# A Novel Efficient Method for Online Calibration in CD-CAT: Entropy-Based Information Gain and EM Perspective

**Authors:** Tan Qingrong, Wang Daxun, Luofen, Cai Yan, Dongbo Tu, Dongbo Tu

**Date:** 2021-07-30T00:00:00+00:00

## Abstract

Item Replenishing plays a crucial role in the maintenance of item banks for Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT), and online calibration represents an important method for item replenishment. Drawing upon the concept of Feature Selection from data mining, this study proposes an efficient online calibration method based on entropy-based information gain (denoted as IGEOCM), which utilizes examinees' responses to both new and existing items to jointly estimate the Q-matrix and item parameters for new items. The study employs Monte Carlo simulation experiments to validate the effectiveness of the newly developed method and simultaneously compares it with existing online calibration methods including SIE (Chen et al., 2015), SIE-R-BIC, and RMSEA-N (谭青蓉，2019). Results indicate that the newly developed IGEOCM demonstrates satisfactory item calibration accuracy and item estimation efficiency across all experimental conditions, and overall outperforms existing methods such as SIE; additionally, IGEOCM requires less time to calibrate new items than SIE and other methods. In summary, this study provides a more efficient and accurate method for item replenishment in CD-CAT item banks.

## Full Text

## A High-Efficiency New Online Calibration Method for CD-CAT: An Information Gain of Entropy and EM Perspective

**TAN Qingrong[1], WANG Daxun[1], LUO Fen[1], CAI Yan[1], TU Dongbo[1]**
[1]School of Psychology, Jiangxi Normal University, Nanchang 330022, China

## Abstract

Item replenishing plays a crucial role in maintaining item banks for cognitive diagnostic computerized adaptive testing (CD-CAT), and online calibration represents an important approach to item replenishment. Drawing inspiration from feature selection techniques in data mining, this study proposes an efficient online calibration method based on information gain of entropy (denoted as IGEOCM) that jointly estimates the Q-matrix and item parameters of new items using examinees'responses to both old and new items. Monte Carlo simulation studies were conducted to evaluate the performance of the newly developed method and compare it with existing online calibration methods including SIE (Chen et al., 2015), SIE-R-BIC, and RMSEA-N (Tan, 2019). Results demonstrate that IGEOCM exhibits strong item calibration accuracy and estimation efficiency across all experimental conditions, outperforming existing methods such as SIE overall. Additionally, IGEOCM requires less time to calibrate new items compared to SIE and other methods. In summary, this research provides a more efficient and accurate method for item replenishment in CD-CAT item banks.

**Keywords:** Cognitive Diagnostic Computerized Adaptive Testing, Item Replenishing, Online Calibration, Q-matrix, Information Gain of Entropy
**Classification Code:** B841

## 1 Introduction

The continuous development of assessment technology and computer technology has led the public to pursue not only test efficiency but also comprehensive test results beyond a single overall score. People desire detailed and thorough test results that enable systematic evaluation of their strengths and weaknesses in the measured content domain, identification of areas needing improvement, and formulation of further learning plans. Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) represents the integration of Cognitive Diagnosis (CD) and Computerized Adaptive Testing (CAT), which can improve testing efficiency and accuracy while providing detailed diagnostic feedback about examinees'strengths and weaknesses in the measured content domain (Wang, 2013; Weiss, 1982). Consequently, targeted instructional remediation can be provided for examinees' weak knowledge points based on their diagnostic results, effectively meeting contemporary demands for efficient and comprehensive testing and demonstrating broad application prospects (Leighton et al., 2004; Liu et al., 2013).

CD-CAT requires a well-constructed item bank as a prerequisite. However, some items in the bank become overexposed or outdated over time, necessitating replacement or replenishment with new items (Chen, 2017). Specifically, experienced domain experts and psychometricians must be invited to develop new items according to diagnostic purposes, after which the parameters of these new items are estimated and placed on the same scale as existing items in the bank.

Online calibration is an effective item replenishment method in traditional CAT that involves having examinees respond to both new and old items during testing and then calibrating the new item parameters based on their responses, with test administrators informing examinees that some items will not count toward their final ability estimates (Chen & Xin, 2011a). Compared to traditional item replenishment methods, online calibration offers several advantages: (1) it eliminates the need for complex post-hoc equating techniques to place new and old item parameters on the same scale (Chen & Wang, 2015); (2) it enables simultaneous estimation of examinee abilities and calibration of new item parameters without requiring external calibration studies, saving substantial human and material resources; and (3) the identical testing format ensures examinees maintain consistent motivation when responding to both new and old items (Chen et al., 2012). To date, researchers have proposed numerous efficient online calibration methods for Unidimensional Computerized Adaptive Testing (UCAT) and Multidimensional Computerized Adaptive Testing (MCAT). In UCAT, Stocking (1988) proposed Method A and Method B, Wainer and Mislevy (1990) recommended a marginal maximum likelihood estimation method with one EM cycle (OEM), and subsequently Ban et al. (2001) proposed a marginal maximum likelihood estimation method with multiple EM cycles (MEM) and the BILOG/Prior method. Additionally, to overcome the theoretical limitation of Method A treating estimated ability values as true ability values, Chen (2016) proposed the FFMLE-Method A and ECSE-Method A methods. In MCAT, Chen et al. (2017) extended Method A, OEM, and MEM, referring to them as M-Method A, M-OEM, and M-MEM, and further recommended M-OEM-BME and M-MEM-BME methods for calibrating item parameters in MCAT (Chen, 2017).

However, research on online calibration methods in CD-CAT remains limited, primarily comprising two categories. The first category includes methods such as CD-Method A, CD-OEM, and CD-MEM proposed by Chen et al. (2012), which are based on Method A, OEM, and MEM. These methods assume the Q-matrix of new items is known and calibrate only the item parameters. In reality, the Q-matrix, as a core component of cognitive diagnosis, is often unknown. In practice, the Q-matrix is typically defined jointly by content domain experts and measurement specialists, requiring substantial human and material resources, and expert-defined Q-matrices are susceptible to subjective factors that can lead to specification errors. Such Q-matrix misspecification ultimately affects the accuracy of item parameter estimation and examinee classification (de la Torre & Chiu, 2016; Rupp & Templin, 2008). Consequently, the second category of online calibration methods has emerged, which simultaneously calibrates both the Q-matrix and item parameters of new items to reduce the resources expended on item calibration and improve calibration efficiency. The Joint Estimation Algorithm (JEA) proposed by Chen and Xin (2011b), the Single-Item Estimation (SIE) method proposed by Chen et al. (2015), and the SIE-R-BIC and RMSEA-N methods proposed by Tan (2019) all belong to this category. The JEA method draws on the joint maximum likelihood estima-

tion (JMLE) approach from item response theory (IRT), treating estimated attribute mastery patterns of examinees in CD-CAT as true values and then using maximum likelihood estimation (MLE) to jointly estimate the Q-matrix and item parameters of new items based on examinees' estimated attribute mastery patterns and their responses to new items. Unlike JEA, SIE uses the posterior distribution of attribute mastery patterns instead of point estimates, calculates each examinee' s posterior predictive distribution, and employs MLE to estimate the Q-matrix of new items. Simultaneously, SIE utilizes the EM algorithm to estimate item parameters. The SIE-R-BIC method builds upon SIE by incorporating model complexity, constructing a BIC criterion to estimate new items' q-vectors through BIC minimization, while RMSEA-N assesses consistency between observed and expected response distributions to calibrate new items' Q-matrices (Tan, 2019). Compared to JEA, SIE, SIE-R-BIC, and RMSEA-N demonstrate improved Q-matrix calibration accuracy, but all suffer from lengthy calibration times and relatively low efficiency. Therefore, developing methods that enhance both calibration accuracy and efficiency in CD-CAT contexts is essential.

Data mining, a hot topic in database and artificial intelligence research, faces the primary challenge of extracting effective information from massive datasets to achieve efficient data utilization (Chandrashekar & Sahin, 2014). Feature selection offers an effective solution by removing redundant or irrelevant features from large datasets to select the most effective feature subset, thereby improving classification accuracy and efficiency (Guyon & Elisseeff, 2003). A critical component of feature selection is the selection criterion, which measures the relationship between features and classification to eliminate irrelevant features. Criteria such as information gain, mutual information, normalized mutual information, and conditional mutual information evaluate classification accuracy to select optimal features (Fleuret, 2004; Lee et al., 2012; Hoque et al., 2014; Pereira et al., 2015). Features that classify examinees more accurately have higher selection probability, while features performing at random classification levels have lower probability.

Inspired by feature selection in data mining, we propose the following logical hypothesis: when calibrating new items in CD-CAT, feature selection methods can be employed to calibrate new items' Q-matrices, and item parameters can be estimated based on these Q-matrices. Treating all possible q-vectors of a new item as candidate features, and with examinee attribute mastery patterns known, each possible q-vector' s classification effectiveness can be evaluated through feature selection criteria. The q-vector yielding optimal criterion values can then be selected as the new item' s q-vector. Based on this hypothesis, this study proposes a novel CD-CAT online calibration method that jointly calibrates Q-matrices and item parameters using feature selection (the detailed process, rationale, and formulas will be presented in Section 3), aiming to provide new perspectives and methods for CD-CAT online calibration and further promote the development and application of cognitive diagnosis, particularly CD-CAT, in practice.

## 2 Existing Online Calibration Methods

Currently, primary online calibration methods in CD-CAT that simultaneously calibrate new items' Q-matrices and item parameters include JEA (Chen & Xin, 2011b), SIE (Chen et al., 2015), SIE-R-BIC, and RMSEA-N (Tan, 2019). The SIE method, proposed under the Deterministic Input, Noisy and Gate (DINA; Junker & Sijtsma, 2001) model, builds upon JEA and considers estimation errors in examinees'attribute mastery patterns, fully utilizing the posterior distribution of attribute mastery patterns when calibrating Q-matrices and item parameters.

The SIE method comprises two components: Q-matrix calibration and item parameter calibration. For Q-matrix calibration, it first calculates the posterior distribution of attribute mastery patterns for examinees who responded to new item j based on their responses to old items. Subsequently, it computes each examinee i' s posterior predictive distribution for a specific response given the posterior distribution of attribute mastery patterns and the probability of correct response to new item j with q-vector :

where K represents the number of attributes measured by the test, denotes the probability that examinee i' s attribute mastery pattern is , calculated based on examinee i' s responses to old items, and represents the probability of correct response to item j for examinees with attribute mastery pattern under the DINA model. Finally, it constructs a likelihood function combining examinees' posterior predictive distributions and their observed responses to new item j, maximizing this likelihood to estimate the new item' s q-vector:

where denotes the set of all possible q-vectors for new item j. Additionally, SIE employs the EM algorithm to estimate item parameters.

The SIE-R-BIC method incorporates model complexity into SIE by constructing a BIC criterion and estimating new items' q-vectors through BIC minimization:

where penalizes model complexity, represents the number of free parameters, and denotes the number of examinees who responded to new item j. Meanwhile, SIE-R-BIC utilizes information from existing items in the bank when calibrating item parameters, using the mean of item parameters from old items sharing the same q-vector as initial values for new items. The RMSEA-N method calibrates item parameters identically to SIE-R-BIC but calibrates Q-matrices by assessing consistency between observed and expected response distributions. Specifically, it selects the q-vector that maximizes consistency between observed and expected response distributions as the estimated q-vector for new item j:

where represents the marginal probability of examinees with the c-th attribute mastery pattern, and and represent the standardized expected and observed correct response probabilities for the c-th attribute mastery pattern, respectively.

### 3.1 Feature Selection Methods and Information Gain of Entropy

In data mining, one purpose of feature selection is to choose features with high discriminative power for data classification. If classification based on a particular feature yields results similar to random classification, that feature contributes little to classification effectiveness (Li, 2012). Information Gain of Entropy-based (IGE) serves as a feature selection criterion; larger IGE values indicate stronger data classification capability (Pereira et al., 2015). The process of selecting optimal features using IGE proceeds as follows: (1) First, identify dataset R and the features for classifying this dataset. (2) Then, calculate the entropy of dataset R:

where is the sample size of dataset R, x represents categories in dataset R, and is the number of samples in dataset R belonging to the x-th category. Entropy measures the uncertainty level of dataset R; larger values indicate greater uncertainty. Uncertainty refers to the consistency level among examinees in dataset R—uncertainty is lowest when all examinees belong to the same category. (3) Next, calculate the conditional entropy of dataset R given feature A:

where H is the number of values for feature A, represents the number of examinees in dataset R belonging to the h-th category, and represents the number of examinees in the h-th subcategory of dataset R belonging to the x-th category. Conditional entropy measures the uncertainty level of dataset R given feature A. Like entropy, larger conditional entropy values indicate greater uncertainty and poorer classification effectiveness based on feature A. (4) Finally, calculate the information gain of entropy:

The information gain of entropy represents the reduction in uncertainty of dataset R given feature A' s information. Larger values indicate better classification effectiveness based on feature A. (5) For all features, repeat steps (3) and (4), compare their information gain values, and select the feature with the maximum value as the optimal feature.

The magnitude of information gain of entropy depends on dataset R' s entropy ( ) and feature A's conditional entropy ( ). As shown in Equation (5), dataset R's entropy calculation is independent of features—in other words, remains constant across all features. Therefore, feature selection based on information gain of entropy essentially relies on conditional entropy; smaller conditional entropy indicates better classification effectiveness and higher likelihood of being the optimal feature.

Estimating new item j' s q-vector can be viewed as a feature selection problem: selecting the best q-vector from all possible q-vectors for new item j. Treating examinees' responses to new item j as dataset R and all possible q-vectors of new item j as features, examinees can be classified based on q-vectors and their estimated attribute mastery patterns. The q-vector that minimizes classification uncertainty for response dataset R represents the optimal feature and can thus

be selected as new item j' s estimated q-vector. Based on this rationale, we propose a new online calibration method—Information Gain of Entropy-based Online Calibration Method (IGEOCM)—which uses information gain of entropy to calibrate new items' Q-matrices and employs the EM algorithm to calibrate item parameters.

## 3.2 Development of the Information Gain of Entropy-Based Online Calibration Method

The DINA model, as one of the most widely applied cognitive diagnostic models, features only two simple and interpretable item parameters per item (slip and guessing parameters) and is frequently used for CD-CAT item bank construction and online calibration (Junker & Sijtsma, 2001; Liu et al., 2013). To facilitate explanation and comparison with existing international methods (the SIE method), the DINA model is used to illustrate the basic rationale and procedure of IGEOCM for calibrating new items.

### 3.2.1 Q-Matrix Calibration in IGEOCM

When the number of attributes measured by new item j (K) is known, the number of possible q-vectors for new item j is , excluding the vector with all zero elements. From a feature selection perspective, estimating new item j' s q-vector involves selecting the most appropriate q-vector from possible candidates.

IGEOCM estimates new item j' s q-vector using information gain of entropy as the feature selection criterion:

where represents the response vector of examinees to new item j (i.e., the response dataset for new item j), is the number of examinees who responded to new item j, x represents examinees' scores on item j (with x = 0 or 1 for dichotomous scoring), is the number of examinees among the respondents who scored x on new item j, and h represents categories for classifying examinees based on q-vectors. Under the DINA model, examinees can be divided into two categories (h = 1 or 0) based on their attribute mastery patterns and item q-vectors: the mastery group and the non-mastery group. The mastery group comprises examinees who have mastered all attributes measured by the item, while the non-mastery group includes examinees lacking mastery of at least one attribute measured by the item. represents the number of examinees among the respondents belonging to the h-th category, and represents the number of examinees among the respondents in the h-th category who scored x on new item j. For all possible q-vectors of new item j, the number of examinees responding to new item j ( ) and each examinee' s score x remain constant. Therefore, calibrating new items' q-vectors based on information gain of entropy essentially involves selecting the q-vector that minimizes conditional entropy when new item j' s q-vector is unknown:

When examinees' attribute mastery patterns are known (i.e., estimated based

on responses to old items in CD-CAT), if new item j' s q-vector is correct and examinees' responses contain no slip or guess errors, all examinees in the mastery group should have observed scores of 1 on new item j, while all examinees in the non-mastery group should have scores of 0. In this case, both groups exhibit high consistency and minimal uncertainty, yielding minimal conditional entropy and thus optimal classification effectiveness for the correct q-vector. If examinees' attribute mastery patterns follow a uniform distribution and responses contain no slip or guess errors, the variation in correct and incorrect q-vectors for new item j is shown in .

The rationale for estimating new item j' s q-vector by minimizing conditional entropy ( ) in IGEOCM is proven as follows:

Under the DINA model, (where represents the number of examinees in the mastery group among the respondents, and represents the number of examinees in the mastery group who scored 0 on new item j), and (where represents the number of examinees in the non-mastery group among the respondents, and represents the number of examinees in the non-mastery group who scored 1 on new item j). Substituting these into the function, taking partial derivatives with respect to and , and setting them equal to zero yields:

Through algebraic operations, we obtain:

The maximum is attained at . When , the function is monotonically increasing.

Additionally, according to Yu and Cheng (2020), when attribute mastery patterns are known and , the following equality holds:

Therefore, when examinees' attribute mastery patterns and their responses to item j are known and , new item j' s q-vector can be estimated by minimizing conditional entropy:

Meanwhile, Example 1 in the appendix further illustrates the rationale for IGEOCM' s q-vector estimation through conditional entropy minimization.

### 3.2.2 Item Parameter Calibration in IGEOCM

IGEOCM employs the EM algorithm to estimate new items' parameters, with each iteration comprising an Expectation Step (E-step) and a Maximization Step (M-step) (Chen et al., 2015). In the E-step, each examinee' s posterior distribution is first calculated based on their response to new item j:

Then, assuming independence among the examinees' responses to new item j, the log-marginal likelihood function is constructed based on their response vector and the posterior distribution of attribute mastery patterns:

The M-step aims to maximize Equation (16) to estimate new item j' s slip parameter and guessing parameter . The EM algorithm iterates between E-step and M-step until a predetermined convergence criterion is met.

The above two components constitute IGEOCM' s calibration of new items' Q-matrices and item parameters. The specific steps for calibrating new items are as follows:

**Step 1: New item q-vector estimation.** For new item j, based on estimated attribute mastery patterns of examinees who responded to new item j and their response data, calculate the conditional entropy for each possible q-vector' s response dataset . Select the q-vector corresponding to the minimum value as new item j' s estimated q-vector.

**Step 2: New item parameter estimation.** Treat the q-vector estimated in Step 1 as new item j's true q-vector. Using the posterior distribution of attribute mastery patterns of examinees who responded to new item j and their responses, employ the EM algorithm to estimate the slip and guessing parameters. New item j' s calibration is then complete.

**Step 3:** For all other new items to be calibrated, repeat Steps 1 and 2 to obtain their estimated Q-matrix and item parameters (slip and guessing parameters) item by item until all new items are calibrated.

IGEOCM is a new online calibration method proposed from a feature selection perspective. Its advantages include requiring only estimated attribute mastery patterns and examinees' responses to new items to estimate Q-matrices, making it a nonparametric approach that is simple, intuitive, and computationally efficient. Furthermore, IGEOCM directly calibrates item parameters using the q-vector estimated via nonparametric methods as the true q-vector, requiring estimation of item parameters for only one determined q-vector regardless of how many possible q-vectors exist for a new item. This substantially reduces calibration time and improves calibration efficiency, differing from SIE methods that require parameter estimation for multiple candidate q-vectors.

## 4 Study 1: Simulation Study Design and Results

### 4.1 Experimental Design

Study 1 aimed to examine IGEOCM' s performance in calibrating new items under different calibration sample sizes (40, 80, 120, 160, 200), attribute mastery pattern distributions (uniform, higher-order, and multivariate normal), and numbers of new items answered by examinees (D = 4, 6, 8), comparing it with SIE, SIE-R-BIC, and RMSEA-N methods. Calibration sample size refers to the number of examinees who responded to new item j, where N is the total number of examinees participating in CD-CAT, D is the number of new items each examinee answers, and m is the number of new items to be calibrated (Chen et al., 2015). SIE, SIE-R-BIC, and RMSEA-N were selected as comparison methods primarily because their new item calibration accuracy is slightly superior to JEA, making them representative. Study 1 employed a four-factor experimental design with simulation conditions, each repeated 500 times to reduce random error.

**4.1.1 Simulation of Examinees and Item Bank** Calibration sample size had five levels: . Examinees'attribute mastery patterns were generated from uniform, higher-order, and multivariate normal distributions. Under the uniform distribution, attribute mastery patterns were generated with equal probability from all possible patterns. Under the higher-order distribution, whether examinee i masters attribute k depends on their general latent ability , with the probability of mastering attribute k for examinee i with ability being:

where are structural parameters with . The study set and for all attributes k, with examinee i' s ability values drawn from N(0,1) (de la Torre & Chiu, 2016). Under the multivariate normal distribution, the correlation between attributes was set to 0.5 (J. Chen, 2017).

Item bank simulation included item parameters (slip parameter s and guessing parameter g) and item Q-matrices. The bank contained 300 items, each measuring at most 3 attributes, with 100 items measuring 1 attribute, 100 measuring 2 attributes, and 100 measuring 3 attributes. With total attributes measured, there were 63 possible item q-vectors: 6 measuring 1 attribute, 15 measuring 2 attributes, and 20 measuring 3 attributes. The 6 q-vectors measuring 1 attribute were repeated 16 times with 4 additional q-vectors randomly selected, the 15 q-vectors measuring 2 attributes were repeated 6 times with 10 additional q-vectors randomly selected, and the 20 q-vectors measuring 3 attributes were repeated 5 times to form a temporary test Q-matrix. All rows in the temporary Q-matrix were then randomly sorted to obtain the final Q-matrix. Each item' s slip parameter s and guessing parameter g were randomly drawn from U(0.05, 0.25).

**4.1.2 Simulation of New Items** New item simulation included slip parameters s, guessing parameters g, and Q-matrices. The study set the number of new items to be calibrated at , making the new items' Q-matrix a matrix. Simulation of new items' Q-matrices and their slip and guessing parameters followed the same procedure as the item bank simulation.

**4.1.3 CD-CAT Simulation and New Item Calibration** The study used a fixed-length termination rule, with each examinee answering 20 old items and D new items (D = 4, 6, 8). The CD-CAT simulation proceeded as follows: At the test' s beginning, (1) one item was randomly selected from the bank as the initial item; (2) the current examinee' s response was simulated, and their attribute mastery pattern was estimated using MLE based on responses to administered items; (3) the Posterior-Weighted Kullback-Leibler (PWKL; Cheng, 2009) item selection strategy selected the most suitable item from the remaining bank based on the current estimated attribute mastery pattern. Steps (2) and (3) were repeated until the test reached the predetermined length.

During CD-CAT simulation, D new items were randomly selected from the 24 new items to be calibrated and placed at random positions in each examinee' s test. After CD-CAT completion, IGEOCM, SIE, SIE-R-BIC, and RMSEA-N

methods were used to calibrate new items'Q-matrices and item parameters based on estimated attribute mastery patterns, posterior distributions, and responses to new items.

**4.1.4 Evaluation Criteria Attribute Vector Correct Estimation Rate (AVCER):** AVCER evaluates the correct estimation rate of new items' Q-matrices:

where r represents the r-th replication of 500 simulation replicates, is the estimated q-vector for new item j in the r-th replication, and is the true q-vector for new item j. is an indicator function assessing whether equals in the r-th replication. Higher AVCER values indicate better Q-matrix estimation accuracy.

**Root Mean Squared Error (RMSE):** RMSE evaluates the accuracy of new item parameter estimation:

where and are estimated slip and guessing parameters for new item j in the r-th replication, and and are the true slip and guessing parameters. Smaller RMSE values indicate higher item parameter estimation accuracy.

**Calibration Efficiency: Average Running Time (ART):** ART evaluates the calibration efficiency of online calibration methods:

where represents the time required for each online calibration method to calibrate new items in the r-th replication. Smaller ART values indicate higher calibration efficiency.

### 4.2 Experimental Results

[Figure 1: see original paper], , and [Figure 2: see original paper] present the item calibration accuracy and efficiency results for SIE, SIE-R-BIC, RMSEA-N, and IGEOCM. Following Chen et al. (2015), a difference in calibration accuracy $\geq 1\%$ between methods indicates superiority. Overall, IGEOCM demonstrates strong item calibration accuracy and estimation efficiency, outperforming SIE, SIE-R-BIC, and RMSEA-N. [Figure 1: see original paper] shows IGEOCM' s Q-matrix estimation accuracy exceeds the other three methods, with more pronounced differences under higher-order and normal distributions. For instance, under uniform distribution, the maximum AVCER difference between SIE and IGEOCM is 2.3%, while under higher-order and normal distributions, the differences reach 6.8% and 9.1%, respectively.

SIE and SIE-R-BIC show similar Q-matrix calibration accuracy across conditions, while RMSEA-N' s accuracy is lower than SIE and SIE-R-BIC under higher-order and normal distributions. Regarding distribution effects, Q-matrix estimation accuracy for all methods is highest under uniform distribution, followed by higher-order distribution, and lowest under normal distribution. For example, IGEOCM' s Q-matrix estimation accuracy ranges from 80.9%–99.8% under uniform, 67.0%–97.3% under higher-order, and 46.0%–76.7% under normal distributions; SIE' s ranges are 79.0%–99.8%, 60.7%–96.9%, and 38.4%–

68.3%, respectively. Calibration sample size substantially affects Q-matrix estimation accuracy, with larger samples yielding higher accuracy. When , average AVCER values are 59.6%, 60.0%, 45.6%, and 65.1% for SIE, SIE-R-BIC, RMSEA-N, and IGEOCM, respectively; when , these increase to 88.1%, 88.2%, 77.2%, and 91.2%. Thus, increasing calibration sample size improves Q-matrix estimation accuracy for all methods.

The number of new items answered by examinees has negligible impact on Q-matrix estimation accuracy for SIE, SIE-R-BIC, RMSEA-N, and IGEOCM.

presents item parameter calibration results. SIE and IGEOCM show similar performance, with maximum RMSE differences $\leq 0.2\%$ and equal RMSE values in most conditions. SIE-R-BIC' s RMSE is slightly lower than SIE and IGEOCM with small calibration samples (e.g., ) and slightly higher with large samples (e.g., ). RMSEA-N' s RMSE exceeds the other three methods in most conditions. Regarding distribution effects, SIE, SIE-R-BIC, and IGEOCM achieve best parameter calibration accuracy under higher-order distribution, while RMSEA-N performs best under uniform distribution. For example, IGEOCM' s average RMSE values are 0.056, 0.066, and 0.071 under higher-order, uniform, and normal distributions, respectively; RMSEA-N' s are 0.093, 0.088, and 0.142.

Item parameter calibration accuracy improves with increasing calibration sample size for all methods. When , average RMSE values for SIE and IGEOCM are both 0.11; when , these decrease to 0.04. Consistent with Q-matrix calibration, the number of new items answered has negligible impact on parameter calibration accuracy.

[Figure 2: see original paper] shows average running times for estimating 24 new items using SIE, SIE-R-BIC, RMSEA-N, and IGEOCM. All methods were run using R 4.0 on identical computer configurations (Intel Core i5-8400 2.81GHz, 20GB RAM), ensuring comparability. Results show SIE, SIE-R-BIC, and RMSEA-N are substantially less efficient than IGEOCM, with average ART values approximately 49 times longer across all conditions. Attribute mastery pattern distribution and number of new items answered have minimal impact on efficiency. All methods' running times increase with calibration sample size. When , average ART values are 106.22, 93.38, 61.39, and 1.74 seconds for SIE, SIE-R-BIC, RMSEA-N, and IGEOCM, respectively; when , these increase to 414.71, 322.40, 286.06, and 6.91 seconds.

## 5 Study 2: Impact of Item Selection Strategies on IGEOCM and Existing Methods

IGEOCM, SIE, SIE-R-BIC, and RMSEA-N calibrate new items based on estimated attribute mastery patterns, posterior distributions, and responses to new items. The estimation accuracy of attribute mastery patterns and posterior distributions affects calibration accuracy (Chen et al., 2015). In CD-CAT, item selection strategies significantly influence attribute mastery pattern estimation

accuracy. Therefore, Study 2 builds upon Study 1 to further examine how item selection strategies affect the performance of online calibration methods.

### 5.1 Experimental Design

Study 2' s design and simulation procedures closely followed Study 1 but added three item selection strategies: MPWKL (modified PWKL), GDI (generalized deterministic inputs, noisy "and" gate model discrimination index), and SHE (Shannon entropy) (Cheng, 2009; Kaplan et al., 2015) to compare the feasibility and accuracy of IGEOCM and SIE under different strategies. Since SIE and SIE-R-BIC show slightly better item calibration accuracy than RMSEA-N, and SIE' s parameter calibration accuracy is slightly better than SIE-R-BIC and RMSEA-N in most conditions, while all three show similar efficiency (ART ratios $\leq 1$), Study 2 selected SIE as the comparison method for IGEOCM. Additionally, based on Study 1' s results showing minimal impact of new items answered, Study 2 fixed this at D = 6. Considering that SIE and IGEOCM' s running times increase with calibration sample size, Study 2 fixed calibration sample size at to reduce experimental duration. Other conditions followed Study 1.

### 5.2 Experimental Results

presents item calibration accuracy and efficiency results for SIE and IGEOCM under different item selection strategies and attribute mastery pattern distributions. Consistent with Study 1, IGEOCM demonstrates higher calibration accuracy and efficiency than SIE across all strategies. Moreover, both methods achieve highest Q-matrix estimation accuracy under uniform distribution, followed by higher-order, and lowest under normal distribution.

CD-CAT item selection strategies moderately affect Q-matrix calibration accuracy. For example, under higher-order distribution, SIE shows higher accuracy with MPWKL (AVCER = 61.7%) than with PWKL (AVCER = 60.7%). Under normal distribution, IGEOCM shows higher accuracy with GDI (AVCER = 46.7%) than with PWKL (AVCER = 45.4%). Item selection strategies have negligible impact on item parameter calibration and efficiency, with RMSE differences $\leq 0.2\%$ and similar average running times across strategies.

## 6 Summary and Discussion

Few online calibration methods in CD-CAT simultaneously calibrate new items' Q-matrices and item parameters, and existing parametric methods suffer from long calibration times and low efficiency. This study proposes IGEOCM, inspired by feature selection in data mining, to provide a more efficient and accurate method for item replenishment in CD-CAT item banks. Unlike existing CD-CAT online calibration methods, IGEOCM uses a nonparametric approach to calibrate new items' Q-matrices, effectively avoiding impacts from item parameter estimation bias, improving calibration accuracy, and enhancing efficiency. Monte Carlo simulation studies validated IGEOCM' s feasibility and

accuracy, comparing it with SIE, SIE-R-BIC, and RMSEA-N. Results indicate: (1) IGEOCM demonstrates strong calibration accuracy and efficiency across all conditions, outperforming SIE, SIE-R-BIC, and RMSEA-N overall. Methods like SIE estimate Q-matrices based on estimated item parameters, where parameter estimation errors affect Q-matrix accuracy and reduce overall calibration precision. IGEOCM directly calibrates Q-matrices using examinees' attribute mastery patterns and responses, independent of item parameter estimation, resulting in fewer influencing factors and higher accuracy. Although SIE and IGEOCM share the same item parameter estimation method, SIE uses parametric Q-matrix calibration while IGEOCM uses nonparametric methods. Nonparametric methods are computationally simpler and faster (Chiu et al., 2018), giving IGEOCM better calibration efficiency. (2) Calibration accuracy for all four methods improves with increasing sample size, while running times increase accordingly. (3) All methods perform better under uniform and higher-order distributions than under normal distribution. (4) The number of new items answered has minimal impact on calibration accuracy and efficiency. (5) CD-CAT item selection strategies affect Q-matrix calibration accuracy for SIE and IGEOCM. Under higher-order and normal distributions, both methods show slightly higher Q-matrix accuracy with MPWKL and GDI strategies than with PWKL. Additionally, Study 2 examined impacts of different and simulation methods under higher-order distribution (drawing from standard normal and from log-normal, versus setting with for all attributes). With D fixed at 6 and other conditions identical to Study 1, IGEOCM still outperformed SIE, further demonstrating its feasibility and advantages (see Appendix Table 1).

Nevertheless, this study has several limitations requiring future improvement. First, IGEOCM' s performance was validated only under the DINA model. Its performance under more complex cognitive diagnostic models such as RRUM (Hartz, 2002) and G-DINA (de la Torre, 2011) remains to be explored. Unlike DINA, which classifies examinees into only mastery and non-mastery groups, more complex models can divide examinees into multiple categories based on attribute mastery patterns and item q-vectors. The information gain of entropy criterion increases with the number of examinee categories, warranting investigation of its effectiveness for q-vector calibration in complex models. Future research should address category count effects, such as penalizing the number of categories to reduce its impact on IGEOCM.

Second, existing CD-CAT online calibration methods are based on dichotomous scoring models. However, psychological and educational assessments contain substantial polytomous data that provide more comprehensive diagnostic information. How to extend the proposed method to polytomous models like the sequential G-DINA model (Ma & de la Torre, 2016) and validate its performance requires further research.

Third, this study randomly selected new items for each examinee, potentially mismatching items with optimal examinees. Future research should consider adaptive methods for selecting optimal examinees for each item, such as optimal

design criteria (He et al., 2020), and examine how different new item selection approaches (random vs. adaptive) affect online calibration methods.

Finally, this study assumed independence among measured attributes. In practice, diagnostic assessments often involve various hierarchical relationships among attributes (e.g., unstructured, linear, branching, convergent; Leighton et al., 2004). Future research should explore how different attribute hierarchies affect online calibration methods. Additionally, while simulation studies provide guidance for practical application, they operate under ideal conditions that ignore many real-world factors. Future studies must evaluate the performance of online calibration methods in authentic settings. In conclusion, online calibration methods that simultaneously calibrate Q-matrices and item parameters in CD-CAT require continued research.

## References

Ban, J. C., Hanson, B. A., Wang, T., Yi, Q., & 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.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16–28.

Chen, J. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement*, 41(4), 277–293.

Chen, P. (2016). Two new online calibration methods for computerized adaptive testing. *Acta Psychologica Sinica*, 48(9), 1184–1198.

Chen, P. (2017). A Comparative Study of Online Item Calibration Methods in Multidimensional Computerized Adaptive Testing. *Journal of Educational and Behavioral Statistics*, 42(5), 559–590.

Chen, P., & Wang, C. (2015). A New Online Calibration Method for Multidimensional Computerized Adaptive Testing. *Psychometrika*, 81(3), 674–701.

Chen, P., Wang, C., Xin, T., & Chang, H. H. (2017). Developing new online calibration methods for multidimensional computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 70(1), 81–117.

Chen, P., & Xin, T. (2011a). Developing on-line calibration methods for cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica*, 43(6), 710–724.

Chen, P., & Xin, T. (2011b). Item replenishing in cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica*, 43(7), 836–850.

Chen, P., Xin, T., Wang, C., & Chang, H. (2012). Online Calibration Methods for the DINA Model with Independent Attributes in CD-CAT. *Psychometrika*, 77(2), 201–222.

Chen, Y., Liu, J., & Ying, Z. (2015). Online item calibration for Q-matrix in CD-CAT. *Applied Psychological Measurement*, 39(1), 5–15.

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

Chiu, C. Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika*, 83(2), 355–375.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2), 179–199.

de la Torre, J., & Chiu, C. Y. (2016). General method of empirical Q-matrix validation. *Psychometrika*, 81(2), 253–273.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5(11), 1531–1555.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(3), 1157–1182.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign.

He, Y., Chen, P., & Li, Y. (2020). New Efficient and Practicable Adaptive Designs for Calibrating Items Online. *Applied Psychological Measurement*, 44(1), 3–16.

Hoque, N., Bhattacharyya, D. K., & Kalita, J. K. (2014). MIFS-ND: A mutual information-based feature selection method. *Expert Systems with Applications*, 41(14), 6371–6385.

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.

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.

Lee, S., Park, Y. T., & d' Auriol, B. J. (2012). A novel feature selection method based on normalized mutual information. *Applied Intelligence*, 37(1), 100–120.

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.

Li, H. (2012). *Statistical Learning Method*. Beijing: Tsinghua University Press.

Liu, H., You, X., Wang, W., Ding, S., & Chang, H. (2013). The Development of Computerized Adaptive Testing with Cognitive Diagnosis for an English

Achievement Test in China. *Journal of Classification*, 30(2), 152–172.

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.

Pereira, R. B., Plastino, A., Zadrozny, B., & Merschmann, L. H. (2015). Information gain feature selection for multi-label classification. *Journal of Information and Data Management*, 6(1), 48–58.

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

Stocking, M. L. (1988). Scale Drift in On-Line Calibration. *ETS Research Report*, 1988(1), 1–122.

Tan, Q. (2019). *The Development of Generalized Online Calibration Methods in CD-CAT* (Unpublished master' s thesis). Jiangxi Normal University, Nanchang.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized Adaptive Testing: A Primer* (Chap. 4, pp. 65–102). Hillsdale, NJ: Erlbaum.

Wang, C. (2013). Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing with Short Test Length. *Educational and Psychological Measurement*, 73(6), 1017–1035.

Weiss, D. J. (1982). Improving Measurement Quality and Efficiency with Adaptive Testing. *Applied Psychological Measurement*, 6(4), 473–492.

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(Suppl), 145–179.

## Appendix

**Example 1:** Assume examinees' attribute mastery patterns are known and uniformly distributed, with no slip or guess errors in responses to new item j. Notably, since the number of examinees responding to new item j ( ) and each examinee' s response remain fixed across all possible q-vectors, the entropy value of response dataset in the information gain formula is equal across all possible q-vectors, with magnitude entirely dependent on conditional entropy . Therefore, this example demonstrates how conditional entropy changes across different q-vectors to illustrate changes in information gain .

Let the number of measured attributes K = 3, yielding possible attribute mastery patterns, each with expected frequency . If new item j' s correct q-vector is , then under the DINA model, examinees with patterns [1 0 0], [1 1 0], [1 0 1], and [1 1 1] are classified into the mastery group ( ), while those with patterns [0 0 0], [0 1 0], [0 0 1], and [0 1 1] are classified into the non-mastery group ( ).

With uniform distribution, both groups contain examinees. If and represent the numbers of examinees in the mastery group who answered incorrectly and correctly, respectively, and and represent these numbers for the non-mastery group, and labeling the numbers of examinees with each attribute mastery pattern in the mastery group who answered incorrectly as and those in the non-mastery group who answered correctly as , then under the assumption of no slip or guess errors, the expected number of incorrect responses in the mastery group is 0 ( ) and the expected number of correct responses in the non-mastery group is 0 ( ). Moreover, when no slip or guess errors exist, (Chiu et al., 2018). Thus, can be calculated as:

If new item j' s q-vector is incorrectly specified as , then examinees with patterns [0 1 1] and [1 1 1] are classified into the mastery group, while those with patterns [0 0 0], [1 0 0], [0 1 0], [0 0 1], [1 1 0], and [1 0 1] are classified into the non-mastery group ( ). This misclassification places examinees who should answer correctly (patterns [1 0 0], [1 1 0], [1 0 1]) into the non-mastery group, with expected correct responses of , and incorrectly places examinees who should answer incorrectly (pattern [0 1 1]) into the mastery group, with expected incorrect responses of . The conditional entropy becomes:

This example demonstrates that when new item j' s q-vector is correct, is minimized at 0, maximizing information gain . Therefore, when the true q-vector is unknown, the q-vector maximizing information gain can be selected as new item j' s estimated q-vector.

**Appendix Table 1.** Item calibration accuracy of SIE and IGEOCM under different and generation methods

*Note: Condition 1: and drawn from normal and log-normal distributions, respectively; Condition 2: with for all attributes k.*

*Note: Figure translations are in progress. See original paper for figures.*