$$

If we define $(\tilde{q}_j,\tilde{s}_j,\tilde{g}_j) = \arg\max_{q_j\neq\hat{q}_j,s_j,g_j} L_j(q_j,s_j,g_j)$, then $\tilde{q}_j$ is the second most probable q vector for item j according to the likelihood. In other words, the statistic η_j is the logarithm of the likelihood ratio between $\hat{q}_j$ and $\tilde{q}_j$, the two most probable **Q**'s for item j. The larger η_j is, the more confident we are in the fit of $\hat{q}_j$.

Suppose that there are m new items to be calibrated. We introduce a new method which is built upon the SIE method and simultaneously calibrates all the new items' parameters. It is described by the following algorithm.

1. Calibrate the unknown parameters of new items 1, 2, ..., m, one at a time via the procedure in the preceding section and obtain $(\hat{q}_j,\hat{s}_j,\hat{g}_j,\eta_j)$, for $j = 1,2,...,m$.

2. The new items with η_j larger than a threshold λ are selected, and sorted in a decreasing order according to η_j. Suppose that there are l items selected, denoted by j_1, ..., j_l. These items are viewed as "good" ones for which we are confident in $\hat{q}_j$'s. We choose λ as half of

the 95% quantile of the χ^2 distribution with one degree of freedom. Although the asymptotic distribution of $2\eta_j$ is not really χ^2 distributed and is unclear, simulation study shows that this λ works well empirically, and it can be tuned in applications.

3. We treat new item j_1 as an additional operational item and treat the calibrated parameters $(\hat{q}_{j_1}, \hat{s}_{j_1}, \hat{g}_{j_1})$ as the true. Then, we update the knowledge state posterior distributions for those examinees who responded to new item j_1, given their responses to both the operational items and this new item j_1. With the new knowledge state posterior distributions, we proceed to recalibrate new item j_2 by applying the procedure in the preceding section, and update $(\hat{q}_{j_2}, \hat{s}_{j_2}, \hat{g}_{j_2})$.

$\vdots$

$l+1$. We treat new items j_1, ..., j_{l-1} as operational items and their calibrated parameters as the true. With new knowledge state posterior distributions by further conditioning on the responses to these $l-1$ new items, we apply the procedure in the preceding section to calculate $(\hat{q}_{j_l}, \hat{s}_{j_l}, \hat{g}_{j_l}, \eta_{j_l})$ and update the knowledge state posterior distributions.

Now, all the l selected new items except item j_1 have been recalibrated, and they all serve as the operational items. We continue the procedure by recalibrating the parameters of new items not selected in step 2.

$l+2$. Using the current posterior distributions of knowledge states given the responses to the operational items and the "good" items, recalibrate the parameters of items not selected in step 2 one at a time, according to the procedure in the preceding section.

$l+3$. With the updated $(\hat{q}_j, \hat{s}_j, \hat{g}_j, \eta_j)$, for $j = 1, 2, ..., m$, the items with their η_j's larger than the threshold are selected. If the selected items are the same as those selected in step 2, the algorithm ends. Otherwise, sort the selected items according to the new η_j's from the largest to the smallest, reset the posterior distributions of knowledge states to the one in step 1 (from the responses to the original operational items), and go to step 3.

$\vdots$

The algorithm ends when the selected "good" estimates do not change in two rounds, which intuitively means that all the "good" items have been utilized to refine the estimation of examinees' knowledge states. Then, we report the calibrated item parameter values. We refer to this method as simultaneous item estimation method (SimIE). Under the setting of simulation study 1 in the

following section, the calibration of 12 new items using the SimIE method takes approximately 7.3 seconds in R (version 2.13.1) on a 2.5 Ghz laptop running Windows 7 Professional.

3 Simulation

3.1 Study 1

The purpose of this study is to evaluate the performance of SIE and SimIE and to compare them to JEA. The comparison is made in terms of the accuracy of the estimation of the **Q**-matrix, s and g, and examinees' knowledge states.

3.1.1 Basic setting

In this study, we consider $N = 400$ examinees, each of whom responds to 12 operational items ($C = 12$) and 6 new items ($D = 6$). It is also assumed that there are 12 new items in total ($m = 12$). In addition, all items are assumed to measure up to 5 skills ($k = 5$), and therefore the possible number of knowledge states is $2^k = 32$. Furthermore, the new items are randomly selected and assigned to each examinee. Thus, each new item is assigned to approximately $(N \times D)/m = 200$ examinees ($n_j \approx 200$).

3.1.2 Item bank generation

Item bank generation includes generating the **Q**-matrix as well as the slipping and guessing parameters under the DINA model setting. With a similar setting as in Chen et al. (2012), we generate an item bank of 240 operational items and the **Q**-matrix is therefore a 240×5 matrix. The generation of the **Q**-matrix is as follows. First we generate basic matrices $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$, where rows of $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$ enumerate all possible q vectors for items measuring one, two and three attribute(s), respectively. Then we generate the 240×5 **Q**-matrix by vertically merging 16 copies of $\mathbf{Q_1}$, 8 copies of $\mathbf{Q_2}$, and 8 copies of $\mathbf{Q_3}$. $\mathbf{Q_1}$, $\mathbf{Q_2}$ and $\mathbf{Q_3}$ are of size 5×5, 10×5 and 10×5 respectively, and are shown as follows.

$$\mathbf{Q_1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{Q_2} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}, \quad \mathbf{Q_3} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

As we can see from the **Q**-matrix structure, all skills are measured by the item bank, with each skill measured by 96 items (40% of the item bank). All slipping and guessing parameters are drawn independently from the uniform distribution over the interval $(0.1, 0.4)$, so that there is a wide variety of items in the item bank.

3.1.3 New item generation

The generation of new items includes the generation of their **Q**-matrix (denoted as $\mathbf{Q_{new}}$) and that of the slipping and guessing parameters. As we assume that there are 12 new items, $\mathbf{Q_{new}}$ is a matrix of size 12×5. 12 rows are randomly drawn from the **Q**-matrix for the original item bank, and compose $\mathbf{Q_{new}}$. The slipping and guessing parameters for each new item are drawn independently and uniformly from the interval (0.1, 0.4). The realizations of the item parameters of the new items are shown in Table 1.

In the simulation study, the parameters of items in the item bank and new items are fixed. For each simulation, the knowledge states and responses of 400 examinees are generated. The comparisons are made based on 1000 independent simulations.

3.1.4 Knowledge state and response generation

For each independent simulation, the knowledge states of 400 examinees are simulated independently, assuming each examinee has a 50% probability of mastering each skill. In addition, it is also assumed that the individual skill masteries are mutually independent.

ID of new item	s	g	$\mathbf{Q_{new}}$
1	0.32	0.30	1 0 0 1 0
2	0.18	0.12	0 0 1 0 0
3	0.39	0.32	1 0 1 0 0
4	0.13	0.18	0 1 1 0 0
5	0.38	0.24	0 1 0 0 0
6	0.37	0.20	1 0 1 0 1
7	0.15	0.15	0 0 0 1 0
8	0.16	0.39	0 0 0 1 0
9	0.39	0.13	0 1 0 0 0
10	0.23	0.39	0 1 0 1 1
11	0.30	0.27	1 0 0 0 0
12	0.14	0.18	0 0 0 1 1

Table 1: The slipping and guessing parameters and the **Q**-matrix of the new items

We select the operational items presented to each examinee in two ways. The first is random selection from the 240-item bank. The other is a CD-CAT method presented in Xu et al. (2003), minimizing the expected Shannon entropy of each subsequent item. The selection of 6 new items for each examinee is performed as a simple random selection among the 12 new items in both cases.

3.1.5 Evaluation criteria

We use the following criteria to compare the three methods (JEA, SIE, and SimIE).

1. Item-specific misspecification rate (IMR): The IMR is used to evaluate the estimation accuracy of the **Q**-matrix for each item.

$$\text{IMR}_j = \frac{\sum_{l=1}^{M} \{\hat{q}_j^{(l)} \neq q_j\}}{M}.$$

   The index l represents the lth simulation. In this simulation study, $M = 1000$, which is the number of times the simulation is repeated. The smaller IMR_j is, the more accurate the estimation of q_j is.

2. Total misspecification rate (TMR): The TMR is defined as the average of the IMR's of all new items, which evaluates the overall performance of the estimation of the **Q**-matrix over all the new items. Again, the smaller the TMR is, the more accurate the estimation of **Q**-matrix is. In other words, the TMR simply summarizes the item-specific misspecification rate by taking average

$$\text{TMR} = \frac{1}{m} \sum_{j=1}^{m} \text{IMR}_j.$$

The above two criteria evaluate the estimation of the **Q**-matrix. In addition, slipping and guessing parameters are also calibrated. We evaluate the accuracy of the calibration of the slipping and guessing parameters by the root mean squared error (RMSE).

3. Root mean squared error (RMSE): For a single item j, the RMSE of the estimates of s_j and g_j is defined as follows.

$$\text{RMSE}_{s_j} = \sqrt{\frac{1}{M}\sum_{l=1}^{M}(s_j^{(l)} - s_j)^2},$$

$$\text{RMSE}_{g_j} = \sqrt{\frac{1}{M}\sum_{l=1}^{M}(g_j^{(l)} - g_j)^2}.$$

4. Total mean squared error (TMSE): To summarize the overall performance of the calibration of the slipping and guessing parameters, the TMSE's for all new items are defined as follows.

$$\text{TMSE}_s = \sqrt{\frac{1}{Mm}\sum_{l=1}^{M}\sum_{j=1}^{m}(s_j^{(l)} - s_j)^2},$$

$$\text{TMSE}_g = \sqrt{\frac{1}{Mm}\sum_{l=1}^{M}\sum_{j=1}^{m}(g_j^{(l)} - g_j)^2}.$$

When developing the SimIE method, we claimed that borrowing information from the responses to the new items improves the measurement accuracy of unknown knowledge states. To empirically verify this, we compare the accuracy of the posterior mode of the examinees' knowledge states before and after the SimIE procedure is applied. Note that the JEA and SIE method do not update the posterior distributions of knowledge states. For the purpose of comparison, the pattern misspecification rate (PMR) is introduced to quantify the estimation accuracy of examinees' entire cognitive profiles.

5. Pattern misspecification rate (PMR): The PMR is defined as follows. According to the definition, the smaller the PMR is, the better examinees' knowledge states are learned.

$$\text{PMR} = \frac{1}{MN}\sum_{l=1}^{M}\sum_{i=1}^{N} I_{\{\hat{\alpha}_i^{(l)} \neq \alpha_i^{(l)}\}}.$$

3.1.6 Results

The results of JEA, SIE, and SimIE are summarized in Tables 2, 3, 4 and 5 and Figure 1. When the operational items are selected adaptively, according to the TMR, which quantifies the overall performance of each method in calibrating the **Q**-matrix, SIE (TMR = 0.124) and SimIE (TMR = 0.119) perform better than JEA (TMR = 0.154). In particular, SIE and SimIE have significantly lower IMR's than JEA on certain items, namely item 5, 9 and 10. When the operational items are selected randomly, SIE and SimIE have more advantage over JEA. More precisely, JEA has higher IMR's on all items and therefore has higher TMR (0.523) than SIM (TMR = 0.338) and SimIE (TMR = 0.297). This is because the good performance of JEA requires accurate estimation of examinees' knowledge states. This is usually not true when the operational items are selected randomly in this simulation study. For the estimation of slipping and guessing parameters, the differences are tiny when the operational items are selected adaptively, while JEA yields significantly higher TMSE_s and TMSE_g when the operational items are selected randomly. In particular, when the items are selected randomly, JEA calibrates the slipping and guessing parameters of certain items (item 2, 4, 7 and 12) much worse than the other two methods. All three methods have comparable performances for other items.

We then compare the SIE and SimIE methods. For the adaptive selection of operational items, we observe that the SimIE method does improve the accuracy of estimation compared to SIE in terms of the calibration of **Q**-matrix. The improvement is obvious on items with low confidence indices. For items with high confidence indices, SimIE method has a comparable performance as the SIE method based on the IMR values. According to the TMR, SimIE (TMR = 0.119) is superior to SIE (TMR = 0.124). When the operational items are randomly selected, the improvement is more significant, because the calibration of the **Q**-matrix is less accurate in this case and there is more room for improvement. As we can see in Table 3, the IMR's of all the new items are reduced when the SimIE method is applied, and the improvement is more clear in this case. The TMR's of the two methods (0.338 versus 0.297) further imply that the SimIE method is preferred to the SIE method.

In addition, SimIE does improve the accuracy of the posterior mode in estimating examinees' knowledge states. This is reflected by the value of PMR (0.216 versus 0.196 in the case of adaptive selection of operational items and 0.721 versus 0.647 in the case of random selection of items). If the examinees are classified according to the posterior distributions of knowledge states only given the responses to operational items, 21.6% (72.1%) of the examinees in the population are misspecified, while if it is updated based on the responses to "good" items selected in the process of applying SimIE (also treating their estimated parameters as true), the misspecification rate is

reduced to 19.6% (64.7%). We point out that both SIE and JEA do not update the posterior distribution of examinees' knowledge states and the PMR values of SIE and JEA are only based on examinees' responses to the operational items. Therefore, the PMR values of SIE and JEA are the same in Tables 4 and 5. We also evaluate the classification of examinees' knowledge states in 1000 simulations by looking at the difference of the number of examinees misclassified according to the posterior distributions with and without the "good" new items. The histograms of the difference for both cases are plotted in Figure 1. For the adaptive selection of operational items, most of the time, the difference is positive, with the mode between 5 and 10. The trend is similar when the operational items are randomly selected but the difference is more significant. The "good" new items help to improve the estimation of examinees' knowledge states, even though their item parameters are unknown. Thus, the SimIE method performs better in the calibration of $\mathbf{Q}$-matrix.

Furthermore, as we can see from Tables 2, 3, 4, and 5, the RMSE's and TMSE of SimIE show improvement in the accuracy of the estimation of slipping and guessing parameters, although the improvement is not substantial.

Item	IMR			RMSE_s			RMSE_g		
	JEA	SIE	SimIE	JEA	SIE	SimIE	JEA	SIE	SimIE
1	0.23	0.23	0.21	0.08	0.10	0.09	0.04	0.04	0.04
2	0.00	0.00	0.00	0.05	0.04	0.04	0.05	0.04	0.04
3	0.45	0.40	0.40	0.10	0.13	0.12	0.04	0.04	0.04
4	0.00	0.00	0.00	0.07	0.05	0.05	0.04	0.03	0.03
5	0.16	0.05	0.04	0.07	0.06	0.06	0.07	0.05	0.05
6	0.32	0.33	0.32	0.12	0.14	0.13	0.03	0.03	0.03
7	0.00	0.00	0.00	0.06	0.04	0.04	0.07	0.04	0.04
8	0.00	0.02	0.01	0.05	0.04	0.04	0.06	0.05	0.05
9	0.11	0.00	0.00	0.07	0.05	0.05	0.07	0.04	0.04
10	0.52	0.44	0.42	0.13	0.12	0.12	0.04	0.04	0.04
11	0.05	0.02	0.02	0.05	0.05	0.05	0.06	0.05	0.05
12	0.01	0.00	0.00	0.08	0.06	0.06	0.04	0.03	0.03

Table 2: The IMR, RMSE values of JEA, SIE and SimIE and the operational items are selected according to Shannon entropy in Study 1

3.2 Study 2

3.2.1 Basic settings

The purpose of this study is to evaluate the performance of JEA, SIE and SimIE under different sample sizes. All the basic settings in this study are the same as those in Study 1, except that data sets of different sample sizes (100, 200, 400, 800, and 1600) are generated. For each sample size,

Item	IMR			$RMSE_s$			$RMSE_g$		
	JEA	SIE	SimIE	JEA	SIE	SimIE	JEA	SIE	SimIE
1	0.72	0.57	0.48	0.13	0.16	0.15	0.06	0.06	0.05
2	0.20	0.05	0.03	0.17	0.07	0.06	0.20	0.07	0.07
3	0.83	0.74	0.70	0.15	0.19	0.18	0.05	0.06	0.06
4	0.36	0.10	0.07	0.23	0.09	0.09	0.10	0.05	0.04
5	0.60	0.33	0.27	0.09	0.13	0.12	0.13	0.08	0.08
6	0.76	0.66	0.61	0.18	0.19	0.18	0.04	0.05	0.05
7	0.15	0.04	0.02	0.17	0.06	0.06	0.20	0.07	0.06
8	0.22	0.21	0.13	0.11	0.08	0.07	0.13	0.09	0.08
9	0.59	0.17	0.13	0.11	0.11	0.09	0.16	0.07	0.07
10	0.84	0.76	0.70	0.17	0.16	0.16	0.04	0.06	0.05
11	0.42	0.26	0.24	0.09	0.11	0.11	0.14	0.09	0.09
12	0.43	0.14	0.10	0.24	0.10	0.10	0.10	0.05	0.05

Table 3: The IMR, RMSE values of JEA, SIE and SimIE and the operational items are selected randomly in Study 1

Method	TMR	$TMSE_s$	$TMSE_g$	PMR
JEA	0.154	0.080	0.052	0.216
SIE	0.124	0.081	0.042	0.216
SimIE	0.119	0.080	0.042	0.196

Table 4: The TMR, TMSE, PMR values of JEA, SIE and SimIE and the operational items are selected according to Shannon entropy in Study 1

we generate 400 data sets independently. Items are selected adaptively according to the Shannon entropy method. The rest of the setting is the same as in Study 1.

3.2.2 Results

The results of Study 2 are shown in Table 6, Figures 2, 3, and 4. As we can see in Table 6 and Figure 2, for all methods, the calibration becomes more accurate as the sample size increases. More precisely, for each method and for each item, the IMR value decreases as the sample size increases. As a consequence, the TMR values have the same pattern. From Table 6, Figures 3 and 4, for all three methods, the estimation of the slipping and guessing parameters also becomes more accurate as the sample size becomes larger. For example, when the sample size is 100, $TMSE_s$ ($TMSE_g$) is around 0.14 (0.09) for JEA and around 0.20 (0.10) for SIE and SimIE; these values are large compared to the true parameters that are uniform in (0.1, 0.4). When the sample size increases to 1600, $TMSE_s$ ($TMSE_g$) decreases to 0.05 (0.03) for JEA and 0.03 (0.02) for SIE and SimIE; these values are relatively small compared to the true parameters.

In addition, based on the TMR values, the simulation results show that JEA performs better in the calibration of **Q**-matrix when the sample size is 100 and the other two methods perform

Method	TMR	TMSE_s	TMSE_g	PMR
JEA	0.523	0.157	0.126	0.721
SIE	0.338	0.129	0.067	0.721
SimIE	0.297	0.120	0.063	0.647

Table 5: The TMR, TMSE, PMR values of JEA, SIE and SimIE and the operational items are selected randomly in Study 1

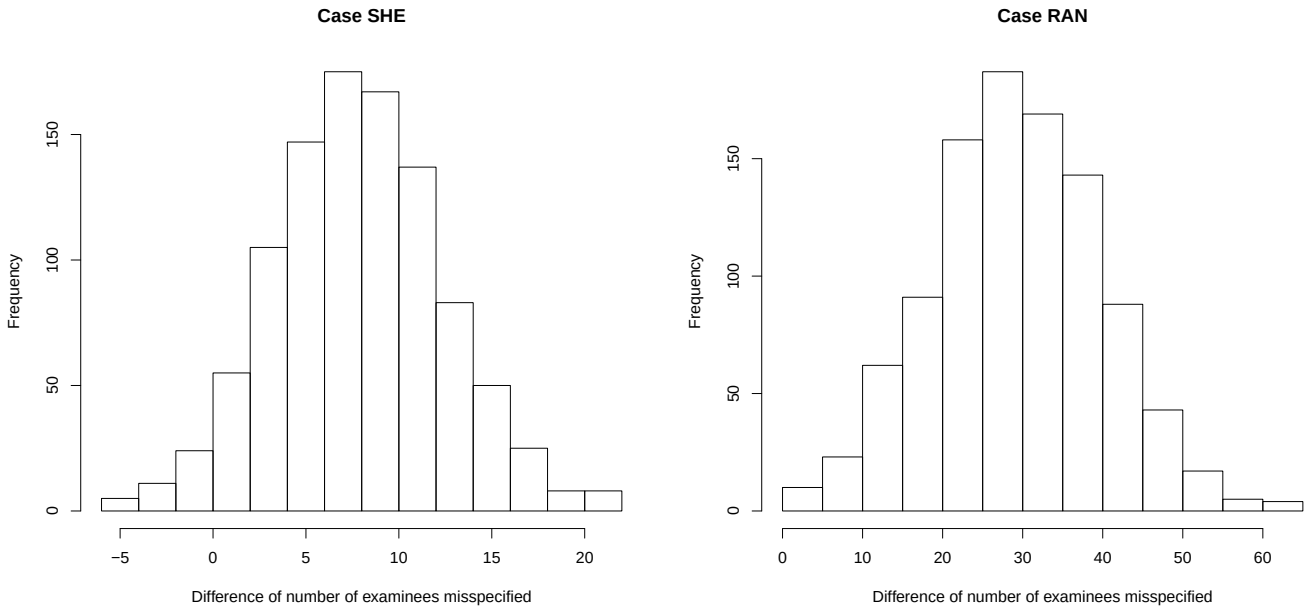

Figure 1: Histogram of difference of the number of misspecification before and after updating the posterior distributions using SimIE method when the operational items are selected according to Shannon entropy (Left) and when the operational items are selected randomly (Right) in Study 1

better when the sample size is greater than or equals to 200. In addition, as sample size increases, the advantage of SIE and SimIE over JEA becomes more significant. For the calibration of the slipping parameters, according to the TMSE values, JEA performs better when the sample size is less than or equal to 200 and the other two methods perform better when the sample size is greater or equal to 800. As for the guessing parameters, SIE and SimIE are superior to JEA except when the sample size is 100.

Furthermore, under all sample sizes except when $N = 100$, SimIE is superior to SIE in the calibration of $\mathbf{Q}$-matrix and the slipping and guessing parameters. In particular, when the sample size is relatively large, both methods yield accurate estimation and 400 simulation trials may not be enough to reflect the difference between them.

Sample size	TMR			$TMSE_s$			$TMSE_g$		
	JEA	SIE	SimIE	JEA	SIE	SimIE	JEA	SIE	SimIE
100	0.41	0.43	0.43	0.14	0.20	0.20	0.09	0.10	0.10
200	0.26	0.25	0.24	0.11	0.13	0.13	0.07	0.07	0.07
400	0.16	0.12	0.12	0.08	0.08	0.08	0.05	0.04	0.04
800	0.06	0.04	0.04	0.06	0.05	0.05	0.04	0.03	0.03
1600	0.02	0.00	0.00	0.05	0.03	0.03	0.03	0.02	0.02

Table 6: The TMR, TMSE values of JEA, SIE and SimIE under different sample sizes in Study 2

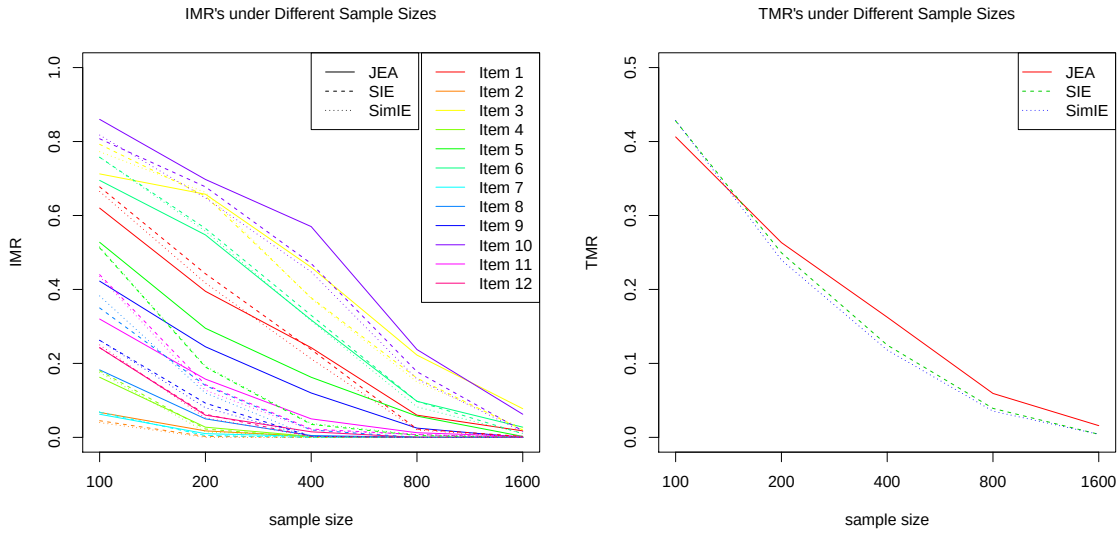

Figure 2: The IMR's for 12 new items of JEA, SIE and SimIE method under 5 different sample sizes (Left) and the TMR's of JEA, SIE and SimIE method under 5 different sample sizes (Right).

4 Conclusion and further discussions

In this paper, we propose new item calibration methods for the **Q**-matrix, a key element in cognitive diagnosis, as well as the slipping and guessing parameters. These methods extend the work of Chen et al. (2012) and are compared with the joint estimation algorithm proposed in Chen and Xin (2011). Under the setting of Study 1, the results show that the proposed SIE and SimIE methods perform better than the JEA method in the calibration of the **Q**-matrix as well as the estimation of slipping and guessing parameters. In addition, JEA is sensitive to the accuracy of the estimation of examinees' knowledge states. The simulation results in Study 1 also show that the SimIE method is superior to the SIE method for the calibration of the **Q**-matrix as well as the estimation of slipping and guessing parameters. Since all three methods can be implemented without much computational burden, the SimIE method is therefore preferred. From the results of Study 2, all three methods tend to estimate the item parameters more accurately as the sample size

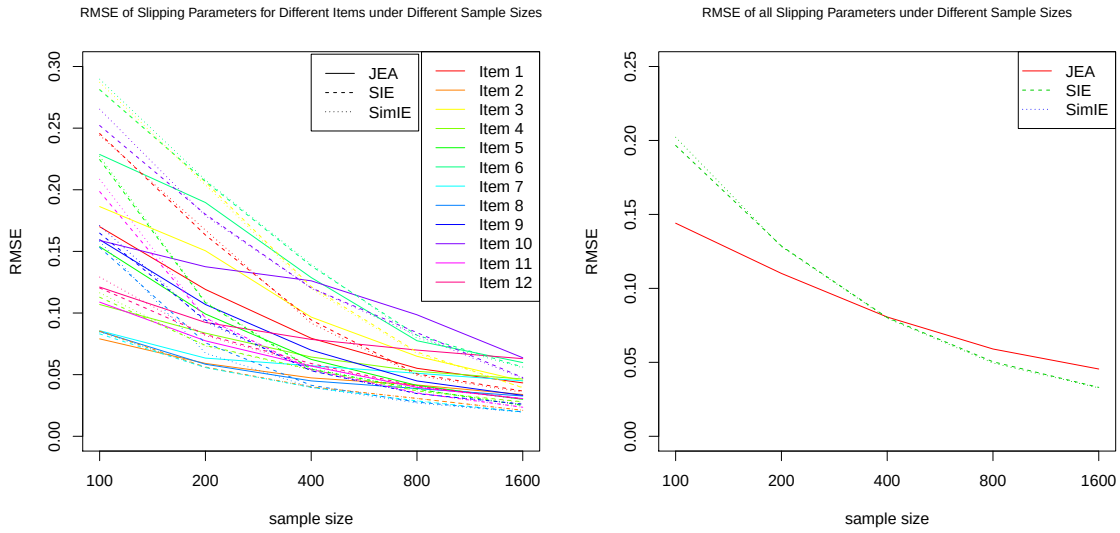

Figure 3: The itemwise RMSE's for slipping parameter for 12 new items of JEA, SIE and SimIE method under 5 different sample sizes (Left) and the overall RMSE's for slipping parameter of JEA, SIE and SimIE method under 5 different sample sizes (Right).

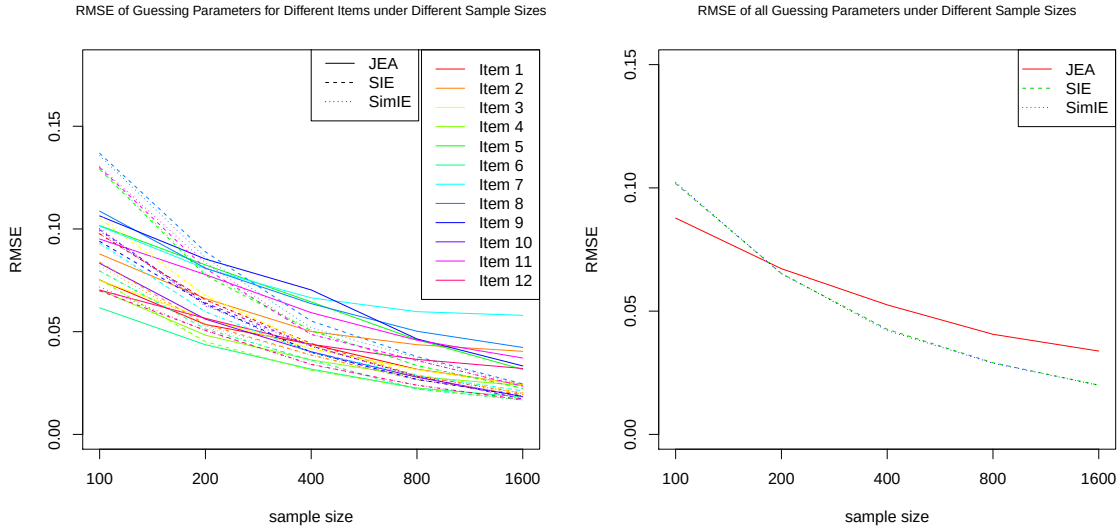

Figure 4: The itemwise RMSE's for guessing parameter for 12 new items of JEA, SIE and SimIE method under 5 different sample sizes (Left) and the overall RMSE's for guessing parameter of JEA, SIE and SimIE method under 5 different sample sizes (Right).

becomes larger. In particular, when the sample size is 1600, under our simulation setting, both the SIE and SimIE methods correctly calibrate the $\mathbf{Q}$-matrix with probability close to 1, and estimate the slipping and guessing parameters with an acceptable accuracy (based on the RMSE values).

Furthermore, we introduce the confidence index η to evaluate the goodness-of-fit for a new item.

When $\mathbf{Q}$ is specified, the estimation accuracy of the slipping and guessing parameters is quantified by the observed Fisher information based on the likelihood function. When $\mathbf{Q}$ is also unknown, the confidence index plays a similar role of the observed Fisher information, since $\mathbf{Q}$ is discrete. Thus, the index itself is of interest in online calibration and, along with the observed Fisher information of the slipping and guessing parameters, summarizes the estimation accuracy of item parameters for a new item. Based on it, a decision may be made as to whether or not the calibration is sufficiently accurate.

There are a number of theoretical issues which require attention. For instance, under what circumstances can the $\mathbf{Q}$-matrix of the new items be consistently estimated? When can the slipping and guessing parameters be consistently estimated? We provide a brief discussion on this issue. Given a known $\mathbf{Q}$-matrix, the identifiability of the slipping and the guessing parameters can be checked by computing the Fisher information with respect to these two parameters. Then, the most important and interesting task is to ensure the identifiability of the $\mathbf{Q}$-matrix. Generally speaking, in order to consistently calibrate all possible $\mathbf{Q}$-matrices, we typically require the following knowledge state patterns exist in the population. For each dimension of the knowledge state α^l, there exist a nonzero proportion of examinees who only master α^l and do not master any other skills. We use $\mathbf{e}_l = (0, ..., 0, 1, 0, ..., 0)$ to denote such a knowledge state vector. Missing one or a few such kind of $\mathbf{e}_l$'s will affect the identification of certain patterns (not all) of $\mathbf{Q}$-matrix.

The above discussion assumes complete and accurate specification of the knowledge state of each examinee. Under the current setting, the knowledge states are not directly observed and are estimated through the responses to the operational items. Therefore, an important issue is the selection of operational items through which enough information about the knowledge states can be obtained. We would like to emphasize that the number of operational items responded to by each examinee is limited. Therefore, we do not require (and it is not necessary) that the knowledge state of each examinee is identified very accurately. On the other hand, we do require the number of examinees to be reasonably large. Even if each of them provides a small amount of information, the new items eventually can be calibrated accurately with a sufficiently large number of people. An important issue for future study is to clarify requirements on the operational items to ensure the consistent calibration of the new item.

We also observe from both simulation studies that the calibration accuracy varies for different $\mathbf{Q}$'s. For example, in Study 1, we observe that the estimation accuracy (of $\mathbf{Q}$) varies for different items. There are at least two aspects affecting the estimation accuracy. The first is the specific value of the slipping and guessing parameters; generally, the smaller the slipping and guessing parameters are, the easier it is to calibrate the $\mathbf{Q}$. This is intuitively easy to understand, because

the slipping and guessing behavior introduces noise which makes the signal (the $\mathbf{Q}$ pattern) harder to recover. The second aspect is related to the knowledge state population. For example, if we look at new items 5 and 6 from Study 1, although item 5 has greater slipping and guessing parameters than item 6, the calibration of $\mathbf{Q}$ for item 5 is much better than that of item 6 according to the corresponding IMR values in Table 2 and 3. Note that $q_5 = (0, 1, 0, 0, 0)$ and $q_6 = (1, 0, 1, 0, 1)$. Considering the way we generate the population, almost half of the examinees are capable of solving item 5, while only 12.5% examinees are capable of solving item 6. In other words, for item 5, the examinees who are able to solve it and those who are not able to are balanced, while this is not the case for item 6. Naturally, this leads to a design problem for how to adaptively assign new items to examinees according to both the current calibration of the new items and our current measurement of the examinees. This becomes extremely important under the situation that the number of examinees is also limited and we'd like to optimize the calibration of all new items.

The current calibration procedure was developed under the DINA model. It is worth pointing out that this method can be extended without difficulty to other core DCMs, such as DINO (deterministic input, noisy-or-gate), NIDA (noisy inputs, deterministic-and-gate), NIDO (noisy input, deterministic-or-gate) model and so on (see Rupp and Templin (2010)). To understand this, we can view core DCMs as special cases of log-linear models and latent classes and different constraints on model parameters Henson et al. (2009). When calibrating a single item under a log-linear model with latent classes, step 2 in the SIE procedure does not change. More specifically, once the auxiliary model parameters are profiled out, $\hat{q}_j$ is obtained by finding the $\mathbf{Q}$ that has the maximum profile likelihood. Step 1 may vary because the EM algorithm may not be realistically feasible for some models. However, the MCMC approach can be applied to estimate auxiliary model parameters when the EM algorithm is not feasible, although it is slower than the EM algorithm; see Chapter 11, Rupp and Templin (2010). Furthermore, the SimIE method can also be generalized to other core DCM's.

The current calibration procedure works under a fixed and known dimension for the latent classes. In practice, a new exam problem, though designed to measure the same set of attributes as the operational items, may possibly be related to additional new attributes. To incorporate this new structure, more column(s) would need to be added to the existing $\mathbf{Q}$-matrix. Another instance that would necessitate additional dimensions is as follows. Suppose that all the operational items require some attribute. Correspondingly, there is one column in $\mathbf{Q}$ containing all ones. Such columns are usually removed and the absence of such an attribute is ascribed to the slipping parameter. If a new item does not need this extra attribute required by all the operational items, then the removed column should be restored to maintain the correctness. In addition, the slipping

and guessing parameters of the operational items need to be recalibrated according to this new **Q**-matrix; in particular, part of the slipping probability is explained by the absence of this extra attribute. Thus, a testing mechanism needs to be developed, so as to determine whether or not an extra dimension should be added to the existing **Q**-matrix during the course of online calibration.

References

Ban, J., Hanson, B., Wang, T., Yi, Q., and Harris, D. (2001). A comparative study of on-line pretest itemcalibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, 38(3):191–212.

Ban, J., Hanson, B., Yi, Q., and Harris, D. (2002). Data sparseness and on-line pretest item calibration-scaling methods in CAT. *Journal of Educational Measurement*, 39(3):207–218.

Chen, P. and Xin, T. (2011). Item replenishing in cognitive diagnostic computerized adaptive testing. In *Annual Meeting of the National Council on Measurement in Education, New Orleans*.

Chen, P., Xin, T., Wang, C., and Chang, H.-H. (2012). Online calibration methods for the dina model with independent attributes in CD-CAT. *Psychometrika*, 77(2):201–222.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74(4):619–632.

Chiu, C., Douglas, J., and Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, 74(4):633–665.

DiBello, L., Stout, W., and Roussos, L. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques. *Cognitively diagnostic assessment. Hillsdale, NJ: Erlbaum*, pages 361–390.

Haertel, E. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*.

Hartz, S. (2002). A bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality. *Doctoral Dissertation, University of Illinois, Urbana-Champaign*.

Henson, R., Templin, J., and Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2):191–210.

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

Leighton, J. P., Gierl, M. J., and Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on tatsuoka's rule-space approach. *Journal of Educational Measurement*, 41:205–237.

Liu, J., Ying, Z., and Zhang, S. (2013). A Rate Function Approach to the Computerized Adaptive Testing for Cognitive Diagnosis. Unpublished.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Erlbaum, Hillsdale.

Lord, F. M. (1971). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement*, 31(1):3–31.

Owen, R. J. (1975). Bayesian sequential procedure for quantal response in context of adaptive mental testing. *Journal of the American Statistical Association*, 70(350):351–356.

Rupp, A. and Templin, J. (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.

Rupp, A. and Templin, J. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspective*, 6:219–262.

Rupp, A. and Templin, J. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Segall, D. (2003). Calibrating CAT pools and online pretest items using MCMC methods. In *Annual Meeting of the National Council on Measurement in Education, Chicago, IL*.

Stocking, M. (1988). Scale drift in on-line calibration. Technical report, DTIC Document.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Applied Statistics (JRSS-C)*, 51:337–350.

Tatsuoka, K. (1983). Rule space: an approch for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20:345–354.

Tatsuoka, K. (2009). *Cognitive assessment: an introduction to the rule space method*. CRC Press.

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

Thissen, D. and Mislevy, R. (2000). Testing algorithms. In et al., H. W., editor, *Computerized adaptive testing: A primer*, pages 101–133. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

van der Linden, W. (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika*, 63:201–216.

van der Linden, W. and Glas, C. (2000). *Computerized adaptive testing: Theory and practice*. Springer.

von Davier, M. (2005). *A general diagnosis model applied to language testing data*. Educational Testing Service, Research Report.

Wainer, H., Dorans, N., Green, B., Steinberg, L., Flaugher, R., Mislevy, R., and Thissen, D. (1990). *Computerized adaptive testing: A primer*. Lawrence Erlbaum Associates.

Wainer, H. and Mislevy, R. (1990). Item response theory, item calibration, and proficiency estimation. In et al., H. W., editor, *Computerized adaptive testing: A primer*, pages 65–102. Lawrence Erlbaum Associates, Hillsdale, NJ.

Xu, X., Chang, H., and Douglas, J. (2003). A simulation study to compare CAT strategies for cognitive diagnosis. In *Annual Meeting of the American Educational Research Association, Chicago*.