Behavior Research Methods (2024) 56:233–257
https://doi.org/10.3758/s13428-022-02036-7

An iterative two-step method for online item calibration in CD-CAT

Xiaofeng Yu[1] · **Ying Cheng**[2]

Accepted: 24 November 2022 / Published online: 29 December 2022

Abstract
The development and maintenance of the item bank is a critical element to a CD-CAT (cognitive diagnostic computerized adaptive testing; Cheng, 2009) system. For continuous testing, it is important to replenish the item bank with new items that have been calibrated. This requires pretesting to estimate the parameters of the new items. For CD-CAT, the structural parameters that need to be estimated include both item parameters and attribute vectors. In this paper, we propose three residual-statistic-based methods: RMA, ROEM, and RMEM, to estimate the attribute vectors and item parameters all together for new items. An iterative two-step online calibration procedure is developed to estimate the attribute vectors for the new items in the first step, and estimate the item parameters in the second step, then proceed iteratively until convergence is reached. An extensive simulation study was conducted to evaluate the performance of the three proposed methods and compare them with two existing methods, namely the Joint Estimation Algorithm (JEA; Chen & Xin, 2011) and Single Item Estimation (SIE; Chen et al., 2015) methods. In terms of the estimation of the attribute vector, the RMEM method performs the best in most of the cases. In terms of item parameter estimation, RMEM still has some advantages, and RMA outperforms JEA and SIE. Taken together, results suggest that the RMEM is superior to the other methods, especially when sample size is relatively small. A real-data example is provided to illustrate the application of RMEM in practice.

Keywords CD-CAT · Residual · Online calibration · DINA model

Cognitive diagnostic computerized adaptive testing (CD-CAT; Cheng, 2009; McGlohen & Chang, 2008) is computerized adaptive testing (CAT) built upon a cognitive diagnostic model (CDM; Rupp & Templin, 2008; Rupp et al., 2010). Cognitive diagnostic models (CDMs) are considered important statistical tools that link item responses to latent cognitive profiles, which capture the strengths and weaknesses of each respondent in terms of their mastery of discrete knowledge points or attributes. Hence, testing programs built on CDMs have both features of model-based measurement and formative assessment (Embretson, 2001).

In a typical adaptive testing system, items are sequentially selected from an item bank, tailored to each respondent according to certain item selection rules, for example, maximizing test information or minimizing the standard error of measurement of the latent trait. In CD-CAT, the goal is to *efficiently* estimate the latent cognitive profiles by sequentially choosing the most suitable items for each candidate (Cheng, 2009; Dai et al., 2016; Yu et al., 2019; Zheng & Chang, 2016; Zheng & Wang, 2017). Given a well-designed item bank, continuous testing can be offered through CD-CAT, which means that efficient formative assessment can be provided to students continuously.

In real applications, any CAT systems that offer continuous testing need to replenish their item banks periodically. This is because repeated use of items may pose a risk to test security and validity. Therefore, retiring flawed, obsolete, or overexposed items and replacing them with new items that have been calibrated, a process called item replenishment, is important for continuous testing (Chen et al., 2012; Chen et al., 2015; Chen & Xin, 2011; Ren et al., 2017). For this reason, new items constantly need to be developed, reviewed, and calibrated for CAT programs.

Online calibration in CAT refers to estimating the parameters of new items that are administered to respondents during the course of their operational testing along with previously calibrated items (Wainer & Mislevy, 2000). Ren et al. (2017) pointed out several main advantages of online calibration. First, new items are calibrated under the exact same

✉ Ying Cheng
ycheng4@nd.edu

[1] Jiangxi Normal University, School of Psychology, Nanchang, China

[2] University of Notre Dame, Department of Psychology, Notre Dame, IN, USA

condition as for their future operational use. Second, the item parameters of the new items are calibrated on the same scale as the operational items, which means linking or rescaling is no longer required. Commonly used methods that have been proposed to calibrate new items include Method A and Method B (Stocking, 1988), marginal maximum likelihood estimation with one expectation maximization (OEM) iteration (Wainer & Mislevy, 2000), and marginal maximum likelihood estimation with multiple EM (MEM) iterations (Ban et al., 2001; Ban et al., 2002).

New items for CD-CAT need to be calibrated in terms of both item parameters and the attribute vectors. In contrast, in traditional CAT, item calibration only refers to the estimation of item parameters. Thus, it is even more challenging to conduct online item calibration for CD-CAT than regular CAT. Chen et al. (2012) considered the online calibration of only the item parameters in CD-CAT and proposed three methods, namely Cognitive Diagnostic-Method A (CD-MA), Cognitive Diagnostic-One EM cycle (CD-OEM), and Cognitive Diagnostic-Multiple EM cycle (CD-MEM). These methods assume known attribute vectors and are analogs to methods described in the preceding paragraph. For online calibration of both item parameters and attribute vectors, literature is relatively scarce. Chen and Xin (2011) proposed a joint estimation algorithm (JEA), which considered jointly estimating the attribute vectors and the item parameters based on the DINA (Deterministic Input, Noisy output "AND" gate; see Junker & Sijtsma, 2001; de la Torre, 2009) model. Their results indicated the JEA can have a promising performance. Chen et al. (2015) considered two Bayesian variations of JEA: the SIE (Single Item Estimation), and the SimIE (Simultaneous Item Estimation) method. As their names suggest, in SIE a single new item is calibrated at a time, while in SimIE multiple new items are calibrated at a time. With a sample size larger than 800, Chen et al. (2015) showed that SIE and SimIE methods perform better than the JEA method in the estimation of both attribute vectors and the item parameters. Due to their iterative nature, SIE and SimIE showed very similar performances in estimating attribute vectors and item parameters. For all three methods, JEA, SIE, or SimIE, the estimation of the item parameters is highly dependent on the estimation of the attribute vectors. However, if the sample size is relatively small (e.g., 400 or fewer), item parameters cannot be estimated well even with known attribute vectors, let alone with unknown attribute vectors (Chen et al., 2015).

Given the limitations of existing methods, in this paper we propose an iterative two-step procedure to estimate both attribute vectors and item parameters with relatively small sample sizes. First, we propose to use a residual-based statistic to estimate the attribute vectors in the context of CD-CAT. This step does not require known or precisely estimated item parameters. In the second step, we treat the estimated attribute vector as true, and estimate the item parameters based on CD-Method A, CD-OEM, or CD-MEM. The procedure proceeds iteratively until convergence is reached.

The rest of this paper is organized as follows. First, we provide a literature review for the existing methods on this topic, which involves two main lines of research: online calibration of the item parameters only, and online calibration of both the item parameters and attribute vectors. Next, we introduce in detail a new method of attribute vector estimation using a residual-based statistic, and the iterative two-step procedure for estimating both item parameters and attribute vectors. A simulation study to assess the performance of the proposed estimation methods is then described. A real-data analysis is provided to illustrate the application of RMEM in practice. Discussions and implications of the results are given in the last section.

Online calibration methods in CD-CAT

In this section, we briefly review several existing methods. For the sake of convenience but without loss of generality, we first introduce the following terms and notations that will be used throughout the remainder of the paper. As discussed earlier, new items refer to the items whose attribute vectors and item parameters are unknown, in contrast to the operational items that have been previously calibrated in the item bank. Let's assume an existing item bank with J operational items. Meanwhile, the item parameters and attribute vectors of M new items need to be estimated. Consider a CD-CAT that targets a total of K attributes. Each of the J operational items require a distinct subset of the K attributes (denoted as $\mathbf{q}_j, j = 1, 2, \ldots, J$) for them to be answered correctly. The stacked $\mathbf{q}_j$s form the item-attribute associations matrix for the item bank, namely the $\mathbf{Q}$-matrix, which is a binary $J \times K$ matrix. The $\mathbf{Q}$-matrix for the m new items is denoted as $\mathbf{Q}_{new}$. The mastery status of each of N test takers is captured by $\boldsymbol{\alpha}_i$ ($i = 1, 2, \ldots, N$), the attribute mastery pattern vector or, AMP. L refers to the fixed test length, and a $N \times L$ matrix $\mathbf{X}$ denotes the item response matrix with its binary element X_{ij}, with $X_{ij} = 1$ indicating a correct response of test taker i on item j, and $X_{ij} = 0$ an incorrect response. Let n_m be the total number of respondents responding to the m^{th} new item.

As a parsimonious and popular CDM model, the DINA model is used here as an example (de la Torre, 2009). An expected or ideal response under the DINA model is characterized by an indicator variable, denoted as $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$, which is used to indicate whether the i^{th} respondent possesses all the required attributes of the j^{th} item or not. Unexpected responses are accounted for by the slipping and guessing parameter, where $s_j = P(X_{ij} = 0 | \eta_{ij} = 1)$ and $g_j = P(X_{ij} = 1 | \eta_{ij} = 0)$, respectively. The probability of a

correct response to the j^{th} item by the i^{th} respondent under the DINA model is therefore defined as

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

For a new item m, its attribute vector $\mathbf{q}_m$ and item parameters (s_m, g_m) are of key interest in online calibration.

Online calibration of item parameters

The following three methods are based on the assumption that the attribute vectors of the new items are known (i.e., $\mathbf{q}_m$'s are available, perhaps through content experts who label each item for the attributes they measure), and only their item parameters need to be estimated.

CD-Method A For a new item m, suppose that there are n_m respondents responding to the item. The CD-method A treats the estimated AMP $\hat{\boldsymbol{\alpha}}_i$ as the true $\boldsymbol{\alpha}_i$, which was obtained based on the operational items answered by the i^{th} respondent, then estimates the slipping and guessing parameters through maximum likelihood (de la Torre, 2009).

$$\frac{\partial l_m}{\partial \mathrm{s}_m}=0, \tag{2}$$

$$\frac{\partial l_m}{\partial g_m}=0, \tag{3}$$

where $l_m(\mathbf{x}_i|\mathbf{q}_m,s_m,g_m)=\log\left(\prod_{i=1}^{n_m}P_{s_m,g_m}(\mathbf{q}_m,\hat{\boldsymbol{\alpha}}_i)^{x_{im}}\left[1-P_{s_m,g_m}(\mathbf{q}_m,\hat{\boldsymbol{\alpha}}_i)\right]^{1-x_{im}}\right)$ is the log-likelihood function, and $\mathbf{q}_m$ is the attribute vector for item m. x_{im} refers to the score of the m^{th} new item answered by the respondent i (0/1), and $P_{s_m,g_m}\left(\mathbf{q}_m,\hat{\boldsymbol{\alpha}}_i\right)$ refers to the response probability to new item m under the DINA model evaluated at $\hat{\boldsymbol{\alpha}}_i$.

CD-OEM. CD-OEM applies a single cycle of an EM algorithm (Chen et al., 2012; de la Torre, 2009) to estimate the item parameters for each new item. For the m^{th} new item, based on the posterior distribution of the AMPs, the CD-OEM method considers one E-step obtaining the expected proportion of respondents who have AMP $\hat{\boldsymbol{\alpha}}_v$ among those who answer the new item m, where $\hat{\boldsymbol{\alpha}}_v$ refers to one of the 2^K possible attribute profiles and $\sum_{v=1}^{2^K}P_m\left(\hat{\boldsymbol{\alpha}}_v\right)=1$. Next, the M-step finds the $\hat{s}_m$ and $\hat{g}_m$ that maximize the logarithm of the corresponding expected likelihood.

CD-MEM By allowing multiple EM cycles, the CD-OEM becomes the CD-MEM. The first EM cycle in CD-MEM is the same as in the CD-OEM method, and the obtained item parameters and attribute vectors are regarded as the initial values of the second EM cycle. The CD-MEM method utilizes both the responses of operational items and new items to calculate the posterior distribution of the AMPs for the E-step from the second EM cycle onward, then fixes the item parameters of the operational items, and adopts the same M-step as that of the CD-OEM method to update the item parameters of the new items (refer to Chen et al., 2012 for further details). The EM cycles are repeated till a stop criterion is met.

Results of Chen et al. (2012) showed that CD-Method A, CD-OEM, and CD-MEM are able to recover item parameters accurately with large sample sizes, and CD-Method A performs the best when the items have smaller slipping and guessing parameters, but its performance is largely affected by the item parameter magnitude.

Online calibration of both item parameters and attribute vectors

The Joint Estimation Algorithm (JEA) Based on the DINA model, Chen and Xin (2011) proposed the JEA to jointly estimate both the attribute vectors and the item parameters, which is the analog of the joint maximum likelihood estimation (JMLE; Baker & Kim, 2004) method in item response theory (IRT). As the extension of CD-Method A, the JEA treats the AMPs estimated from operational items as true, and then estimates the item parameters and the attribute vectors for the new items, one item at a time. For the m^{th} new item, the JEA maximizes $l_m(\mathbf{q}_m,s_m,g_m)$ with respect to $\mathbf{q}_m$ given (s_m,g_m), then consider the estimated $\mathbf{q}_m$ as true and optimizes $l_m(\mathbf{q}_m,s_m,g_m)$ with respect to (s_m,g_m). This is done iteratively until convergence is reached. Convergence can be defined as a very small difference of the log-likelihood between one iteration and the next.

To account for the uncertainty of the estimated AMPs, the SIE and SimIE are two Bayesian versions of the JEA.

The Single Item Estimation Method (SIE) Instead of plugging in the estimates of the AMPs of the respondents who answered the m^{th} new item, the SIE method considers the expected log likelihood

$$\begin{aligned}&\mathrm{E}\left(l_m(\mathbf{x}_i|\mathbf{q}_m,s_m,g_m)\right)\\&=\sum_{i=1}^{n_m}\sum_{\boldsymbol{\alpha}_i}\pi_i(\boldsymbol{\alpha}_i;s_m,g_m)\left[x_{im}\log P_{s_m,g_m}(\mathbf{q}_m,\boldsymbol{\alpha}_i)+(1-x_{im})\left(1-\log P_{s_m,g_m}(\mathbf{q}_m,\boldsymbol{\alpha}_i)\right)\right],\end{aligned} \tag{4}$$

where $\pi_i(\boldsymbol{\alpha}_i;s_m,g_m)$ is the posterior distribution of $\boldsymbol{\alpha}_i$ based on the operational items (in the first EM cycle), or both the operational items and new items (in the remaining EM cycles). By doing so, SIE takes the uncertainty of $\hat{\boldsymbol{\alpha}}_i$ into account. The SimIE further considers calibrating multiple new items at a time.

The Simultaneous Item Estimation Method (SimIE) As noted by Chen et al. (2015), the more accurate the information about the AMP is, the better the calibration will be. Therefore, the motivation of the SimIE is to borrow some useful information from the new items to improve the estimation of

the unknown AMPs. However, borrowing information from those inadequately calibrated items may have a detrimental effect on the estimation of AMPs. In order to address this issue, Chen et al. (2015) proposed an index, here denoted as ω_m (denoted as η_j in the original paper, but as ω_m here to avoid confusion), to evaluate the confidence in the fit of $\hat{\mathbf{q}}_m$. ω_m was defined as the difference between the log-likelihood function for the two most probable $\hat{\mathbf{q}}'_m s$ for the m^{th} item. Half of the 95th percentile of the χ^2 distribution with one degree of freedom, i.e., 1.92, was chosen as the empirical cutoff for the "good" new items in Chen et al. (2015). Then treating the first chosen new item, which has the maximum ω_m and $\omega_m > 1.92$, as an additional operational item, SimIE updates the posterior distribution of the AMP of the respondents based on all operational items, and recalibrates the second chosen new item. This process is repeated until all the chosen new items are treated as additional operational items. Then, new items which are not selected in the preceding step are calibrated one at a time. This is one estimation cycle. The algorithm proceeds until the chosen items do not change in two consecutive cycles.

Attribute vector estimation based on a residual-based statistic

In this section, we first briefly introduce the residual-based statistic (please refer to Yu and Cheng (2020) for more details) to measure the appropriateness of the attribute vector of an item. Then we present the theoretical proof that under the DINA model, the proposed residual-based statistic can be used to identify the true attribute vector of the m^{th} new item with arbitrarily chosen item parameters under certain assumptions. This may help liberate the dependency on large sample size for existing methods.

Let $E(X_{im}|\boldsymbol{\alpha}_i)$ be the expected score for the i^{th} respondent with AMP $\boldsymbol{\alpha}_i$, and $P(X_{im}=x_{im}|\boldsymbol{\alpha}_i)$, denoted by $P(x_{im}|\boldsymbol{\alpha}_i)$ for short, be the probability for the respondent obtaining score x_{im}, x_{im} being 0 or 1. Then the appropriateness index of the attribute vector for the m^{th} item can be defined as

$$R_m(\boldsymbol{\alpha}, \mathbf{q}_m, s_m, g_m) = \sum_{i=1}^{n_m} \log\left[\frac{x_{im} - E(X_{im}|\boldsymbol{\alpha}_i)}{P(x_{im}|\boldsymbol{\alpha}_i)}\right]^2, or \sum_{i=1}^{n_m} \log\left|\frac{x_{im} - E(X_{im}|\boldsymbol{\alpha}_i)}{P(x_{im}|\boldsymbol{\alpha}_i)}\right|, \quad (5)$$

where $\boldsymbol{\alpha}$ is a matrix of vertically stacked $\boldsymbol{\alpha}'_i$s, i.e., attribute profiles of those respondents who answered the m^{th} new item. The squared form is numerically two times the absolute form, so the performance of the method based on these two forms are equivalent. The squared form will be used in all our simulation conditions just for coding consistency. Under the DINA model, according to the values of η_{im} and the response x_{im}, each respondent is classified into one of the four groups, G_1, G_2, G_3 and G_4, where respondents in G_1 have $\eta_{im}=1$ and $x_{im}=1$, respondents in G_2 have $\eta_{im}=1$ and $x_{im}=0$, respondents in G_3 have $\eta_{im}=0$ and $x_{im}=1$, respondents in G_4 have $\eta_{im}=0$ and $x_{im}=0$, respectively. Hence, formula 5 can be expanded to

$$R_m(\boldsymbol{\alpha}, \mathbf{q}_m, s_m, g_m) = 2\sum_{i=1}^{n_m} \log \left[\eta_{im}\left(\frac{s_m}{1-s_m}\right)^{x_{im}}\left(\frac{1-s_m}{s_m}\right)^{1-x_{im}} + (1-\eta_{im})\left(\frac{g_m}{1-g_m}\right)^{1-x_{im}}\left(\frac{1-g_m}{g_m}\right)^{x_{im}}\right], \quad (6)$$

where $\eta_{im} = \prod_{k=1}^{K} \alpha_{ik}^{q_{mk}}$ is the ideal response of the i^{th} examine (with attribute profile $\boldsymbol{\alpha}_i$) to the m^{th} item (with attribute vector $\mathbf{q}_m$). We expect that given $\hat{\boldsymbol{\alpha}}$ from operational items, $R_m(\hat{\boldsymbol{\alpha}}, \mathbf{q}_m, s_m, g_m)$ as a function of $\mathbf{q}_m$ is minimized when $\mathbf{q}_m$ is at its true value.

Theorem 1. Consider an infinite sample, that is $N \to \infty$, and the true item parameters $s_m, g_m \in (0, 0.5)$. Denote $\hat{\boldsymbol{\alpha}}$ as the estimate of $\boldsymbol{\alpha}$. Furthermore, assume its true value $\boldsymbol{\alpha}$ is known in advance. Given the provisional item parameters for the m^{th} item as (s_m^0, g_m^0), where s_m^0, g_m^0 are two arbitrarily chosen real numbers within the range of $(0, 0.5)$, and denote $\hat{R}_m^0(\boldsymbol{\alpha}, \mathbf{q}_m, s_m^0, g_m^0)$ as the value of the residual-based statistic evaluated at (s_m^0, g_m^0), then $\hat{R}_m^0(\boldsymbol{\alpha}, \mathbf{q}_m, s_m^0, g_m^0)$ reaches its minimum only when $\mathbf{q}_m$ is correctly specified.

Theorem 1 is the basis of our proposed iterative two-step online calibration method leveraging the residual statistic. According to the Theorem 1, we can obtain the attribute vector for each new item by arbitrarily assigning item parameters to it, e.g., $s_m^0 = 0.25$, $g_m^0 = 0.25$, and minimizing the residual statistic. In other words, it is not necessary to jointly estimate the attribute vector and the item parameters for each new item, and the vector of the new item can be obtained based on the fixed item parameters as long as $\boldsymbol{\alpha}$ is known. Then one can estimate the item parameters based on the vector obtained in the preceding step. This is very useful for situations where existing joint estimation methods suffer, e.g., when the sample size is small, which is oftentimes the case for a diagnostic test. For conciseness, the proof is presented in Appendix A.

The iterative two-step online item calibration method

Based on the preceding theorem, we propose an iterative two-step method for online item calibration. A flow chart describing the procedure is presented in Fig. 1. As we can see, by fixing the new item parameters at 0.25 (or any value between 0 and .5), the estimated attribute vector for the m^{th} new item can be obtained based on the attribute profiles estimated from the responses of the operational items. In the second step, assume the estimated vector for each new item as true, the CD-MA, CD-OEM, and CD-MEM can be applied to calibrate item parameters as

described in Chen et al. (2012). Accordingly, the resulting three variations of the iterative two-step online calibration methods based on the residual statistic are denoted as RMA, ROEM, and RMEM, respectively. Let $\hat{R}\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{Q}}_{\boldsymbol{new}}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{g}}\right)$ denote the sum of the R for all new items, that is, $\hat{R}\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{Q}}_{\boldsymbol{new}}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{g}}\right) = \sum_{m=1}^{M} \hat{R}_m\left(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{q}}_m, \hat{s}_m, \hat{g}_m\right)$, and let $\hat{R}^t_{\hat{\mathbf{Q}}_{new}}$ be the shorthand of $\hat{R}\left(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{Q}}_{\boldsymbol{new}}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{g}}\right)$ in the t^{th} iteration. The iterative algorithm stops till the number of iterations reaches its prespecified maximum or the difference between two adjacent iterations, $\hat{R}^t_{\hat{\mathbf{Q}}_{new}}$ and $\hat{R}^{t-1}_{\hat{\mathbf{Q}}_{new}}$, is smaller than a preset threshold.

The process of the calibration of the m^{th} item can be described as follows:

Step 1: Estimate the attribute vector for the m^{th} new item:

(1) Obtain $\hat{\boldsymbol{\alpha}}$ of each examinee based on their responses to the operational items;
(2) Assigning the initial slipping and guessing parameter as 0.25, estimate the attribute vector for each new item based on the proposed R statistic.

Step 2: Based on the estimated attribute-vectors obtained from the last step, apply the CD-MA, CD-OEM, or CD-

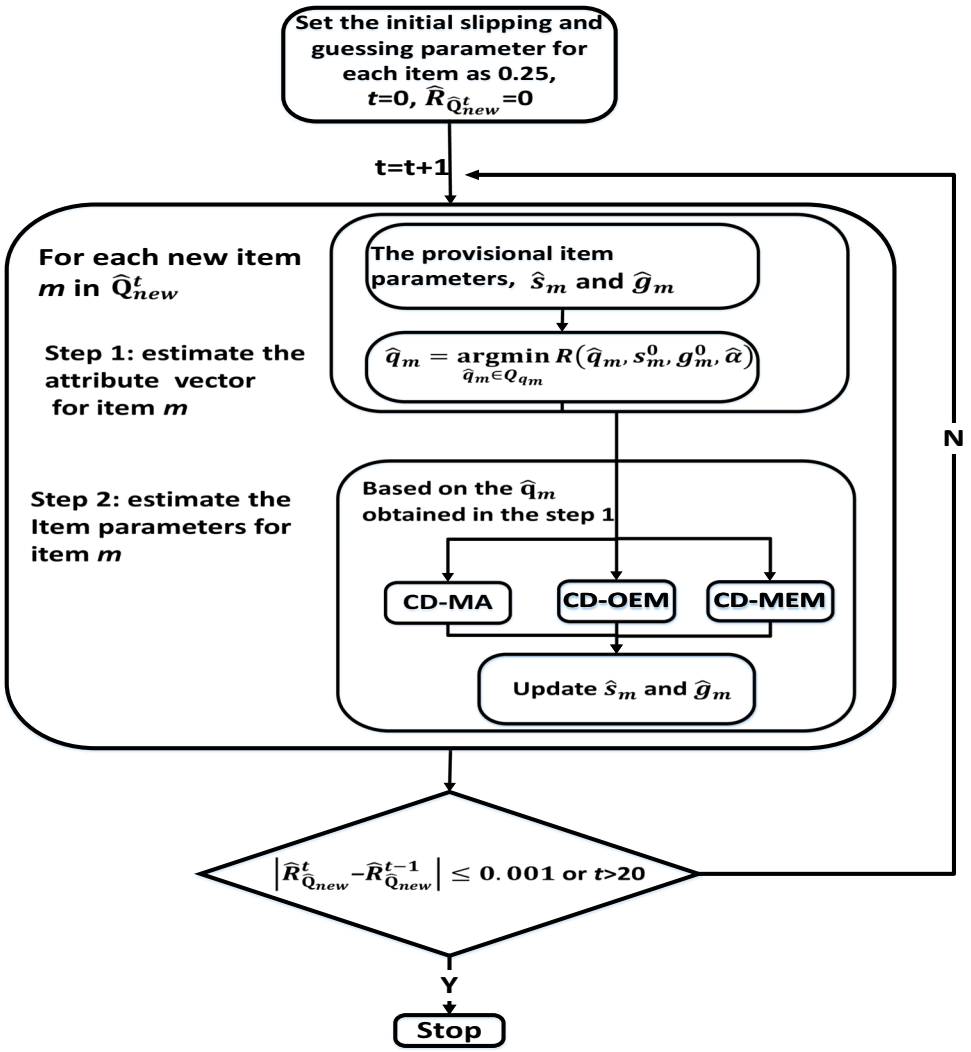

Fig. 1 The flow chart of the iterative two-step online item calibration method. Note. $\hat{\mathbf{Q}}^t_{new}$ is the attribute vector definition of the new items in the t^{th} iteration. $\hat{R}^t_{\hat{\mathbf{Q}}_{new}}$ and $\hat{R}^{t-1}_{\hat{\mathbf{Q}}_{new}}$ refer to the sum of the R statistic for all new items in the t^{th} and the $(t-1)^{th}$ iteration, respectively. $\mathbf{Q}_{q_m}$ refers to the set of the possible attribute vectors of the m^{th} new item. and $\hat{\mathbf{q}}_m$ is the estimate of the attribute vector for item m. $\hat{\boldsymbol{\alpha}}$ refers to the AMP estimates of those respondents who were administered the new item. $\hat{s}_m$ and $\hat{g}_m$ refer to estimates of the slipping and the guessing parameters, s^0_m and g^0_m are their initial values, respectively. In the context of cognitive diagnosis, CD-MA, CD-OEM, CD-MEM refer to the online calibration of item parameters based on method A, OEM, and MEM, respectively

MEM method to update the slipping and guessing parameters for the m^{th} item.

Two practical concerns arise when using the iterative two-step procedure in real applications. One is that the true AMPs are unknown, and the AMPs based on responses to operational items are used in their place. The other is that theorem 1 holds only when $N \rightarrow \infty$. Therefore, robustness of the proposed procedure in presence of unknown AMPs and limited sample size remains to be examined. In order to evaluate the performance and the robustness of the proposed two-step method under the condition of unknown AMPs and a relatively small sample size, a simulation study is conducted. According to the results of Chen et al. (2015), SIE and SimIE have almost the same performance with sample sizes smaller than 1600. Since our main goal is to compare the online item calibration methods in the context of CD-CAT with a relatively small sample size, only the JEA, SIE, and the three residual-based methods are involved in the following simulation study. The purpose of this article is twofold: (a) to introduce three residual-based methods implemented in an iterative algorithm for online calibration in CDA, and (b) to examine how the performance of these methods compares to that of the JEA and SIE under a wide range of conditions by means of a simulation study.

Simulation study

Diagnostic assessment sees great promise in classroom assessment, which calls for considerations of a small sample size and short test length. Furthermore, the AMP distributions are most likely different for respondents in different classes. Therefore, in a comprehensive simulation study, we evaluate the performance of the proposed method under various conditions, e.g., different sample sizes, test lengths, distribution of AMPs, and proportion of the new items to the operational items. The performance of the proposed methods is compared against two existing methods, JEA and SIE. For each condition, the simulation is replicated 1000 times. The same as Chen et al. (2012), the number of attributes measured by the test is set as $K = 6$. Therefore, the number of possible AMPs is $2^6 = 64$. The comparison is made in terms of the accuracy of the estimation of the attribute vectors for the new items, slipping and guess parameters, as well as the respondents' AMPs.

Sample Size Six sample sizes (200, 400, 600, 800, and 1000) are considered. The first three are small sample sizes, and the last two are medium sample sizes.

Test Length Three test lengths (20, 30, and 40) are considered. Each test consists of a certain number of operational items and new items, with the total test length being 20, 30, or 40. For each test length, the rate of new to operational items (denoted by λ) is 1:4, 1:3, or 1:2. For example, at the test length of 30, there could be six new and 24 operational items, or roughly eight new and 22 operational items, or ten new and 20 operational items.

Respondent Generation We use a similar method to Chen et al. (2012) and Chen et al. (2015) to generate the AMPs of respondents. Two independent groups of respondents are simulated. The first group assumes each respondent has a 50% probability of mastering each attribute, i.e., all attributes are equally "difficult". The second group assumes that the probability of mastery varies from one attribute to another. More specifically, the probability of mastery is set at 0.65, 0.25, 0.75, 0.45, 0.55, and 0.35 for attribute 1 to 6, where 0.65 and 0.75 refer to low difficulty, 0.45 and 0.55 refer to medium difficulty, and 0.25 and 0.35 refer to high difficulty.

Item Bank Generation Similar to Chen et al. (2012) and Chen et al. (2015), two item banks are simulated based on the ranges of the item parameters. The slipping and guessing parameters are all randomly drawn from U(0.05, 0.25) for the first item bank, which feature items with high discrimination (Kaplan et al., 2015), and drawn from U(0.15, 0.35) for the second item bank, resulting in an item bank of low discrimination (Kaplan et al., 2015). A total of 360 items with the same **Q**-matrix as in Chen et al. (2012) are generated. Typically, high discrimination items involve less noise (as represented by slipping and guessing), and lead to better measurement outcomes.

New Item Generation The same as Chen et al. (2012) and Chen et al. (2015), suppose the number of the new items as 20, which indicates that there are 20 items in the $\mathbf{Q}_{new}$, the associated attribute vectors for them are randomly drawn from the operational item banks. The set of the new items will be drawn either from the low-discrimination bank or high-discrimination bank, denoted as New_1 or New_2, respectively. Table 1 presents detailed information of the new items.

Simulation of CD-CAT and Online Calibration For each respondent, the CD-CAT and the online calibration proceed as follows: (1) Generate the initial AMP estimate randomly, with each attribute having an equal probability of being mastered or not mastered; (2) Select the next item based on the most recent AMP estimate; (3) Generate the response to the selected item and update the AMP estimate according to the responses to the previously administered items. Steps 2 and step 3 are repeated until the stopping rule is satisfied. During the process, a certain number of new items (1/3, 1/4, or 1/5 of the test length) are randomly seeded in the test of each respondent. Three fixed

Table 1 The settings of the new items

ID	New_1								New_2							
	(0.05, 0.25)		$\mathbf{Q}_{New_1}$						(0.15, 0.35)		$\mathbf{Q}_{New_2}$					
	s	g							s	g						
1	0.226	0.233	1	1	0	1	0	0	0.239	0.333	0	1	0	1	0	1
2	0.082	0.173	0	0	0	0	1	1	0.267	0.306	1	0	1	0	0	0
3	0.215	0.180	1	0	0	1	0	0	0.284	0.173	0	0	0	0	1	0
4	0.169	0.214	0	0	0	0	1	0	0.198	0.269	0	0	0	0	1	0
5	0.171	0.128	0	0	0	1	0	1	0.224	0.164	0	0	1	0	0	0
6	0.241	0.186	0	1	1	0	0	1	0.270	0.308	0	1	0	1	1	0
7	0.207	0.115	0	1	0	1	0	1	0.232	0.324	1	0	0	0	0	0
8	0.071	0.109	0	0	0	0	0	1	0.243	0.292	0	0	0	1	0	0
9	0.244	0.080	1	0	0	0	0	0	0.268	0.289	0	0	1	0	0	0
10	0.196	0.234	0	0	1	0	0	0	0.340	0.333	1	1	0	1	0	0
11	0.246	0.174	0	0	0	0	1	1	0.306	0.191	1	0	0	1	0	1
12	0.208	0.173	0	1	0	0	1	0	0.176	0.345	0	1	0	0	1	1
13	0.166	0.095	0	1	0	1	0	1	0.253	0.285	1	0	0	0	0	0
14	0.076	0.053	1	0	0	0	0	0	0.247	0.315	0	0	0	0	0	1
15	0.206	0.226	0	1	0	0	0	0	0.176	0.288	1	1	0	1	0	0
16	0.144	0.212	0	0	0	1	0	1	0.333	0.278	0	0	0	0	1	0
17	0.109	0.051	0	1	0	0	1	1	0.164	0.188	1	1	1	0	0	0
18	0.151	0.144	1	0	1	0	0	0	0.271	0.316	0	0	0	1	0	1
19	0.122	0.071	0	1	1	1	0	0	0.306	0.257	0	0	0	1	0	0
20	0.185	0.088	0	0	1	0	0	0	0.326	0.162	0	0	1	1	0	1

New_1 and New_2 are the two settings of the seeded new items. s and g refer to the slipping and the guessing parameters, respectively. $\mathbf{Q}_{New_1}$ and $\mathbf{Q}_{New_2}$ are the $\mathbf{Q}$-matrices based on the settings of New_1 and New_2, respectively

test lengths $L = 20$, 30, and 40 are simulated, and the item selection strategy for operational items is the Shannon Entropy method (SHE; Cheng, 2009; Tatsuoka, 2002, Xu et al., 2003). The prior distribution of the AMP is assumed to be the uniform distribution. It should be noted that the AMP estimates of CD-Method A, CD-OEM, and CD-MEM are based on the operational items, while those of SIE and SimIE are based on both the operational items and new items.

Update of the AMP In the simulation, the Maximum A Posterior (MAP; Huebner & Wang, 2011) method is used to update the AMP estimates of respondents:

$$\hat{\boldsymbol{\alpha}}_i = \underset{v=1,2,\cdots,2^K}{\operatorname{argmax}} P\left(\boldsymbol{\alpha}_v|\mathbf{X}_i\right), \tag{7}$$

where $\mathbf{X}_i$ refers to the response pattern for the i^{th} respondent. As noted by Chen et al. (2012), the AMP is estimated after each operational item is answered. The test is terminated as soon as the test length reaches L.

Evaluation Criteria For each condition, the following eight criteria are applied to evaluate the performance of online calibration methods. The first three indices are used to evaluate the estimation of the AMPs, while the remaining indices address the estimation accuracy of the item parameters and the attribute vectors for the new items.

Person Pattern Accuracy Rate (PPAR) The *PPAR* represents the proportion of respondents whose AMPs are correctly estimated, which is defined as follows:

$$PPAR = \frac{\sum_{i=1}^{N} I\left(\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i\right)}{N}, \tag{8}$$

where $I\left(\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i\right)$ is an indicator function which equals 1 if the estimate AMP $\hat{\boldsymbol{\alpha}}_i$ for the i^{th} respondent equates to its true value $\boldsymbol{\alpha}_i$, and 0 otherwise.

Person Attribute Accuracy Rate (*PAAR*). The $PAAR_k$ quantifies the estimation accuracy rate for attribute k:

$$PAAR_k = \frac{\sum_{i=1}^{N} I\left(\alpha_{ik} = \hat{\alpha}_{ik}\right)}{N}. \tag{9}$$

Average Person Attribute Accuracy Rate (APAAR) The *APAAR* summarizes the average attribute estimation accuracy at the person level for the CD-CAT, which can be determined as follows

$$APAAR = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K} I\left(\alpha_{ik} = \hat{\alpha}_{ik}\right)}{NK}. \quad (10)$$

The following five indexes evaluate the estimation of the new items.

Root Mean Squared Error (RMSE) The *RMSE* summarizes the overall performance of the calibration accuracy of the slipping and guessing parameters of the *M* new items (Chen et al., 2012; Chen et al., 2015):

$$s_{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(s_m - \hat{s}_m\right)^2}, \quad (11)$$

$$g_{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(g_m - \hat{g}_m\right)^2}. \quad (12)$$

Item Pattern Accuracy Rate (IPAR) The *IPAR* indicates the calibration accuracy for the attribute vector of the new items, which is defined as follows:

$$IPAR = \frac{\sum_{m=1}^{M} I\left(\hat{\mathbf{q}}_m = \mathbf{q}_m\right)}{M}, \quad (13)$$

where $I(\bullet)$ is an indicator function: $I\left(\hat{\mathbf{q}}_m = \mathbf{q}_m\right)$ returns a value of 1 when $\hat{\mathbf{q}}_m$ and $\mathbf{q}_m$ are equal, and returns a 0 otherwise.

Item Attribute Accuracy Number (IAAN) The *IAAN* quantifies the average number of attributes per item that are specified correctly for the new items:

$$IAAN = \frac{\sum_{m=1}^{M}\sum_{k=1}^{K} I\left(\hat{q}_{mk} = q_{mk}\right)}{M}. \quad (14)$$

Among the preceding indices: The *PPAR*, *PAAR*, and *APAAR* are used to summarize the estimation accuracy of AMPs. Higher value indicates better estimation. s_{RMSE} and g_{RMSE} are used to evaluate the item parameter estimation accuracy for the new items. Smaller s_{RMSE} and g_{RMSE} indicate more accurate estimation of item parameters. The *IPAR* and *IAAN* quantify the attribute vector estimation accuracy of the new items, with larger values representing a more accurate estimation of attribute vector.

Results

Figure 2 and Table 2 provide the indices of the AMP estimation accuracy for the CD-CAT, which includes *PPAR*, *PAAR*, *APAAR*, under the condition of the sample size of 200 (Results for other sample sizes show similar patterns and are omitted to save space. They are available upon request). It should be noted that these three indices are calculated only based on the operational items. The two uppercase letters in the first column of the tables refer to the range of item parameters and attribute mastery probability. The letters "L" and "H" denote the low- and high-discrimination items with parameters' range [0.15, 0.35] and [0.05, 0.25], respectively. The letters "S" and "D" refer to respondents with the same and different mastery probabilities, respectively. Results indicate that the test with highly discriminating items is indeed better for the estimation of respondents' attribute profile, consistent with expectation. For example, the test with 13 high-discrimination operational items (i.e., in the 20-item highly discriminating test, with 1:2 new to operational item ratio) can reach a comparable *PPAR* of the test with 22 low-discrimination operational items (i.e., in the low-discrimination test with test length of 30, with 1:4 new to operational item ratio). Similar results between HS and HD, as well as between LS and LD suggest that the attribute mastery probabilities show little effect on the estimation of the respondents' attribute profiles. Due to the fixed test length of the CD-CAT, the AMP estimation precision will decrease with the number of seeded new items, because AMP estimation depends on the responses to the operational items. For example, at the length of the 20-item test with the rate of new to operational items being 1:4, 1:3, and 1:2, the *PPAR*s are 0.944, 0.906, and 0.810, respectively.

The six columns below *PAAR* in Table 2 are the estimation accuracy index of six attributes, which indicates that the test with high-discrimination items result in a higher *PAAR*, and the test with more operational items also lead to a higher *PPAR*, which can be easily seen in Fig. 2. When the test length reaches as high as 40, the difference caused by the ratio of new items and operational item becomes less pronounced (see Fig. 2). On the other hand, the distribution of the attribute mastery probability shows a small effect on the estimation of respondents' attribute profiles. Table 2 also shows that the *PPAR* and *APAAR* indices have the same trend as *PAAR*.

Tables 3, 4, 5, 6 and 7 present the *IPAR* index of the new items. Based on the results, more discriminating items, i.e., items with lower guessing and slipping parameters, are beneficial for online calibration. The proposed residual-based (*R*-based) methods outperform the JEA and SIE method in attribute vector estimation of the new items. When all attributes are equally likely to be mastered, RMEM has the highest *IPAR* in most cases. Between JEA and SIE, there does not seem to be a consistent winner in terms of the *IPAR* index, suggesting that the Bayesian version of the JEA could not always borrow enough information to help the item calibration. For the *R*-based methods, RMA and ROEM have close performances. Results also suggest that a higher *IPAR* index can be obtained with more seeded new items. For example,

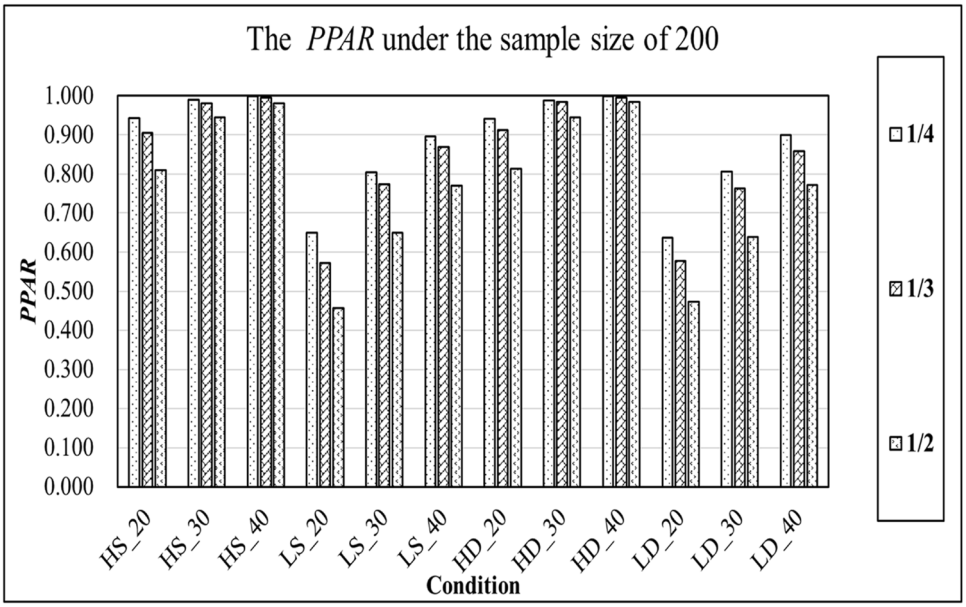

Fig. 2 The *PPAR* (Person Pattern Accuracy Rate) of the new items. Note. The first letter 'H' or 'L' in the labels for the *x*-axis refers to items with high- or low-discrimination, the second letter 'S' or 'D' refer to respondents with the same or different attribute mastery probability (ies). The number after the underscore refers to the test length. For example, HS_20 refers to the test with highly discriminative items and test length of 20. The numbers in the legend refer to the ratio of the number of seeded new items to the number of operational items

consider a sample size of 200 respondents and test length of 20, the *IPAR* index for RMEM under three different ratios of seeded new items and operational items are 0.464, 0.524, 0.538 (see Table 3). The increase of seeded new items leads to more responses to each new item, and subsequently leads to better estimation of new items' attribute vectors. Consider the sample size of 400 and a 20-item test, if five new items (corresponding to 1:3 new to operational item ratio) are seeded in the test, about 400×5/20 = 100 respondents answer each new item on average. Nevertheless, if seven new items (i.e., the ratio of new to operational items is 1:2) are seeded, about 400×7/20 = 140 respondents answer each new item on average. Meanwhile, the decrease of the operational items will lead to lower *PPAR* index, which is harmful to the calibration. Therefore, a trade-off between the number of seeded new items and operational items needs to be considered.

All five methods have better performances with more discriminating items, which is consistent with the findings of Chen et al. (2012). For example, for the RMEM method in the 20 items test with 200 respondents, the values of the *IPAR* index for the HS and LS condition with a 1:4 new to operational item ratio are .790 and .464. For the two distributions of respondents' attribute mastery probability, each of the methods has better performance in terms of the *IPAR* index under the condition of respondents with the same attribute mastery probability of 0.5. Also, take the 20-item test, 200-respondents condition as an example, under the HS and HD conditions, with a new to operational ratio of 1:4, the *IPAR*s of the RMEM method are .790 and .708, respectively.

Across five samples, the same trend for the *IAAN* index is observed. Hence, we only provide the results under the condition of the sample size of 200 and 400, which are presented in Tables 8 and 9. For this index, 6 means that all attributes of the item are estimated correctly, and the closer to 6 the better. As we can see, RMEM performs better in most of the conditions. RMA and ROEM have comparable *IAAN* in some cases. For example, 4897 attributes can be correctly recovered on average under the condition of 20-item test with 1/4 seeded new items, and respondents with uniform attribute mastery probability.

Consider the item parameter estimation of the new items, the RMA and RMEM lead to comparable *RMSE*s for both the slipping and guessing parameters, and they together outperform the other three methods. As shown in Tables 10, 11, 12, 13 and 14, ROEM results in higher s_{RMSE} and g_{RMSE} than RMEM and RMA. As discussed before, information borrowed from the respondents' posterior distribution may not be enough to improve the online item calibration, and in most cases, the JEA has the largest s_{RMSE} and g_{RMSE}. The same as the attribute vector estimation, each method has better or comparable performances when the respondents have the same attribute mastery probability. With more seeded new items, estimation of the new items become better, as more seeded new items for each respondent mean

Table 2 Estimation accuracy of the respondents under the sample size of 200

Condition	*L*	λ	*PPAR*	*PAAR*						*APAAR*
				A1	A2	A3	A4	A5	A6	
HS	20	1/4	**0.944**	**0.990**	**0.990**	**0.990**	**0.988**	**0.990**	**0.986**	**0.989**
		1/3	0.906	0.983	0.984	0.981	0.975	0.985	0.980	0.981
		1/2	0.810	0.965	0.963	0.961	0.957	0.966	0.953	0.961
	30	1/4	**0.989**	**0.998**	**0.999**	**0.998**	**0.998**	**0.998**	**0.998**	**0.998**
		1/3	0.982	0.996	0.997	0.996	0.996	0.997	0.996	0.996
		1/2	0.945	0.989	0.992	0.989	0.986	0.991	0.986	0.989
	40	1/4	**0.998**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**	**1.000**
		1/3	0.995	0.999	0.999	0.999	0.999	0.999	0.999	0.999
		1/2	0.981	0.997	0.997	0.996	0.995	0.997	0.995	0.996
LS	20	1/4	**0.649**	**0.906**	**0.930**	**0.920**	**0.908**	**0.931**	**0.909**	**0.917**
		1/3	0.573	0.889	0.912	0.900	0.887	0.915	0.882	0.897
		1/2	0.457	0.843	0.878	0.873	0.850	0.883	0.838	0.861
	30	1/4	**0.805**	**0.952**	**0.961**	**0.954**	**0.956**	**0.964**	**0.954**	**0.957**
		1/3	0.774	0.943	0.956	0.945	0.947	0.959	0.945	0.949
		1/2	0.649	0.908	0.929	0.920	0.910	0.929	0.902	0.916
	40	1/4	**0.896**	**0.975**	**0.983**	**0.976**	**0.980**	**0.982**	**0.974**	**0.978**
		1/3	0.869	0.969	0.977	0.973	0.973	0.974	0.969	0.973
		1/2	0.771	0.945	0.955	0.948	0.944	0.959	0.941	0.949
HD	20	1/4	**0.942**	**0.989**	**0.992**	**0.987**	**0.987**	**0.989**	**0.989**	**0.989**
		1/3	0.913	0.984	0.986	0.982	0.978	0.985	0.981	0.983
		1/2	0.813	0.966	0.971	0.960	0.955	0.963	0.960	0.962
	30	1/4	**0.988**	**0.998**	**0.999**	**0.998**	**0.998**	**0.998**	**0.997**	**0.998**
		1/3	0.984	0.997	0.998	0.998	0.996	0.997	0.996	0.997
		1/2	0.944	0.989	0.993	0.989	0.986	0.990	0.988	0.989
	40	1/4	**0.998**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.999**	**1.000**
		1/3	0.996	0.999	0.999	0.999	1.000	0.999	0.999	0.999
		1/2	0.985	0.998	0.998	0.997	0.997	0.997	0.997	0.997
LD	20	1/4	**0.637**	**0.895**	**0.939**	**0.910**	**0.908**	**0.922**	**0.909**	**0.914**
		1/3	0.578	0.874	0.922	0.885	0.899	0.909	0.896	0.898
		1/2	0.472	0.832	0.894	0.862	0.861	0.877	0.859	0.864
	30	1/4	**0.805**	**0.946**	**0.968**	**0.948**	**0.957**	**0.963**	**0.959**	**0.957**
		1/3	0.763	0.933	0.961	0.935	0.951	0.956	0.947	0.947
		1/2	0.639	0.893	0.938	0.903	0.918	0.916	0.917	0.914
	40	1/4	**0.901**	**0.974**	**0.987**	**0.973**	**0.980**	**0.980**	**0.981**	**0.979**
		1/3	0.859	0.961	0.980	0.962	0.970	0.971	0.970	0.969
		1/2	0.772	0.939	0.961	0.936	0.951	0.954	0.952	0.949

The indices in the table were obtained only based on the operational item, where λ refers to the rate of new to operational items, *L* refers to the test length. The first letters in 'HS', 'LS', 'LS', 'LD', which are 'H' or 'L', refer to items with high- or low-discrimination, the second letter 'S' or 'D' refers to respondents with the same or different attribute mastery probability (ies). A1–A6 refers to the six simulated attributes, respectively. *PPAR*, *PAAR*, and *APAAR* are the person pattern accuracy rate, the person attribute accuracy rate, and the average person attribute accuracy rate, respectively. Boldfaced values indicate the best performance across λ levels

more responses can be collected for each new item. Though the estimation accuracy of the respondents' AMP decreases in the test with more seeded new items, the increase of the respondents for each new item can improve the calibration of the new item, again pointing to a tradeoff.

Figure 3 illustrates the *IPAR* index of the condition of sample size 200 with different test lengths. As we can see, on one hand, the *IPAR* becomes better with more seeded new items (with a new to operational item ratio of 1:4 to 1:2 within each specific test length). On the other hand, the

Table 3 The *IPAR* (Item Pattern Accuracy Rate) for the new items with the sample size of 200

Group	*L*	λ	*s*, *g*~*U*(0.15,0.35)					*s*, *g*~*U*(0.05,0.25)				
			RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE
Uniform attribute mastery probability	20	1/4	**0.464**	**0.464**	**0.464**	0.431	0.403	0.789	**0.790**	**0.790**	0.768	0.775
		1/3	0.518	0.522	**0.524**	0.481	0.463	0.873	**0.876**	**0.876**	0.856	0.857
		1/2	**0.538**	**0.538**	**0.538**	0.508	0.481	0.917	0.918	**0.919**	0.895	0.874
	30	1/4	0.653	0.655	**0.656**	0.629	0.642	**0.922**	0.921	**0.922**	0.908	0.906
		1/3	0.711	0.711	**0.712**	0.686	0.688	**0.942**	**0.942**	**0.942**	0.935	0.937
		1/2	**0.774**	**0.774**	**0.774**	0.736	0.735	**0.981**	**0.981**	**0.981**	0.976	0.977
	40	1/4	**0.767**	**0.767**	**0.767**	0.735	0.753	**0.950**	**0.950**	**0.950**	0.943	0.944
		1/3	0.839	**0.840**	**0.840**	0.829	0.831	**0.979**	**0.979**	**0.979**	0.971	0.972
		1/2	0.878	0.878	**0.898**	0.858	0.892	0.990	0.990	**0.996**	0.989	0.990
Uneven attribute mastery probability	20	1/4	0.407	0.409	**0.441**	0.402	0.403	0.704	0.704	**0.708**	0.706	0.699
		1/3	0.444	0.449	**0.455**	0.452	0.441	**0.770**	0.769	0.767	0.759	0.749
		1/2	0.458	0.460	**0.476**	0.463	0.453	**0.801**	0.798	0.797	0.773	0.764
	30	1/4	0.565	**0.567**	**0.567**	0.559	0.560	0.830	0.831	**0.833**	0.825	0.813
		1/3	0.578	0.578	**0.599**	0.583	0.587	**0.864**	0.863	**0.864**	0.859	0.854
		1/2	0.625	0.626	**0.644**	0.624	0.623	0.912	0.910	**0.919**	0.908	0.913
	40	1/4	0.666	0.666	**0.697**	0.673	0.684	0.864	**0.867**	**0.867**	0.865	0.866
		1/3	0.713	0.714	**0.731**	0.715	0.720	0.915	**0.916**	**0.916**	0.909	0.911
		1/2	0.762	0.762	**0.796**	0.764	0.783	**0.949**	0.948	**0.949**	0.946	0.947

λ is the rate of new to operational items, and *L* refers to the test length. RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA, and SIE refer to the joint estimation algorithm and the single item estimation method, respectively. *s* and *g* refer to the slipping and the guessing parameters, respectively. Boldfaced values indicate the best performance across estimation methods

Table 4 The *IPAR* (Item Pattern Accuracy Rate) for the new items with the sample size of 400

Group	*L*	λ	*s*, *g*~*U*(0.15,0.35)					*s*, *g*~*U*(0.05,0.25)				
			RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE
Uniform attribute mastery probability	20	1/4	**0.638**	0.637	0.636	0.600	0.593	**0.945**	0.944	**0.945**	0.924	0.934
		1/3	**0.715**	0.714	**0.715**	0.674	0.617	0.967	0.966	**0.977**	0.965	0.971
		1/2	**0.717**	0.716	**0.717**	0.674	0.657	0.981	0.981	**0.987**	0.979	0.972
	30	1/4	0.857	0.859	**0.866**	0.843	0.857	**0.983**	**0.983**	**0.983**	0.978	0.978
		1/3	0.887	0.889	**0.899**	0.878	0.891	0.995	0.995	**0.996**	0.993	0.994
		1/2	0.922	0.922	**0.932**	0.906	0.892	**0.996**	**0.996**	**0.996**	0.996	0.996
	40	1/4	0.924	0.923	**0.925**	0.913	0.922	0.984	0.984	**0.992**	0.987	0.987
		1/3	0.959	0.960	**0.966**	0.954	0.959	**0.998**	**0.998**	**0.998**	**0.998**	**0.998**
		1/2	0.982	0.982	**0.986**	0.980	0.980	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Uneven attribute mastery probability	20	1/4	0.534	**0.538**	0.537	0.525	0.529	0.843	0.843	**0.847**	0.840	0.841
		1/3	0.563	0.561	**0.586**	0.581	0.566	**0.910**	0.907	0.909	0.892	0.892
		1/2	0.592	0.593	**0.594**	0.593	**0.594**	0.908	0.908	**0.928**	0.894	0.899
	30	1/4	0.729	0.723	**0.742**	0.709	0.739	0.933	0.932	**0.935**	0.932	0.929
		1/3	0.752	0.753	**0.759**	0.750	0.758	0.950	0.950	**0.955**	0.953	0.950
		1/2	0.767	**0.769**	**0.769**	0.764	0.758	0.973	0.972	**0.977**	0.972	0.976
	40	1/4	0.823	0.823	**0.828**	0.804	0.821	**0.952**	**0.952**	**0.952**	0.944	0.945
		1/3	0.871	0.872	**0.877**	0.858	0.870	0.972	0.972	**0.973**	0.971	0.972
		1/2	0.889	0.889	**0.909**	0.885	0.887	0.988	0.988	**0.989**	0.987	0.989

λ is the rate of new to operational items, and *L* refers to the test length. RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA, and SIE refer to the joint estimation algorithm and the single item estimation method, respectively. *s* and *g* refer to the slipping and the guessing parameters, respectively. Boldfaced values indicate the best performance across estimation methods

Table 5 The *IPAR* (Item Pattern Accuracy Rate) for the new items with the sample size of 600

Group	*L*	λ	*s*, *g*~*U*(0.15,0.35)					*s*, *g*~*U*(0.05,0.25)				
			RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE
Uniform attribute mastery probability	20	1/4	0.743	0.740	**0.768**	0.719	0.742	**0.971**	**0.971**	**0.971**	0.966	0.970
		1/3	**0.787**	**0.787**	**0.787**	0.769	0.690	**0.989**	**0.989**	**0.989**	0.988	**0.989**
		1/2	0.794	0.794	**0.796**	0.774	0.757	**0.994**	**0.994**	**0.994**	0.991	0.991
	30	1/4	**0.928**	**0.928**	**0.928**	0.909	0.925	0.995	0.995	**0.996**	**0.996**	**0.996**
		1/3	0.938	0.939	**0.940**	0.925	0.927	0.997	0.997	**0.998**	**0.998**	**0.998**
		1/2	**0.961**	**0.961**	**0.961**	0.944	0.933	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	40	1/4	**0.969**	0.968	**0.969**	0.962	0.963	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**
		1/3	0.988	0.988	**0.989**	0.984	0.988	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		1/2	**0.990**	**0.990**	**0.990**	0.986	**0.990**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Uneven attribute mastery probability	20	1/4	0.620	0.620	**0.630**	0.620	0.614	0.910	0.910	**0.912**	0.903	0.911
		1/3	0.624	0.624	**0.654**	**0.654**	**0.654**	0.933	0.933	**0.935**	0.923	0.934
		1/2	0.651	0.655	**0.659**	**0.659**	**0.659**	0.941	0.941	**0.950**	0.933	0.940
	30	1/4	0.802	0.804	**0.815**	0.798	0.812	**0.961**	**0.961**	**0.961**	0.960	0.959
		1/3	0.831	0.835	**0.848**	0.823	0.831	0.981	0.981	**0.982**	0.981	0.981
		1/2	0.838	0.839	**0.851**	0.842	0.832	0.987	0.987	**0.989**	0.985	**0.989**
	40	1/4	0.883	0.882	**0.886**	0.871	0.874	**0.981**	**0.981**	**0.981**	0.977	0.976
		1/3	0.926	0.926	**0.939**	0.915	0.931	**0.986**	**0.986**	**0.986**	0.985	**0.986**
		1/2	0.931	0.930	**0.954**	0.931	0.937	0.994	0.994	**0.995**	0.993	0.994

λ is the rate of new to operational items, and *L* refers to the test length. RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA, and SIE refer to the joint estimation algorithm and the single item estimation method, respectively. *s* and *g* refer to the slipping and the guessing parameters, respectively. Boldfaced values indicate the best performance across estimation methods

Table 6 The *IPAR* (Item Pattern Accuracy Rate) for the new items with the sample size of 800

Group	*L*	λ	*s*, *g*~*U*(0.15,0.35)					*s*, *g*~*U*(0.05,0.25)				
			RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE
Uniform attribute mastery probability	20	1/4	0.844	0.843	**0.845**	0.825	0.825	**0.990**	**0.990**	**0.990**	**0.990**	**0.990**
		1/3	0.862	0.864	**0.865**	0.850	0.851	0.995	0.997	**0.998**	**0.998**	0.997
		1/2	**0.879**	**0.879**	**0.879**	0.865	0.867	0.997	0.995	**0.998**	0.997	**0.998**
	30	1/4	0.957	0.957	**0.966**	0.948	0.959	**0.998**	**0.998**	**0.998**	**0.998**	**0.998**
		1/3	0.979	**0.981**	**0.981**	0.970	0.976	**0.999**	**0.999**	**0.999**	**0.999**	**0.999**
		1/2	0.979	0.980	**0.988**	0.980	0.986	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	40	1/4	**0.989**	0.988	**0.989**	0.987	**0.989**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		1/3	**0.998**	**0.998**	**0.998**	0.995	0.995	**1.000**	**1.000**	**1.000**	0.999	0.999
		1/2	**1.000**	**1.000**	**1.000**	0.996	0.996	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Uneven attribute mastery probability	20	1/4	0.677	0.693	**0.694**	0.689	0.689	**0.950**	**0.950**	**0.950**	0.945	0.947
		1/3	0.691	0.707	**0.719**	0.712	0.713	0.958	0.958	**0.966**	0.957	0.964
		1/2	0.711	0.717	**0.727**	0.722	0.724	0.967	0.967	**0.987**	0.964	0.965
	30	1/4	**0.873**	**0.873**	**0.873**	0.860	0.865	**0.981**	0.980	**0.981**	0.977	0.976
		1/3	0.883	0.886	**0.895**	0.881	0.894	0.990	0.990	**0.992**	0.987	0.987
		1/2	0.886	0.886	**0.899**	0.888	0.865	0.995	0.995	**0.998**	0.994	0.998
	40	1/4	0.923	0.923	**0.932**	0.913	0.927	**0.987**	**0.987**	**0.987**	0.984	0.983
		1/3	**0.949**	**0.949**	**0.949**	0.943	0.945	0.995	0.995	**0.995**	0.995	0.995
		1/2	0.956	0.956	**0.976**	0.949	0.966	0.999	0.999	**1.000**	1.000	1.000

λ is the rate of new to operational items, and *L* refers to the test length. RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA, and SIE refer to the joint estimation algorithm and the single item estimation method, respectively. *s* and *g* refer to the slipping and the guessing parameters, respectively. Boldfaced values indicate the best performance across estimation methods

Table 7 The *IPAR* (Item Pattern Accuracy Rate) for the new items with the sample size of 1000

Group	*L*	λ	*s*, *g*~*U*(0.15,0.35)					*s*, *g*~*U*(0.05,0.25)				
			RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE
Uniform attribute mastery probability	20	1/4	0.860	0.860	**0.861**	0.843	0.846	0.994	0.994	**0.996**	0.994	**0.996**
		1/3	0.887	**0.888**	**0.888**	0.878	0.881	0.996	0.996	**0.999**	0.997	0.997
		1/2	0.897	**0.898**	**0.898**	0.880	0.881	**0.999**	**0.999**	**0.999**	0.997	0.998
	30	1/4	0.975	0.977	**0.978**	0.971	0.975	**0.999**	**0.999**	**0.999**	0.998	0.998
		1/3	**0.986**	**0.986**	**0.986**	0.982	0.984	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		1/2	**0.988**	0.987	**0.988**	0.983	0.985	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	40	1/4	**0.995**	**0.995**	**0.995**	0.992	0.994	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		1/3	**0.999**	**0.999**	**0.999**	0.997	0.997	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
		1/2	**1.000**	**1.000**	**1.000**	0.999	0.998	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Uneven attribute mastery probability	20	1/4	0.723	0.722	**0.730**	0.727	0.728	**0.962**	0.961	0.961	0.952	0.957
		1/3	0.739	0.741	**0.762**	0.746	0.746	0.974	0.973	**0.987**	0.968	0.973
		1/2	0.740	0.743	**0.774**	0.764	0.765	0.977	0.977	**0.992**	0.973	0.979
	30	1/4	0.888	0.887	**0.896**	0.871	0.888	0.986	0.986	**0.989**	**0.989**	**0.989**
		1/3	0.909	0.908	**0.918**	0.906	0.910	**0.993**	**0.993**	**0.993**	0.991	0.992
		1/2	0.917	0.917	**0.927**	0.918	0.918	**0.999**	**0.999**	**0.999**	0.997	0.997
	40	1/4	0.944	0.944	**0.946**	0.935	0.940	**0.996**	**0.996**	**0.996**	0.995	0.995
		1/3	0.969	0.969	**0.970**	0.965	0.964	0.997	0.997	**0.998**	**0.998**	**0.998**
		1/2	0.969	0.969	**0.987**	0.967	0.973	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

λ is the rate of new to operational items, and *L* refers to the test length. RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA, and SIE refer to the joint estimation algorithm and the single item estimation method, respectively. *s* and *g* refer to the slipping and the guessing parameters, respectively. Boldfaced values indicate the best performance across estimation methods

Table 8 The *IAAN* (Item Attribute Accuracy Number) for the new items with the sample size of 200

Group	*L*	λ	*s*, *g*~*U*(0.15,0.35)					*s*, *g*~*U*(0.05,0.25)				
			RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE
Uniform attribute mastery probability	20	1/4	4.888	4.888	**4.897**	4.838	4.816	5.644	5.645	**5.646**	5.618	5.623
		1/3	5.004	5.013	**5.035**	4.948	5.015	5.806	**5.811**	**5.811**	5.787	5.792
		1/2	5.065	**5.073**	**5.073**	5.001	5.064	5.878	5.878	**5.882**	5.853	5.837
	30	1/4	5.344	5.336	**5.384**	5.305	5.341	**5.884**	**5.884**	**5.884**	5.864	5.870
		1/3	**5.486**	5.485	**5.486**	5.443	5.458	**5.923**	**5.923**	**5.923**	5.914	5.917
		1/2	5.618	**5.619**	5.617	5.561	5.578	**5.977**	**5.977**	**5.977**	5.974	5.975
	40	1/4	**5.602**	5.601	5.599	5.566	5.598	**5.935**	**5.935**	**5.935**	5.927	5.927
		1/3	5.721	**5.726**	**5.726**	5.714	5.720	**5.973**	**5.973**	**5.973**	5.965	5.965
		1/2	5.807	5.807	**5.808**	5.788	5.843	5.988	5.988	**5.989**	5.987	5.988
Uneven attribute mastery probability	20	1/4	4.788	4.788	**4.791**	4.719	4.759	5.504	5.506	**5.508**	5.490	5.462
		1/3	4.888	4.900	**4.918**	4.854	4.896	**5.628**	5.624	5.621	5.582	5.578
		1/2	4.925	4.935	**4.937**	4.883	4.942	**5.685**	5.679	5.680	5.651	5.630
	30	1/4	5.187	**5.199**	**5.199**	5.145	5.153	5.732	5.732	**5.735**	5.720	5.701
		1/3	5.234	5.234	**5.238**	5.202	5.216	**5.792**	5.791	**5.792**	5.780	5.774
		1/2	5.330	5.330	**5.333**	5.314	5.366	5.885	5.881	**5.888**	5.876	5.884
	40	1/4	**5.413**	**5.413**	**5.413**	5.397	5.425	5.789	**5.792**	**5.792**	5.779	5.771
		1/3	5.528	5.530	**5.532**	5.513	5.522	5.874	**5.876**	**5.876**	5.866	5.859
		1/2	5.618	5.617	**5.623**	5.606	5.635	**5.936**	5.935	**5.936**	5.935	5.935

λ is the rate of new to operational items, and *L* refers to the test length. RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA, and SIE refer to the joint estimation algorithm and the single item estimation method, respectively. *s* and *g* refer to the slipping and the guessing parameters, respectively. Boldfaced values indicate the best performance across estimation methods

Table 9 The *IAAN* (Item Attribute Accuracy Number) for the new items with the sample size of 400

Group	*L*	λ	*s*, *g*~*U*(0.15,0.35)					*s*, *g*~*U*(0.05,0.25)				
			RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE
Uniform attribute mastery probability	20	1/4	5.329	**5.328**	5.326	5.274	5.310	5.922	5.922	**5.923**	5.899	5.910
		1/3	5.475	**5.478**	5.468	5.428	5.449	5.951	5.951	**5.953**	5.952	5.952
		1/2	5.491	**5.489**	5.488	5.430	5.458	5.975	5.975	**5.979**	5.976	5.969
	30	1/4	5.767	5.770	**5.771**	5.765	5.769	5.982	5.982	**5.983**	5.976	5.976
		1/3	5.823	**5.825**	**5.825**	5.803	5.840	**5.995**	**5.995**	**5.995**	5.993	5.994
		1/2	5.886	5.884	**5.887**	5.864	5.859	**5.996**	**5.996**	**5.996**	**5.996**	**5.996**
	40	1/4	5.890	5.886	**5.892**	5.876	5.884	5.983	5.983	**5.987**	5.986	5.986
		1/3	5.942	5.946	**5.947**	5.941	5.946	**5.998**	**5.998**	**5.998**	**5.998**	**5.998**
		1/2	5.974	5.974	**5.977**	5.972	5.975	**6.000**	**6.000**	**6.000**	**6.000**	**6.000**
Uneven attribute mastery probability	20	1/4	5.156	**5.172**	5.157	5.120	5.146	5.768	5.767	**5.769**	5.759	5.762
		1/3	5.244	5.250	**5.255**	5.227	5.254	**5.872**	5.868	**5.872**	5.851	5.856
		1/2	5.188	5.190	**5.193**	**5.193**	**5.193**	5.871	5.872	**5.873**	5.856	5.852
	30	1/4	5.550	5.536	**5.554**	5.502	5.546	5.908	5.907	**5.909**	5.902	5.900
		1/3	5.605	5.602	**5.603**	5.585	5.600	5.934	5.931	**5.939**	5.938	5.933
		1/2	5.632	5.634	**5.634**	**5.622**	5.629	5.969	5.968	**5.974**	5.968	5.972
	40	1/4	5.725	5.728	**5.732**	5.690	5.721	**5.939**	**5.939**	**5.939**	5.928	5.929
		1/3	5.803	5.804	**5.813**	5.784	5.803	5.967	5.967	**5.968**	5.967	5.966
		1/2	5.839	5.837	**5.844**	5.828	5.837	5.986	5.986	**5.988**	5.985	5.986

λ is the rate of new to operational items, and *L* refers to the test length. RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA, and SIE refer to the joint estimation algorithm and the single item estimation method, respectively. *s* and *g* refer to the slipping and the guessing parameters, respectively. Boldfaced values indicate the best performance across estimation methods

IPAR increases with the test length, and the full range of *IPAR* gets tighter. Figure 4 shows the *IPAR* index under the 20-item test, and 1:4 new to operational ratio condition with different sample sizes. It is clear that the R-based statistics have higher *IPAR* indices, JEA outperforms SIE when the sample size is smaller than 600, and SIE has an equal or higher *IPAR* index than JEA when the sample size is 600 or higher. Figure 5 only provides the *IPAR* for the 20-item test with different sample sizes for the RMEM method, which shows that the proposed method performs better both when the items are highly discriminative and when the attribute mastery probability is uniform across attributes.

It is worth pointing out that although this method has promising performance in calibrating new items in small samples and theoretically does not depend on the initial value of the item parameters, it relies on accurate estimation of respondents' AMP. Therefore, the premise that the method does not depend on the initial item parameters is that AMPs are estimated sufficiently well based on the operational items. For that reason, the number of operational items taken by respondents and the number of respondents who take each new item should not be too small.

Real data example Due to the unavailability of a real dataset for CD-CAT, the real data example based on a dataset collected from a non-adaptive test is used for illustrative purposes for the proposed iterative two-step method. It is important to note that this does not mean that the proposed method is restricted to non-adaptive testing. One can view the application to non-adaptive testing as a special case where attribute profiles of test takers can be obtained based on the responses to the items with known attribute vectors (these items correspond to the operational items in adaptive testing), and the items that need to be estimated corresponds to new items in online calibration of CD-CAT. In fact, though the motivation for this approach was to develop an online calibration method for adaptive testing, the method can be used both for adaptive and non-adaptive tests.

The real dataset used here was collected from a learning experiment at the University of Tuebingen in Germany. The dataset contained responses from 504 examinees to 12 elementary probability theory problems that measure the following four attributes: (A1) calculate the classic probability of an event, (A2) calculate the probability of the complement of an event, (A3) calculate the probability of the union of two disjoint events, and (A4) calculate the probability of two independent events. The Q-matrix was initially produced by content experts and response data are available in the R package *pks* (Heller & Wickelmaier, 2013). Wang et al. (2020) applied several methods to estimate the Q-matrix by

Table 10 The *RMSE* (root mean squared error) of the item parameters for the new items with the sample size of 200

Group	L	λ	$s, g \sim U(0.15,0.35)$										$s, g \sim U(0.05,0.25)$									
			s_{RMSE}					g_{RMSE}					s_{RMSE}					g_{RMSE}				
			RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE
Uniform attribute mastery probability	20	1/4	0.118	0.172	**0.116**	0.149	0.138	**0.078**	0.093	0.090	0.108	0.081	**0.108**	0.121	0.111	0.120	**0.108**	**0.062**	0.084	0.063	0.086	0.063
		1/3	0.113	0.145	**0.102**	0.142	0.126	**0.071**	0.082	0.076	0.090	0.074	**0.095**	0.110	0.097	0.106	0.096	**0.055**	0.073	**0.055**	0.076	0.056
		1/2	0.108	0.144	**0.092**	0.140	0.120	0.072	0.085	**0.069**	0.078	0.075	0.088	0.108	**0.083**	0.094	0.091	0.052	0.068	**0.048**	0.066	0.053
	30	1/4	**0.100**	0.121	0.108	0.121	0.114	**0.060**	0.069	0.063	0.075	0.063	**0.086**	0.093	0.089	0.093	0.087	**0.050**	0.066	**0.050**	0.066	0.050
		1/3	**0.093**	0.114	0.097	0.112	0.104	**0.056**	0.064	0.057	0.069	0.058	0.081	0.086	0.081	0.087	**0.080**	**0.044**	0.059	**0.044**	0.059	**0.044**
		1/2	0.087	0.118	**0.084**	0.101	0.094	0.052	0.061	**0.050**	0.055	0.054	**0.063**	0.072	0.065	0.071	0.064	**0.036**	0.047	**0.036**	0.048	**0.036**
	40	1/4	**0.087**	0.102	0.096	0.104	0.096	**0.053**	0.062	0.055	0.065	0.055	**0.075**	0.082	0.077	0.082	**0.075**	**0.045**	0.059	**0.045**	0.059	**0.045**
		1/3	**0.080**	0.093	0.086	0.095	0.086	**0.045**	0.051	0.046	0.055	0.047	**0.066**	0.072	0.068	0.072	0.067	**0.037**	0.048	**0.037**	0.048	**0.037**
		1/2	**0.073**	0.090	0.077	0.086	0.076	0.042	0.048	**0.040**	0.047	0.043	**0.053**	0.058	**0.053**	0.058	**0.053**	**0.032**	0.040	**0.032**	0.040	**0.032**
Uneven attribute mastery probability	20	1/4	0.131	0.181	**0.120**	0.155	0.145	**0.090**	0.107	0.095	0.117	0.091	0.117	0.134	0.118	0.129	**0.114**	**0.070**	0.092	0.071	0.093	0.071
		1/3	0.124	0.156	**0.104**	0.152	0.136	**0.086**	0.101	**0.086**	0.099	0.088	0.108	0.131	0.114	0.117	**0.107**	**0.064**	0.083	**0.064**	0.083	0.066
		1/2	0.120	0.155	**0.099**	0.152	0.130	0.091	0.105	**0.084**	0.090	0.093	0.107	0.123	**0.100**	0.113	0.106	0.065	0.084	**0.056**	0.072	0.066
	30	1/4	**0.110**	0.136	0.115	0.128	0.124	0.070	0.081	**0.069**	0.085	0.071	0.098	0.103	0.100	0.102	**0.097**	**0.055**	0.072	**0.055**	0.072	**0.055**
		1/3	**0.103**	0.127	0.106	0.121	0.117	0.068	0.079	**0.066**	0.075	0.070	0.090	0.096	0.092	0.096	**0.089**	**0.049**	0.062	0.050	0.062	0.050
		1/2	0.099	0.122	**0.093**	0.115	0.108	0.072	0.086	**0.061**	0.067	0.074	0.076	0.084	0.077	0.083	**0.074**	**0.041**	0.052	**0.041**	0.052	**0.041**
	40	1/4	**0.101**	0.115	0.111	0.116	0.111	**0.060**	0.068	**0.060**	0.071	**0.060**	**0.090**	0.096	0.093	0.097	0.091	**0.051**	0.065	**0.051**	0.065	**0.051**
		1/3	**0.090**	0.104	0.096	0.103	0.097	0.055	0.064	**0.053**	0.062	0.057	**0.078**	0.083	0.081	0.083	0.079	**0.042**	0.054	**0.042**	0.054	**0.042**
		1/2	**0.083**	0.105	0.087	0.095	0.090	0.058	0.069	**0.049**	0.053	0.059	**0.067**	0.073	0.069	0.072	**0.067**	0.036	0.043	**0.035**	0.043	0.036

λ is the rate of new to operational items, and L refers to the test length. RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA, and SIE refer to the joint estimation algorithm and the single item estimation method, respectively. s and g refer to the slipping and the guessing parameters, respectively. s_{RMSE} and s_{RMSE} are the *RMSE*s of the slipping and the guessing parameters, respectively. Boldfaced values indicate the best performance across estimation methods

Table 11 The *RMSE* (root mean squared error) of the item parameters for the new items with the sample size of 400

Group	L	λ	$s, g \sim U(0.15,0.35)$										$s, g \sim U(0.05,0.25)$									
			s_{RMSE}					g_{RMSE}					s_{RMSE}					g_{RMSE}				
			RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE
Uniform attribute mastery probability	20	1/4	0.101	0.174	**0.098**	0.122	0.110	**0.062**	0.071	**0.062**	0.074	0.064	**0.082**	0.091	0.084	0.091	0.083	**0.045**	0.059	**0.045**	0.061	0.046
		1/3	0.101	0.137	**0.086**	0.117	0.104	0.063	0.068	**0.054**	0.062	0.060	**0.070**	0.080	0.072	0.079	0.071	0.040	0.051	**0.039**	0.055	0.040
		1/2	0.094	0.130	**0.082**	0.117	0.108	0.057	0.077	**0.053**	0.058	0.065	0.069	0.084	**0.063**	0.069	0.070	0.039	0.050	**0.034**	0.048	0.039
	30	1/4	**0.078**	0.093	0.083	0.095	0.084	0.047	0.053	**0.044**	0.054	0.047	**0.061**	0.068	0.062	0.068	0.061	0.035	0.045	**0.035**	0.045	**0.035**
		1/3	0.077	0.089	**0.073**	0.085	0.076	0.043	0.049	**0.041**	0.050	0.044	**0.055**	0.060	**0.055**	0.060	0.055	**0.032**	0.041	**0.032**	0.041	**0.032**
		1/2	0.073	0.115	**0.067**	0.078	0.078	0.045	0.054	**0.035**	0.040	0.046	0.049	0.054	**0.047**	0.054	0.049	0.026	0.033	**0.025**	0.034	0.026
	40	1/4	**0.066**	0.075	0.069	0.076	0.069	**0.036**	0.041	0.037	0.043	0.037	**0.054**	0.059	0.055	0.059	0.054	**0.031**	0.040	**0.031**	0.040	**0.031**
		1/3	**0.061**	0.070	**0.061**	0.071	0.062	0.034	0.038	**0.033**	0.040	0.034	**0.048**	0.053	**0.048**	0.053	0.048	**0.025**	0.034	**0.025**	0.034	**0.025**
		1/2	0.056	0.074	**0.053**	0.061	0.056	0.033	0.039	**0.028**	0.033	0.033	**0.040**	0.044	**0.040**	0.044	**0.040**	0.023	0.028	**0.022**	0.028	0.023
Uneven attribute mastery probability	20	1/4	0.107	0.138	**0.100**	0.134	0.121	0.078	0.091	**0.071**	0.085	0.080	**0.094**	0.104	0.095	0.102	0.092	0.050	0.065	**0.049**	0.067	0.050
		1/3	0.106	0.149	**0.091**	0.126	0.113	0.081	0.094	**0.070**	0.077	0.083	**0.081**	0.093	0.083	0.090	0.079	0.045	0.056	**0.043**	0.057	0.045
		1/2	0.107	0.186	**0.092**	0.131	0.117	0.087	0.104	**0.069**	0.070	0.089	0.080	0.103	**0.077**	0.082	0.080	0.053	0.066	**0.040**	0.053	0.053
	30	1/4	**0.088**	0.108	0.094	0.106	0.098	0.057	0.066	**0.051**	0.059	0.058	**0.072**	0.077	0.075	0.077	0.073	**0.038**	0.048	**0.038**	0.048	**0.038**
		1/3	**0.079**	0.104	0.086	0.095	0.088	0.058	0.068	**0.048**	0.054	0.059	**0.065**	0.071	0.067	0.071	0.065	**0.034**	0.044	**0.034**	0.044	**0.034**
		1/2	0.082	0.128	**0.079**	0.091	0.088	0.066	0.082	**0.049**	0.048	0.067	**0.055**	0.060	0.056	0.060	0.056	0.030	0.038	**0.029**	0.037	0.030
	40	1/4	**0.076**	0.086	0.083	0.087	0.083	0.045	0.052	**0.044**	0.052	0.047	0.063	0.068	0.063	0.068	**0.062**	**0.034**	0.043	**0.034**	0.043	**0.034**
		1/3	**0.066**	0.078	0.071	0.078	0.071	0.044	0.050	**0.039**	0.044	0.044	**0.053**	0.059	0.055	0.058	**0.053**	0.030	0.036	**0.029**	0.036	0.030
		1/2	**0.067**	0.089	**0.067**	0.072	0.070	0.051	0.063	**0.037**	0.038	0.051	**0.044**	0.048	0.045	0.048	**0.044**	**0.025**	0.031	**0.025**	0.031	**0.025**

λ is the rate of new to operational items, and L refers to the test length. RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA, and SIE refer to the joint estimation algorithm and the single item estimation method, respectively. s and g refer to the slipping and the guessing parameters, respectively. s_{RMSE} and s_{RMSE} are the *RMSE*s of the slipping and the guessing parameters, respectively. Boldfaced values indicate the best performance across estimation methods

Table 12 The *RMSE* (root mean squared error) of the item parameters for the new items with the sample size of 600

Group	L	λ	$s, g \sim U(0.15,0.35)$										$s, g \sim U(0.05,0.25)$									
			s_{RMSE}					g_{RMSE}					s_{RMSE}					g_{RMSE}				
			RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE
Uniform attribute mastery probability	20	1/4	0.105	0.172	**0.087**	0.108	0.096	0.055	0.064	**0.049**	0.059	0.056	**0.067**	0.075	0.068	0.075	0.067	**0.037**	0.048	**0.037**	0.050	0.038
		1/3	0.090	0.141	**0.076**	0.105	0.110	0.062	0.077	**0.045**	0.047	0.063	0.059	0.067	**0.057**	0.065	0.059	0.033	0.042	**0.031**	0.045	0.033
		1/2	0.090	0.124	**0.078**	0.101	0.094	0.056	0.066	**0.045**	0.050	0.056	0.062	0.077	**0.052**	0.060	0.063	0.035	0.044	**0.027**	0.040	0.035
	30	1/4	0.073	0.113	**0.071**	0.081	0.072	0.039	0.044	**0.036**	0.043	0.039	**0.051**	0.055	0.052	0.056	0.052	**0.029**	0.037	**0.029**	0.037	**0.029**
		1/3	0.068	0.081	**0.063**	0.074	0.067	0.038	0.044	**0.032**	0.039	0.038	**0.047**	0.050	**0.047**	0.050	**0.047**	**0.025**	0.032	**0.025**	0.032	**0.025**
		1/2	0.064	0.081	**0.058**	0.066	0.073	0.042	0.051	**0.031**	0.034	0.043	0.038	0.043	**0.037**	0.044	0.038	0.022	0.028	**0.021**	0.028	0.022
	40	1/4	**0.057**	0.063	0.058	0.064	0.058	0.032	0.035	**0.031**	0.037	0.032	**0.044**	0.049	**0.044**	0.049	**0.044**	**0.025**	0.033	**0.025**	0.033	**0.025**
		1/3	0.051	0.059	**0.049**	0.057	0.051	0.029	0.033	**0.027**	0.033	0.029	0.038	0.042	**0.038**	0.042	**0.038**	**0.022**	0.028	**0.022**	0.028	**0.022**
		1/2	0.053	0.072	**0.043**	0.051	0.053	0.030	0.036	**0.023**	0.028	0.030	**0.032**	0.036	**0.032**	0.036	**0.032**	**0.018**	0.024	**0.018**	0.024	**0.018**
Uneven attribute mastery probability	20	1/4	0.111	0.140	**0.096**	0.120	0.111	0.075	0.089	**0.062**	0.069	0.076	0.078	0.085	0.080	0.085	**0.079**	0.043	0.053	**0.042**	0.054	0.043
		1/3	0.102	0.157	**0.089**	0.118	0.106	0.079	0.094	**0.060**	0.065	0.081	0.072	0.080	**0.072**	0.076	0.073	0.040	0.050	**0.036**	0.050	0.040
		1/2	0.099	0.189	**0.089**	0.121	0.117	0.089	0.104	**0.065**	0.064	0.090	0.075	0.093	**0.065**	0.069	0.076	0.050	0.062	**0.035**	0.043	0.050
	30	1/4	**0.077**	0.095	0.080	0.088	0.084	0.053	0.063	**0.044**	0.050	0.054	0.057	0.061	0.059	0.061	**0.058**	**0.031**	0.039	**0.031**	0.039	**0.031**
		1/3	**0.073**	0.096	0.074	0.079	0.077	0.056	0.067	**0.044**	0.047	0.057	0.054	0.057	0.054	0.058	**0.053**	**0.028**	0.035	**0.028**	0.035	**0.028**
		1/2	0.075	0.134	**0.074**	0.081	0.086	0.069	0.085	0.046	**0.043**	0.069	0.046	0.051	**0.045**	0.050	0.047	0.027	0.033	**0.025**	0.032	0.027
	40	1/4	**0.064**	0.074	0.072	0.074	0.069	0.040	0.045	**0.037**	0.043	0.040	**0.053**	0.057	**0.053**	0.057	**0.053**	**0.028**	0.035	**0.028**	0.035	**0.028**
		1/3	**0.057**	0.070	0.060	0.068	0.060	0.040	0.047	**0.033**	0.036	0.040	**0.043**	0.046	**0.043**	0.046	**0.043**	**0.024**	0.030	**0.024**	0.030	**0.024**
		1/2	0.059	0.086	**0.057**	0.058	0.061	0.051	0.064	0.035	**0.032**	0.051	**0.036**	0.040	**0.036**	0.040	**0.036**	**0.020**	0.025	**0.020**	0.025	**0.020**

λ is the rate of new to operational items, and L refers to the test length. RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA, and SIE refer to the joint estimation algorithm and the single item estimation method, respectively. s and g refer to the slipping and the guessing parameters, respectively. s_{RMSE} and s_{RMSE} are the *RMSE*s of the slipping and the guessing parameters, respectively. Boldfaced values indicate the best performance across estimation methods

Table 13 The *RMSE* (root mean squared error) of the item parameters for the new items with the sample size of 800

Group	L	λ	s, g~U(0.15,0.35)										s, g~U(0.05,0.25)									
			s_{RMSE}					g_{RMSE}					$sRMSE$					g_{RMSE}				
			RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE
Uniform attribute mastery probability	20	1/4	0.083	0.117	**0.079**	0.098	0.087	0.049	0.058	**0.044**	0.053	0.050	**0.058**	0.065	**0.058**	0.065	0.059	**0.033**	0.041	**0.033**	0.044	**0.033**
		1/3	0.082	0.133	**0.070**	0.094	0.086	0.051	0.061	**0.039**	0.045	0.051	0.050	0.058	**0.049**	0.057	0.051	0.029	0.036	**0.028**	0.039	0.029
		1/2	0.095	0.175	**0.070**	0.090	0.099	0.059	0.073	**0.041**	0.041	0.059	0.053	0.068	**0.043**	0.052	0.054	0.032	0.040	**0.024**	0.036	0.032
	30	1/4	**0.058**	0.072	0.059	0.070	0.060	0.034	0.039	**0.031**	0.039	0.035	**0.044**	0.048	**0.044**	0.048	**0.044**	**0.025**	0.031	**0.025**	0.031	**0.025**
		1/3	0.059	0.076	**0.052**	0.063	0.058	0.032	0.038	**0.027**	0.034	0.032	**0.039**	0.044	**0.039**	0.044	**0.039**	**0.023**	0.028	**0.023**	0.028	**0.023**
		1/2	0.067	0.108	**0.051**	0.057	0.067	0.039	0.048	**0.027**	0.030	0.039	0.034	0.039	**0.033**	0.039	0.034	**0.019**	0.024	**0.019**	0.024	**0.019**
	40	1/4	**0.048**	0.055	**0.048**	0.056	0.049	0.028	0.031	**0.027**	0.033	0.028	**0.040**	0.045	**0.040**	0.045	**0.040**	**0.023**	0.029	**0.023**	0.029	**0.023**
		1/3	0.045	0.053	**0.042**	0.051	0.045	0.025	0.029	**0.023**	0.028	0.025	**0.034**	0.038	**0.034**	0.038	**0.034**	**0.019**	0.024	**0.019**	0.024	**0.019**
		1/2	0.047	0.067	**0.037**	0.046	0.047	0.028	0.034	**0.020**	0.026	0.028	**0.028**	0.030	**0.028**	0.030	**0.028**	**0.016**	0.020	**0.016**	0.020	**0.016**
Uneven attribute mastery probability	20	1/4	0.092	0.137	**0.090**	0.106	0.100	0.069	0.084	**0.056**	0.062	0.070	**0.065**	0.073	0.069	0.073	0.067	0.036	0.046	**0.034**	0.047	0.036
		1/3	0.095	0.154	**0.083**	0.106	0.100	0.076	0.092	**0.054**	0.058	0.077	**0.062**	0.070	0.063	0.066	0.063	0.035	0.043	**0.031**	0.043	0.035
		1/2	0.105	0.189	**0.087**	0.107	0.113	0.087	0.104	**0.062**	0.058	0.088	0.068	0.085	**0.057**	0.058	0.069	0.047	0.058	**0.030**	0.037	0.047
	30	1/4	**0.068**	0.085	0.072	0.078	0.072	0.047	0.057	**0.037**	0.042	0.047	**0.051**	0.056	0.052	0.056	**0.051**	0.026	0.034	**0.026**	0.034	0.026
		1/3	**0.066**	0.090	0.069	0.072	0.070	0.051	0.062	**0.038**	0.040	0.051	**0.044**	0.049	0.045	0.049	**0.044**	**0.025**	0.031	**0.025**	0.031	**0.025**
		1/2	0.075	0.130	**0.065**	0.067	0.078	0.067	0.084	0.043	**0.039**	0.067	0.040	0.044	**0.038**	0.044	0.040	0.024	0.029	**0.022**	0.029	0.024
	40	1/4	**0.061**	0.069	0.062	0.069	0.063	0.035	0.041	**0.032**	0.037	0.036	**0.044**	0.047	0.045	0.047	**0.044**	**0.024**	0.031	**0.024**	0.031	**0.024**
		1/3	**0.051**	0.063	0.052	0.058	0.052	0.037	0.044	**0.029**	0.032	0.037	**0.038**	0.042	**0.038**	0.042	**0.038**	0.020	0.025	**0.019**	0.025	0.020
		1/2	0.055	0.082	**0.048**	0.052	0.057	0.050	0.063	**0.033**	0.030	0.050	**0.033**	0.036	**0.033**	0.036	**0.033**	0.018	0.021	**0.017**	0.021	0.018

λ is the rate of new to operational items, and L refers to the test length. RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA, and SIE refer to the joint estimation algorithm and the single item estimation method, respectively. s and g refer to the slipping and the guessing parameters, respectively. s_{RMSE} and s_{RMSE} are the *RMSE*s of the slipping and the guessing parameters, respectively. Boldfaced values indicate the best performance across estimation methods

Table 14 The *RMSE* (root mean squared error) of the item parameters for the new items with the sample size of 1000

Group	L	λ	$s, g \sim U(0.15,0.35)$										$s, g \sim U(0.05,0.25)$									
			s_{RMSE}					g_{RMSE}					s_{RMSE}					g_{RMSE}				
			RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE	RMA	ROEM	RMEM	JEA	SIE
Uniform attribute mastery probability	20	1/4	0.082	0.119	**0.073**	0.085	0.084	0.047	0.057	**0.037**	0.045	0.048	**0.051**	0.057	**0.051**	0.057	0.052	**0.029**	0.037	**0.029**	0.038	**0.029**
		1/3	0.087	0.139	**0.065**	0.083	0.089	0.051	0.061	**0.036**	0.041	0.051	0.046	0.053	**0.044**	0.052	0.046	0.027	0.034	**0.024**	0.034	0.027
		1/2	0.104	0.183	**0.066**	0.080	0.108	0.059	0.074	**0.038**	0.035	0.059	0.054	0.068	**0.039**	0.048	0.054	0.030	0.039	**0.021**	0.032	0.030
	30	1/4	0.057	0.070	**0.055**	0.064	0.058	0.032	0.037	**0.027**	0.033	0.033	**0.039**	0.043	**0.039**	0.043	0.039	**0.021**	0.028	**0.021**	0.028	**0.021**
		1/3	0.057	0.077	**0.050**	0.058	0.058	0.033	0.039	**0.025**	0.031	0.033	**0.035**	0.039	**0.035**	0.039	**0.035**	**0.020**	0.025	**0.020**	0.025	**0.020**
		1/2	0.067	0.112	**0.044**	0.049	0.068	0.040	0.049	**0.024**	0.027	0.040	0.030	0.034	**0.029**	0.034	0.030	0.017	0.022	**0.016**	0.022	0.017
	40	1/4	0.046	0.052	**0.045**	0.053	0.046	0.025	0.029	**0.024**	0.030	0.025	**0.035**	0.039	**0.035**	0.039	**0.035**	**0.020**	0.026	**0.020**	0.026	**0.020**
		1/3	0.042	0.050	**0.039**	0.048	0.042	0.023	0.027	**0.019**	0.025	0.023	**0.028**	0.032	**0.028**	0.032	**0.028**	**0.017**	0.022	**0.017**	0.022	**0.017**
		1/2	0.046	0.067	**0.033**	0.042	0.046	0.028	0.034	**0.018**	0.023	0.028	**0.024**	0.027	**0.024**	0.027	**0.024**	**0.014**	0.018	**0.014**	0.018	**0.014**
Uneven attribute mastery probability	20	1/4	0.089	0.135	**0.085**	0.103	0.097	0.071	0.086	**0.053**	0.058	0.072	**0.059**	0.066	0.061	0.066	0.060	0.032	0.041	**0.031**	0.042	0.032
		1/3	0.093	0.156	**0.079**	0.099	0.097	0.078	0.093	**0.055**	0.056	0.079	0.056	0.064	**0.055**	0.061	0.056	0.035	0.041	**0.030**	0.039	0.035
		1/2	0.109	0.191	**0.087**	0.104	0.116	0.088	0.104	**0.062**	0.057	0.089	0.065	0.082	**0.052**	0.055	0.065	0.047	0.056	**0.029**	0.034	0.047
	30	1/4	**0.064**	0.084	0.066	0.072	0.068	0.048	0.058	**0.037**	0.040	0.048	**0.045**	0.049	0.047	0.049	0.046	**0.024**	0.030	**0.024**	0.030	**0.024**
		1/3	0.065	0.090	**0.063**	0.069	0.068	0.050	0.062	**0.035**	0.036	0.050	**0.040**	0.045	0.041	0.045	0.040	**0.022**	0.027	**0.022**	0.027	**0.022**
		1/2	0.075	0.132	**0.062**	0.064	0.078	0.066	0.083	**0.042**	0.037	0.067	0.037	0.040	**0.035**	0.040	0.037	0.021	0.025	**0.019**	0.025	0.021
	40	1/4	**0.053**	0.061	0.056	0.061	0.055	0.033	0.039	**0.029**	0.035	0.033	**0.039**	0.043	0.040	0.043	0.039	**0.022**	0.028	**0.022**	0.028	**0.022**
		1/3	0.049	0.061	**0.047**	0.055	0.050	0.036	0.044	**0.027**	0.029	0.036	**0.034**	0.037	**0.034**	0.037	**0.034**	**0.018**	0.023	**0.018**	0.023	**0.018**
		1/2	0.052	0.080	**0.043**	0.047	0.054	0.048	0.061	**0.031**	0.026	0.048	**0.029**	0.032	**0.029**	0.032	**0.029**	**0.016**	0.019	**0.016**	0.019	**0.016**

λ is the rate of new to operational items, and L refers to the test length. RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA, and SIE refer to the joint estimation algorithm and the single item estimation method, respectively. s and g refer to the slipping and the guessing parameters, respectively. s_{RMSE} and s_{RMSE} are the *RMSE*s of the slipping and the guessing parameters, respectively. Boldfaced values indicate the best performance across estimation methods

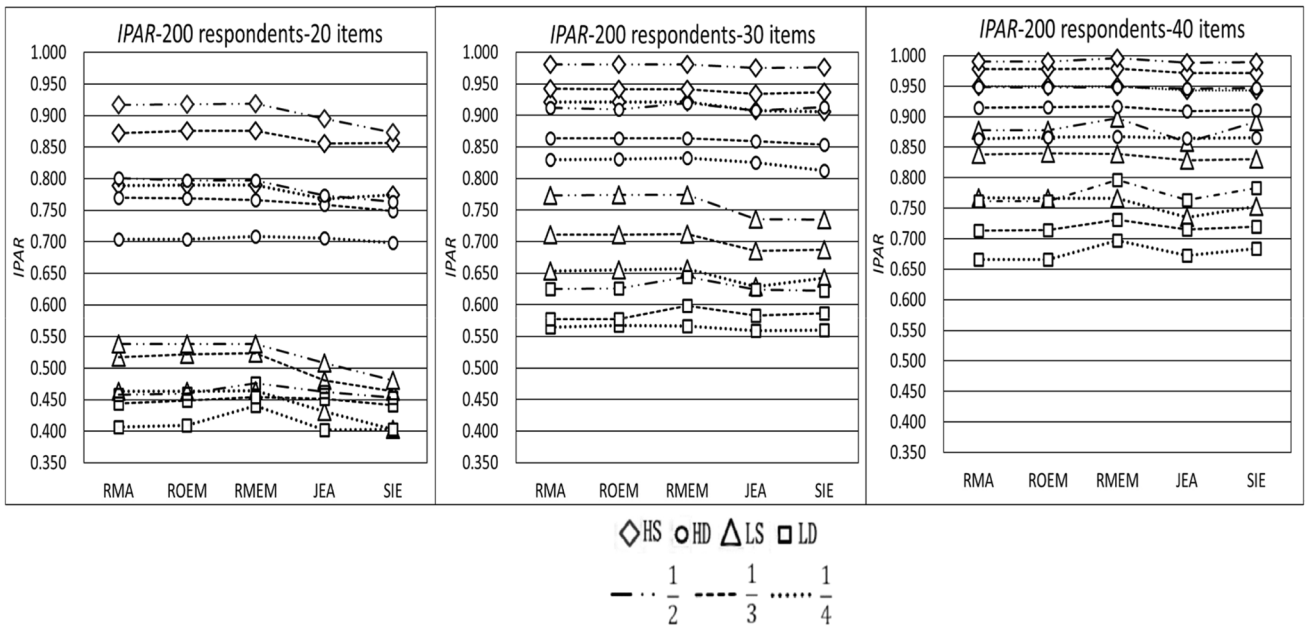

Fig. 3 The *IPAR* (Item Pattern Accuracy Rate) in different test lengths with 200 respondents. Note. The first letter 'H' or 'L' in the legend refer to items with high- or low-discrimination, the second letter 'S' or 'D' refer to respondents with the same or different attribute mastery probability (ies), and $\frac{1}{2}$, $\frac{1}{3}$, or $\frac{1}{4}$ denote the rate of new to operational items. RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA and SIE refer to the joint estimation algorithm and the single item estimation method, respectively

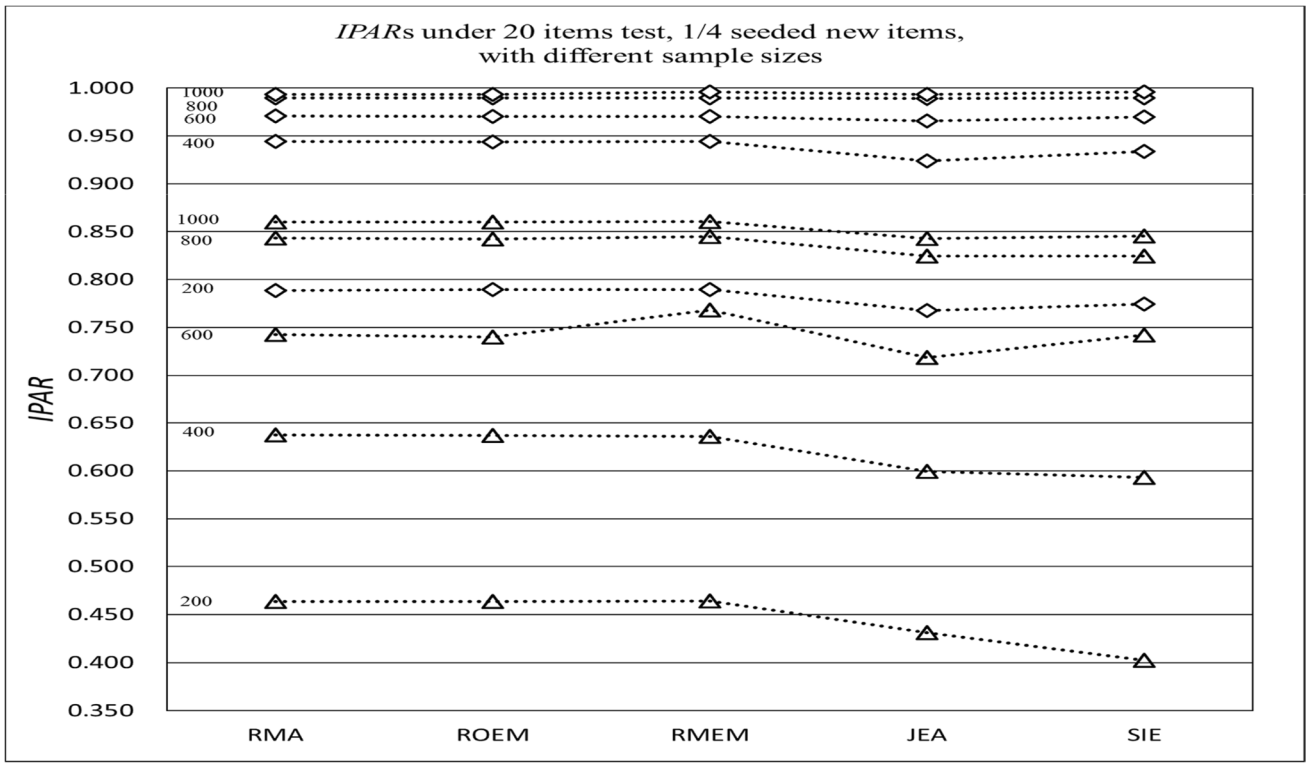

Fig. 4 The *IPAR* (Item Pattern Accuracy Rate) in the 20-item test with 1/4 seeded new items under different sample sizes. Note. The first letter 'H' or 'L' in the legend refer to items with high- or low-discrimination, and the second letter 'S' or 'D' refer to respondents with the same or different attribute mastery probability (ies). RMA, ROEM, and RMEM are variations of CD-MA, CD-OEM, and CD-MEM, respectively. JEA, and SIE refer to the joint estimation algorithm and the single item estimation method, respectively

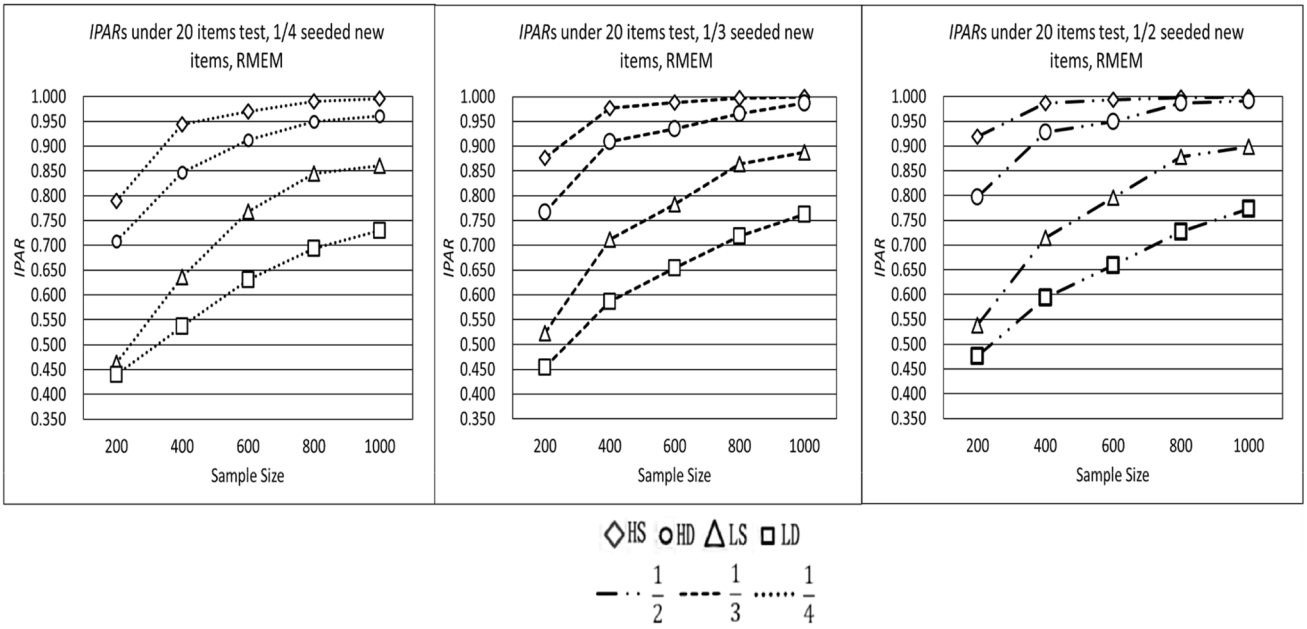

Fig. 5 The *IPAR* (Item Pattern Accuracy Rate) for the RMEM method with different sample sizes in the 20-item test. Note. The first letter 'H' or 'L' in the legend refer to items with high- or low-discrimination, the second letter 'S' or 'D' refer to respondents with the same or different attribute mastery probability (ies), and $\frac{1}{2}$, $\frac{1}{3}$, or $\frac{1}{4}$ denotes the rate of new to operational items. RMEM is a variation of the CD-MEM method

treating eight of the 12 items as operational items and the remaining four as new. Here we follow a similar strategy, i.e., we consider eight of 12 items as the operational items, which are items 1, 2, 3, 4, 6, 7, 9, 11, and the remaining four items (items 5, 8, 10, 12) as new. The Q-matrix for the operational items and the original Q-matrix for the new items in the package *pks* are given in Table 15.

Responses to the eight operational items are referred to as $\mathbf{X}^O$, and responses to the four new items are referred to as $\mathbf{X}^N$. Based on the two-step on-line item calibration method, we follow the process below to obtain the Q-matrix for the new items:

(1) Obtain the estimates of the attribute profile $\hat{\boldsymbol{\alpha}}$ of each examinee based on the $\mathbf{X}^O$,
(2) Assign the initial slipping and guessing parameter as 0.25, estimate the attribute vector for each new item based on the proposed *R* statistic,
(3) Based on the attribute-vectors obtained from the last step, apply the CD-MEM method to estimate the slipping and guessing parameters,
(4) Repeat step 2 to step 3 till the convergence condition reaches.

The estimated $\boldsymbol{Q}$-matrix for the new items is presented at the bottom of Table 15. The proposed method suggested four changes to the original Q-matrix, which are all from 1 to 0.

Table 15 The Q-matrix for the operational items, and the original and suggested Q-matrix for the new items

	Item	A1	A2	A3	A4
Operational items	1	1	0	0	0
	2	0	1	0	0
	3	0	0	1	0
	4	0	0	0	1
	6	1	1	0	0
	7	1	0	1	0
	9	1	0	0	1
	11	1	1	0	1
New items (original)	5	1	1	0	0
	8	1	0	1	0
	10	0	1	0	1
	12	1	0	1	1
New items (Suggested by RMEM)	5	1	**0***	0	0
	8	**0***	0	1	0
	10	0	**0***	0	1
	12	1	0	**0***	1

The entries with an asterisk in bold are different from the original Q-matrix. The four attributes in the table are: A1 - Calculate the classic probability of an event; A2 - Calculate the probability of the complement of an event; A3 - Calculate the probability of the union of two disjoint events; and A4 - Calculate the probability of two independent events

This seems to indicate that the proposed method tends to assign fewer attributes to each new item. Take the first new item (its name is p105 in the *pks* package) as an example, whose stem is "Given a standard deck containing 32 different cards, what is the probability of not drawing a heart?" The RMEM suggests that it only measures the attribute A1, which is to calculate the classic probability of an event. The original attribute specification of this item is A1 and A2, where A2 refers to "calculate the probability of the complement of an event". Based on our analysis, it does not seem to require mastery of A2 to answer this item. The estimated **Q**-matrix could serve as a reference for domain experts, who can further review the changes.

Conclusions and further discussion

In this paper, we proposed a method based on a residual-based statistic to estimate attribute vectors of new items in the online calibration of CD-CAT. The rationale of the use of the residual-based statistic in online calibration is presented in Appendix A. Essentially, the residual statistic is minimized when the attribute vector of a new item is at its true value, regardless of the item parameters. An iterative two-step online calibration method was thus developed in the context of CD-CAT in which the attribute vectors and item parameters are estimated in separate steps iteratively. By coupling CD-MA, CD-OEM, and CD-MEM with the residual-based statistic, three new online calibration methods: RMA, ROEM, and RMEM, are developed. The analytical result in Appendix A holds when $N \to \infty$ and the AMPs of respondents are known. When the AMPs need to be estimated, and the sample size is limited, the performance of RMA, ROEM, and RMEM are not guaranteed to be optimal, but could still be superior to existing methods.

The results from the simulation study indicate that the methods based on the proposed statistics do work well in terms of item-parameter recovery, and attribute-vector recovery, even under a small sample size. Compared to the JEA and SIE methods, the methods based on the residual statistic show some advantages, especially in the situation of a small sample size. Results also suggest that RMA and ROEM perform similarly in the estimation of the attribute vector of the new items, and RMA and SIE have similar performance in the estimation of the item parameters of the new items, especially in the test with highly discriminative items. For a CD-CAT system, quality of items (operational items and new items) is very important because it can seriously affect the efficiency and accuracy of the test, and the online calibration as well.

Several future directions for research need to be considered. First, the **Q**-matrix in this study is generated assuming that attributes are independent. However, in more realistic conditions, some relationships may exist among the attributes, such as hierarchical relationships (Leighton et al., 2004). Non-independence may impact the performance of the proposed methods, which is worthy of investigation in the future. Second, the proposed methods were evaluated under the DINA model, and it should be adapted to many other CDMs, such as RRUM (Hartz, 2002), DINO (Templin & Henson, 2006) and more general models (e.g., Ma & de la Torre, 2016, 2019) such as the G-DINA model (de la Torre, 2011). Under the G-DINA model, each respondent is classified into one of the $2^{k_m^*}$ groups, where $k_m^* = \sum_{k=1}^{K} q_{mk}$. Then the residual statistic defined in Eq. (6) can be adapted as follows for the G-DINA model:

$$R_m(\boldsymbol{\alpha}, \mathbf{q}_m, s_m, g_m) = 2\sum_{l=1}^{2^{k_m^*}} \sum_{i=1}^{n_{lm}} \log\left\{\left[\frac{1-p(\alpha_{lm}^*)}{p(\alpha_{lm}^*)}\right]^{x_{im}} + \left[\frac{p(\alpha_{lm}^*)}{1-p(\alpha_{lm}^*)}\right]^{1-x_{im}}\right\}, \quad (15)$$

where n_{lm} refers to the number of respondents with attribute vector $\boldsymbol{\alpha}_{lm}^*$, and $\boldsymbol{\alpha}_{lm}^* = \left(\alpha_{lm_1}, \cdots, \alpha_{lm_{k_m^*}}\right)$. The probability that respondents with attribute pattern $\boldsymbol{\alpha}_{lm}^*$ will answer item m correctly is denoted by $p\left(X_{im} = 1 | \boldsymbol{\alpha}_{lm}^*\right) = p\left(\boldsymbol{\alpha}_{lm}^*\right)$. By defining an appropriate residual statistic, the proposed method in this paper is potentially applicable to other models. That said, it remains to be investigated how well the adapted residual statistic works, and whether nice statistical properties such as what is demonstrated in Theorem 1 still holds true for other models.

Third, the study assumes that the attribute vectors and item parameters of all the operational items are known. In reality, those must have been estimated or specified by content experts at some point. How will the proposed methods perform when the attribute vectors or the item parameters or both for some of the operational items are misspecified? How badly will different methods react to the misspecification? These are issues yet to be investigated. Finally, recent popularity of online learning environments has prompted advances in continuous item calibration that may not require any operational items to begin with (Fink et al., 2018) for CAT. The same philosophy may be applicable to CD-CAT and is certainly an interesting direction to pursue.

Appendix A

Theorem 1. Consider an infinite sample, that is $N \to \infty$, and the true item parameters $s_j, g_j \in (0, 0.5)$. Denote $\hat{\boldsymbol{\alpha}}$ as the estimate of $\boldsymbol{\alpha}$. Furthermore, assume its true value $\boldsymbol{\alpha}$ is known in advance. Given the provisional item parameters for the j^{th} item $\left(s_j^0, g_j^0\right)$, where s_j^0, g_j^0 are two arbitrary real numbers within the range of $(0, 0.5)$, denote $R_j\left(\boldsymbol{\alpha}, \mathbf{q}_j^*, s_j^0, g_j^0\right)$ as the value of the residual-based statistic when the item parameters

and attribute vector are assigned as $\left(s_j^0, g_j^0\right)$ and $\mathbf{q}_j^*$, respectively. Then $R_j\left(\boldsymbol{\alpha}, \mathbf{q}_j^*, s_j^0, g_j^0\right)$ reaches its minimum only when the attribute vector of the j^{th} item, $\mathbf{q}_j^*$, is correctly specified.

Proof.

Based on the $\mathbf{q}_j$ and $\boldsymbol{\alpha}$, the respondents can be categorized into four groups G_1, G_2, G_3, and G_4, with N_{11}^j, N_{10}^j, N_{01}^j, and N_{00}^j respondents, respectively. The two numbers in the subscript of N_{11}^j, N_{10}^j, N_{01}^j, or N_{00}^j are the values of the ideal response η_{ij} and the response X_{ij}. Respondents in G_1 and G_2 possess all the required attributes of item j, while respondents in G_3 and G_4 miss at least one of the required attributes of item j. Respondents in G_1 and G_3 answer the item correctly, but not the respondents in G_2 and G_4. Eq. (5) can be transformed to

$$\begin{aligned} &R_j(\boldsymbol{\alpha}, \mathbf{q}_j, s_j, g_j) \\ &= 2\left[N_{11}^j \log\left(\tfrac{s_j}{1-s_j}\right) + N_{10}^j \log\left(\tfrac{1-s_j}{s_j}\right) + N_{01}^j \log\left(\tfrac{1-g_j}{g_j}\right) + N_{00}^j \log\left(\tfrac{g_j}{1-g_j}\right)\right]. \end{aligned} \tag{A-1}$$

When $N \to \infty$, it is expected that $N_{11}^j > N_{10}^j$, and $N_{00}^j > N_{01}^j$. Substituting s_j and g_j in A-1with s_j^0 and g_j^0, we get

$$\begin{aligned} &R_j\left(\boldsymbol{\alpha}, \mathbf{q}_j, s_j^0, g_j^0\right) \\ &= 2\left[N_{11}^j \log\left(\tfrac{s_j^0}{1-s_j^0}\right) + N_{10}^j \log\left(\tfrac{1-s_j^0}{s_j^0}\right) + N_{01}^j \log\left(\tfrac{1-g_j^0}{g_j^0}\right) + N_{00}^j \log\left(\tfrac{g_j^0}{1-g_j^0}\right)\right] \\ &= 2\left[\left(N_{11}^j - N_{10}^j\right) \log\left(\tfrac{s_j^0}{1-s_j^0}\right) + \left(N_{00}^j - N_{01}^j\right) \log\left(\tfrac{g_j^0}{1-g_j^0}\right)\right], \end{aligned} \tag{A-2}$$

Given s_j^0 and $g_j^0 \in (0, 0.5)$, $R_j\left(\boldsymbol{\alpha}, \mathbf{q}_j, s_j^0, g_j^0\right)$ should always be negative. The number of respondents in the four groups may change with the value that $\mathbf{q}_j$ takes. For $\mathbf{q}_j = \mathbf{q}_j^*$, the corresponding number of respondents in each group can be denoted as N_{11}^{j*}, N_{10}^{j*}, N_{01}^{j*}, and N_{00}^{j*}, respectively. The difference between $R_j\left(\boldsymbol{\alpha}, \mathbf{q}_j^*, s_j^0, g_j^0\right)$ and $R_j\left(\boldsymbol{\alpha}, \mathbf{q}_j, s_j^0, g_j^0\right)$, can then be defined as

$$\Delta = R_j\left(\boldsymbol{\alpha}, \mathbf{q}_j^*, s_j^0, g_j^0\right) - R_j\left(\boldsymbol{\alpha}, \mathbf{q}_j, s_j^0, g_j^0\right), \tag{A-3}$$

where

$$R_j\left(\boldsymbol{\alpha}, \mathbf{q}_j^*, s_j^0, g_j^0\right) = 2\left[\left(N_{11}^{j*} - N_{10}^{j*}\right) \log\left(\frac{s_j^0}{1-s_j^0}\right) + \left(N_{00}^{j*} - N_{01}^{j*}\right) \log\left(\frac{g_j^0}{1-g_j^0}\right)\right]. \tag{A-4}$$

In the following discussion, we only consider the conditions when $\mathbf{q}_j^*$ does not match the true value of $\mathbf{q}_j$, which includes the following three cases.

Case 1: In addition to all the required attributes in $\mathbf{q}_j$, $\mathbf{q}_j^*$ also contains some unnecessary attributes, for example, $\mathbf{q}_j = [1\ 1\ 0]$, $\mathbf{q}_j^* = [1\ 1\ 1]$. In this case, some of the respondents in G_1 and G_2 will be wrongfully categorized into G_3 (denote as ΔN_{11}) and G_4 (denote as ΔN_{10}). When $N \to \infty$, $\Delta N_{11} > \Delta N_{10}$. Also note the increase in G_3 is exactly the decrease in G_1, and the increase in G_4 is exactly the decrease in G_2. That is, $N_{11}^{j*} = N_{11}^j - \Delta N_{11}$, $N_{10}^{j*} = N_{10}^j - \Delta N_{10}$, $N_{01}^{j*} = N_{01}^j + \Delta N_{11}$, and $N_{00}^{j*} = N_{00}^j + \Delta N_{10}$. Therefore, A-3 becomes

$$\begin{aligned} \Delta &= 2\left[\left(N_{11}^{j*} - N_{10}^{j*}\right) \log\left(\tfrac{s_j^0}{1-s_j^0}\right) + \left(N_{00}^{j*} - N_{01}^{j*}\right) \log\left(\tfrac{g_j^0}{1-g_j^0}\right)\right] \\ &-2\left[\left(N_{11}^j - N_{10}^j\right) \log\left(\tfrac{s_j^0}{1-s_j^0}\right) + \left(N_{00}^j - N_{01}^j\right) \log\left(\tfrac{g_j^0}{1-g_j^0}\right)\right] \\ &= 2\left(\Delta N_{10} - \Delta N_{11}\right)\left[\log\left(\tfrac{s_j^0}{1-s_j^0}\right) + \log\left(\tfrac{g_j^0}{1-g_j^0}\right)\right] \\ &= 2\left(\Delta N_{10} - \Delta N_{11}\right) \log\left(\tfrac{s_j^0}{1-s_j^0} \tfrac{g_j^0}{1-g_j^0}\right). \end{aligned} \tag{A-5}$$

On one hand,$\log\left(\frac{s_j^0}{1-s_j^0}\frac{g_j^0}{1-g_j^0}\right)$ is a constant and negative; on the other hand, $\Delta N_{10} - \Delta N_{11} \le 0$, therefore, $\Delta \ge 0$. In other words, the misspecification in this case is expected to lead to an increase of the R statistic, except when the $\mathbf{q}^*$ is correctly specified (i.e., $\mathbf{q}_j = \mathbf{q}_j^*$), in which case $\Delta = 0$.

Case 2: In this case $\mathbf{q}^*$ lacks some required attributes of $\mathbf{q}_j$, for example, $\mathbf{q}_j = [1\ 1\ 0]$, $\mathbf{q}^* = [1\ 0\ 0]$. This means that ΔN_{01} respondents in G_3 will be wrongfully categorized into G_1, and ΔN_{00} will be wrongfully categorized from G_4 to G_2, respectively. When $N \to \infty$, $\Delta N_{01} < \Delta N_{00}$. That is, $N_{11}^* = N_{11}^j + \Delta N_{01}$, $N_{10}^* = N_{10}^j + \Delta N_{00}$, $N_{01}^* = N_{01}^j - \Delta N_{01}$, and $N_{00}^* = N_{00}^j - \Delta N_{00}$. Therefore, A-3 becomes

$$\begin{aligned} \Delta &= 2\left(\Delta N_{01} - \Delta N_{00}\right)\left[\log\left(\frac{s_j^0}{1-s_j^0}\right) + \log\left(\frac{g_j^0}{1-g_j^0}\right)\right] \\ &= 2\left(\Delta N_{01} - \Delta N_{00}\right) \log\left(\frac{s_j^0}{1-s_j^0}\frac{g_j^0}{1-g_j^0}\right). \end{aligned} \tag{A-6}$$

Because $\Delta N_{01} - \Delta N_{00} \le 0$, $\Delta \ge 0$. The misspecification in this case is also expected to lead to an increase of the R statistic, except when the $\mathbf{q}^*$ is correctly specified (i.e., $\mathbf{q}_j = \mathbf{q}_j^*$), in which case $\Delta = 0$.

Case 3: We consider a more complex situation where $\mathbf{q}^*$ lacks some of the required attributes, while containing some unnecessary attributes, for example, $\mathbf{q}_j = [1\ 1\ 0]$, $\mathbf{q}^* = [1\ 0\ 1]$. In this case, ΔN_{11} respondents in G_1 will be wrongly categorized into G_3, and ΔN_{10} respondents from G_2 to G_4, and $\Delta N_{11} > \Delta N_{10}$. Meanwhile, some respondents in G_3 will be wrongly categorized into G_1 (ΔN_{01} respondents), and from G_4 to G_2 (ΔN_{00} respondents), and $\Delta N_{01} < \Delta N_{00}$. Then, $N_{11}^* = N_{11} - \Delta N_{11} + \Delta N_{01}$, $N_{10}^* = N_{10} - \Delta N_{10} + \Delta N_{00}$, $N_{01}^* = N_{01} - \Delta N_{01} + \Delta N_{11}$, $N_{00}^* = N_{00} - \Delta N_{00} + \Delta N_{10}$. Thus, A-3 becomes

$$\Delta = 2\left(\Delta N_{10} + \Delta N_{01} - \Delta N_{00} - \Delta N_{11}\right) \log\left(\frac{s_j^0}{1-s_j^0}\frac{g_j^0}{1-g_j^0}\right), \tag{A-7}$$

and $\Delta \ge 0$. Again, the misspecification in this case is expected to lead to an increase of the R statistic, except

when the $\mathbf{q}^*$ is correctly specified (i.e., $\mathbf{q}_j = \mathbf{q}_j^*$), in which case $\Delta = 0$.

It should be noted that Case 3 can be considered as the general case that covers Case 1 and Case 2. In other words, both Cases 1 and Case 2 are special cases of Case 3. Altogether, any misspecifications in the $\mathbf{q}^*$ will lead to a larger residual statistic. In other words, one can estimate $\mathbf{q}_j$ can be obtained by minimizing $R_j\left(\boldsymbol{\alpha}, \mathbf{q}_j, s_j^0, g_j^0\right)$, with s_j^0, g_j^0being arbitrarily chosen in the range of (0, .5). This indicates that we can estimate the attribute vector of the j^{th} new item without knowing its item parameters.

References

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques (2nd ed.)*. New York, NY: Marcel Dekker. https://doi.org/10.1201/9781482276725

Ban, J. C., Hanson, B. A., Wang, T. Y., & Harris, D. J. (2001). A comparative study of on-line pretest item–calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*(3), 191–212. https://doi.org/10.1111/j.1745-3984.2001.tb01123.x

Ban, J. C., Hanson, B. A., Yi, Q., & Harris, D. J. (2002). Data sparseness and on-line pretest item calibration-scaling methods in CAT. *Journal of Educational Measurement, 39*(3), 207–218. https://doi.org/10.2307/1435078

Chen, P., & Xin, T. (2011). *Item replenishing in cognitive diagnostic computerized adaptive testing*. Paper presented at the annal meeting of the National Council on Measurement in Education, New Orleans, LA.

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. X., Liu, J. C., & Ying, Z. L. (2015). Online item calibration for Q-Matrix in CD-CAT. *Applied Psychological Measurement, 39*(1), 5–15. https://doi.org/10.1177/0146621613513065

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

Dai, B. Y., Zhang, M. Q., & Li, G. M. (2016). Exploration of item selection in dual-purpose cognitive diagnostic computerized adaptive testing: Based on the RRUM. *Applied Psychological Measurement, 40*(8), 625–640. https://doi.org/10.1177/0146621616666008

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

Embretson, S. E. (2001). The second century of ability testing: Some predictions and speculations. Retrieved from http://www.ets.org/Media/Research/pdf/PICANG7.pdf

Fink, A., Born, S., Spoden, C., & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling, 60*(3), 327–346.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* (Doctoral dissertation), University of Illinois. http://hdl.handle.net/2142/87393

Heller, J., & Wickelmaier, F. (2013). Minimum discrepancy estimation in probabilistic knowledge structures. *Electronic Notes in Discrete Mathematics, 42*, 49–56. https://doi.org/10.1016/j.endm.2013.05.145

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

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

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*(3), 167–188. https://doi.org/10.1177/0146621614554650

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

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*(3), 253–275. https://doi.org/10.1111/bmsp.12070

Ma, W., & de la Torre, J. (2019). Digital module 05: Diagnostic measurement — The G-DINA framework. *Educational Measurement: Issues and Practice, 38*(2), 114–115. https://doi.org/10.1111/emip.12262

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

Ren, H., Van der Linden, W. J., & Diao, Q. (2017). Continuous online item calibration: Parameter recovery and item utilization. *Psychometrika, 82*(2), 498–522. https://doi.org/10.1007/s11336-017-9553-1

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*(4), 219–262. https://doi.org/10.1080/15366360802490866

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods and application*. Guilford. https://doi.org/10.1111/j.1751-5823.2011.00134_21.x

Stocking, M. L. (1988). *Scale drift in on-line calibration* (Research Report No. 88-28). Retrieved from Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1988.tb00284.x

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 51*(3), 337–350. http://www.jstor.org/stable/3592656

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer & N. J. Dorans (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 61–100). Lawrence Erlbaum Retrieved from https://link.springer.com/content/pdf/10.1023/A:1016834001219.pdf

Wang, D., Cai, Y., & Tu, D. (2020). Q-matrix estimation methods for cognitive diagnosis models: Based on partial known Q-matrix. *Multivariate Behavioral Research, 1–13*. https://doi.org/10.1080/00273171.2020.1746901

Xu, X. L., Chang, H. H., & Douglas, J. (April, 2003). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the Annual Meeting of American Educational Research Association, Chicago, IL.

Yu, X., & Cheng, Y. (2020). Data-driven Q-matrix validation using a residual-based statistic in cognitive diagnostic assessment. *British Journal of Mathematical and Statistical Psychology., 73*(S1), 145–179. https://doi.org/10.1111/bmsp.12191

Yu, X., Cheng, Y., & Chang, H. (2019). Recent developments in cognitive diagnostic computerized adaptive testing (CD-CAT): A comprehensive review. In von Davier & Lee (Eds.), *Handbook of Diagnostic Classification Models* (pp. 307–331). Springer. https://doi.org/10.1007/978-3-030-05584-4_15

Zheng, C., & Chang, H. (2016). High-efficiency response distribution-based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 40*(6), 608–624. https://doi.org/10.1177/0146621616665196

Zheng, C., & Wang, C. (2017). Application of binary searching for item exposure control in cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 41*(7), 561–576. https://doi.org/10.1177/0146621617707509

Open practices statement The data in the real analysis section of the paper are available through the R package pks (https://cran.r-project.org/web/packages/pks/pks.pdf). The codes to estimate the Q-matrix using our two-step procedure are available at an OSF depository: https://osf.io/d7ehf/?view_only=78a85c10c3be4f2ca27a0e1e7588b939.