Development of an Online Calibration Method Based on SCAD Penalty and EM Perspective in CD-CAT: The G-DINA Model

Authors: Tan Qingrong, Cai Yan, Daxun Wang, Luo Fen, Tu Dongbo, Cai Yan, Wang Daxun, Tu Dongbo

Date: 2023-11-22T00:00:00+00:00

Abstract

The G-DINA (generalized deterministic input, noisy and gate) model features minimal constraints and broad applicability, satisfying the requirements of large-scale psychological and educational assessment data. This study proposes SCADOCM, a novel online calibration method for cognitive diagnostic computerized adaptive testing (CD-CAT) that simultaneously calibrates the Q-matrix and item parameters of new items for G-DINA and similar models, aiming to facilitate the practical dissemination and application of CD-CAT. Employing both simulated and real item banks, the results indicate that SCADOCM achieves relatively ideal calibration accuracy and efficiency across all experimental conditions compared to the traditional SIE method, demonstrating promising application prospects. The SIE method is unsuitable for saturated G-DINA and other models, yielding low Q-matrix calibration accuracy under all experimental conditions.

Full Text

Development of an Online Calibration Method Based on SCAD Penalty and EM Perspective in CD-CAT: A Study Based on the G-DINA Model

TAN Qingrong[1],[2], **CAI Yan**[2], **WANG Daxun**[2], **LUO Fen**[3], **TU Dongbo**[2]
([1] Department of Basic Psychology, College of Psychology, Army Medical University, Chongqing 400000, China)
([2] School of Psychology, Jiangxi Normal University, Nanchang 330022, China)
([3] College of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China)

Abstract

The G-DINA (generalized deterministic input, noisy and gate) model imposes few restrictions and has broad applicability, satisfying the requirements of numerous psychological and educational assessment datasets. This study proposes a novel online calibration method, SCADOCM, for simultaneously calibrating the Q-matrix and item parameters of new items in cognitive diagnostic computerized adaptive testing (CD-CAT), applicable to models such as G-DINA, with the aim of promoting the practical implementation and application of CD-CAT. Using both simulated and real item banks, the results demonstrate that SCADOCM achieves satisfactory calibration accuracy and efficiency across all experimental conditions and outperforms the traditional SIE method. The SIE method is not suitable for saturated models such as G-DINA, exhibiting low Q-matrix calibration accuracy under all experimental conditions.

Keywords: Cognitive Diagnostic Computerized Adaptive Testing, Online Calibration, Q-matrix, G-DINA Model, SCAD Penalty
Classification Code: B841

1 Introduction

How can examinees be provided with detailed and valuable diagnostic information about their mastery of assessed content in an efficient and accurate manner to meet their testing needs? This question has garnered considerable attention from researchers and practitioners in psychological and educational measurement in recent years. In psychological assessment, if a test can quickly, accurately, and efficiently provide clinical psychologists—especially novice clinicians—with specific symptom manifestations of a client's psychological problems, helping them better understand the complex relationships between psychological issues and specific symptoms, clinicians can promptly develop effective prevention and intervention strategies to advance the therapeutic process (e.g., de la Torre et al., 2018; Tan et al., 2023). In educational assessment, if a test can rapidly, accurately, and efficiently provide teachers with information about which knowledge points students have mastered and which they have not, teachers can focus classroom instruction on areas where students need improvement, and students can engage in targeted learning to address their weaknesses, thereby reducing student burden, improving instruction, and enhancing teaching effectiveness (e.g., Tang & Zhan, 2021).

Cognitive diagnostic computerized adaptive testing (CD-CAT) emerged against this backdrop, combining the advantages of two recently developed measurement technologies: cognitive diagnosis (CD) and computerized adaptive testing (CAT). CD-CAT represents an ideal approach for achieving these measurement objectives (Cheng, 2009; Lin & Chang, 2019; Xu et al., 2016). The rapid development of cognitive diagnosis is largely driven by the practical demand for formative assessment. Unlike summative assessment, which only provides total test scores, cognitive diagnosis provides each examinee with an attribute mastery

pattern that details their mastery of the assessed concepts or content, offering important reference points for further remediation and intervention after testing (de la Torre, 2011; Junker & Sijtsma, 2001). CAT is favored by researchers and practitioners for its tailored and efficient characteristics. CAT customizes a test for each examinee based on their latent trait level, ensuring that most items administered match their ability level, thus providing more effective and precise latent trait estimates. CD-CAT possesses both the features of CAT and the functions of cognitive diagnosis. Through personalized, "tailor-made" testing, CD-CAT quickly and accurately identifies examinees' strengths and weaknesses in the assessed content, providing timely and detailed diagnostic feedback. This approach improves the accuracy of test results while significantly reducing the response burden on test-takers (Chen et al., 2012; Chen et al., 2015; Lin & Chang, 2019; Liu et al., 2013). This aligns with the spirit and requirements of policies such as "double reduction" and meets the current practical needs of national and social development, promoting precise, adaptive, and personalized psychological and educational assessment, as well as the digital transformation of examinations.

The effectiveness of CD-CAT depends on a high-quality item bank. However, after continuous use over time, some items in the bank become outdated or lose their functionality and must be promptly replaced with new items to maintain test and bank quality (Chen et al., 2012; Chen et al., 2015; Kang et al., 2020). Specifically, experienced domain experts and psychometricians must be invited to develop new items (i.e., items to be added to the bank but with uncalibrated parameters), after which the parameters of these new items are estimated and placed on the same scale as existing bank items. Online calibration is an effective item replenishment method in CAT, referring to the process of having examinees respond to both new and old items (items with calibrated parameters already in the bank) during testing and using their responses to calibrate the new item parameters (Chen & Xin, 2011a). In addition to saving resource investment and ensuring equivalent motivation for responding to new and old items due to identical test administration modes, another important advantage of online calibration is that it eliminates the need for complex equating techniques to address challenging issues such as test equating encountered in large-scale bank construction (Chen & Wang, 2015; Chen et al., 2012). To date, researchers have proposed various efficient online calibration methods for unidimensional CAT (UCAT) and multidimensional CAT (MCAT), including Method A (Stocking, 1988), marginal maximum likelihood estimation with one EM cycle (OEM; Wainer & Mislevy, 1990), marginal maximum likelihood estimation with multiple EM cycles (MEM; Ban et al., 2001), FFMLE-Method A (Chen, 2016), M-Method A (Chen et al., 2017), and M-MEM-BME (Chen, 2017).

While online calibration can be used to calibrate new item parameters in CD-CAT, an important question arises: Is equating necessary in cognitive diagnostic testing, and is online calibration needed to calibrate new items? de la Torre and Lee (2010) noted that when the model fits the data perfectly, the deterministic

input, noisy and gate (DINA; Junker & Sijtsma, 2001) model's item parameters exhibit invariance. Bradshaw and Madison (2015) and Madison and Bradshaw (2018) also indicated that the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009) and the Transition Diagnostic Classification Model (TDCM) developed based on LCDM show parameter invariance when the model fits the data. Under these conditions, equating is not required to ensure that examinee parameter estimates are on the same scale. However, their research also noted that parameter invariance is difficult to observe when the model does not fit the data perfectly, and even when model-data fit is adequate, parameter invariance diminishes as calibration sample size decreases (Bradshaw & Madison, 2015; de la Torre & Lee, 2010; Madison & Bradshaw, 2018). This suggests that parameter invariance requires certain conditions: perfect model-data fit and sufficiently large calibration samples (e.g., no fewer than 1,000). When these conditions are met, equating may be unnecessary. However, in practical testing situations, perfect model-data fit is not always achieved, and obtaining sufficiently large calibration samples within a single test administration is often difficult. Both factors can lead to biased item parameter estimates, affecting classification accuracy and Q-matrix calibration correctness. Therefore, online calibration is necessary for CD-CAT bank construction, as it helps reduce the impact of item parameter estimation bias and improves the quality of CD-CAT banks and tests.

Currently, research on online calibration methods in CD-CAT remains limited. Moreover, unlike UCAT and MCAT, calibrating new items in CD-CAT requires consideration of both item parameter calibration and Q-matrix calibration. As a core component of cognitive diagnosis, the Q-matrix is often unknown in practice. In real testing situations, the Q-matrix is typically defined jointly by domain experts and psychometricians, requiring substantial human and material resources. Additionally, expert-defined Q-matrices are susceptible to subjective influences, and misspecification can ultimately affect item parameter estimation accuracy and examinee classification accuracy (de la Torre & Chiu, 2016; Rupp & Templin, 2008). Therefore, Q-matrix calibration for new items is an aspect that cannot be ignored when calibrating new items in CD-CAT.

To date, some studies have explored simultaneous calibration of Q-matrices and item parameters for new items in CD-CAT. For example, the joint estimation algorithm (JEA) proposed by Chen and Xin (2011b), the single-item estimation (SIE) method by Chen et al. (2015), the Information Gain of Entropy-based Online Calibration Method (IGEOCM) by Tan et al. (2021), and the Gini-based method by Tan et al. (2022) are all online calibration methods that simultaneously calibrate Q-matrices and item parameters for new items. Existing research indicates that JEA, SIE, IGEOCM, and the Gini-based method achieve satisfactory item calibration accuracy under the DINA model. However, their performance under other models, particularly saturated cognitive diagnosis models with broader applicability and fewer constraints (such as the generalized DINA model, or G-DINA; de la Torre, 2011), remains to be further investigated.

Compared with the DINA model, models such as G-DINA have broader applicability due to fewer restrictions and can meet the requirements of most test data in psychological and educational assessment (de la Torre, 2011; de la Torre et al., 2018; Tu et al., 2017; Xi et al., 2020), with increasing practical application. For instance, in psychological clinical diagnostic assessment, clinical diagnosis can be achieved if examinees meet some of the symptoms in the diagnostic criteria for a psychological disorder. Taking Internet addiction as an example, the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) defines nine symptom criteria for Internet addiction, and examinees meeting five or more symptoms can be diagnosed as having Internet addiction. The DINA model is clearly unsuitable for such tests, as it assumes that examinee responses to an item are only influenced by the interaction of all attributes measured by the item, not by main effects or other types of interactions. Forcing this model to analyze the entire test may lead to model-data misfit, subsequently affecting the credibility and accuracy of diagnostic results (Hou, 2013). In contrast, the G-DINA model does not have these strict assumptions, positing that examinee responses can be jointly influenced by main effects of each attribute measured by the item and various types of interaction effects. If the coefficient estimates for main effects (or interaction effects) are zero or near zero, the main effects (or interaction effects) are not significant, meaning they do not exist. However, if coefficients are significantly non-zero, this indicates the presence of main effects (or interaction effects). Therefore, the G-DINA model is more flexible and better suited for such tests.

However, to date, no published journal articles have investigated simultaneous online calibration methods for Q-matrices and item parameters applicable to models with few constraints such as G-DINA. This has somewhat limited the practical application scope of CD-CAT and hindered its further promotion in real testing situations.

In light of this, this study introduces the idea of feature selection using the SCAD (smoothly clipped absolute deviation penalty; Fan & Li, 2001) method from data mining to propose a simultaneous online calibration method for Q-matrices and item parameters applicable to models such as G-DINA. The aim is to provide efficient and accurate methodological support for the further promotion and application of CD-CAT in practice.

2.1 The G-DINA Model

Among existing cognitive diagnosis models, the G-DINA model, developed as an extension of the DINA model, imposes few restrictions and has broader applicability, meeting the requirements of numerous psychological and educational assessment datasets and receiving increasing attention from researchers. More and more researchers are developing cognitive diagnostic tests based on the G-DINA model (e.g., de la Torre et al., 2018; Tu et al., 2017; Xi et al., 2020). Therefore, this study introduces the new online calibration method and validates it within the G-DINA model framework, though the method can also be

applied to other cognitive diagnosis models.

Let K be the number of attributes measured by the test, and $\mathbf{q}_j$ be the q-vector of item j, representing the jth row of the test Q-matrix. If mastering the kth attribute is required for correctly answering item j, then $q_{jk} = 1$; otherwise, $q_{jk} = 0$. Let Y_{ij} denote examinee i's response to item j, and α_c denote the cth attribute mastery pattern, where α_{ck} indicates whether examinee i with the cth attribute mastery pattern has mastered the kth attribute. If attribute k is mastered, $\alpha_{ck} = 1$; otherwise, $\alpha_{ck} = 0$. The G-DINA model posits that examinees with different attribute mastery patterns have different probabilities of correctly answering an item, classifying examinees into $2^{K_j^*}$ categories, where K_j^* is the number of attributes measured by item j.

Depending on the link function used, the G-DINA model has different mathematical expressions. The most commonly used link functions are the log link function, logit link function, and identity link function. The G-DINA model under the identity link function represents a more general form of the G-DINA model (de la Torre, 2011), with the mathematical expression:

$$P(Y_{ij} = 1|\alpha_{ci}) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{ck} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{ck}\alpha_{ck'} + \cdots + \delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ck}, \quad (1)$$

where α_{ci}^* represents the reduced attribute mastery pattern based on the attributes measured by item j, with $\alpha_{ci}^* = (\alpha_{c1}^*, \alpha_{c2}^*, \cdots, \alpha_{cK_j^*}^*)$. For example, if the test measures three attributes and item j measures the first two attributes, then $\alpha_{ci}^* = (\alpha_{c1}, \alpha_{c2})$. The parameter δ_{j0} represents the intercept parameter for item j, also known as the baseline probability—the probability of correctly answering the item without mastering any of its measured attributes, which is non-negative. The parameter δ_{jk} represents the main effect of attribute k for item j, indicating the increase in the probability of correctly answering the item due to mastering attribute k, generally taking non-negative values. Larger values indicate greater contribution of mastering that attribute to correctly answering the item. The parameter $\delta_{jkk'}$ represents the interaction effect between attributes k and k' for item j, and $\delta_{j12\cdots K_j^*}$ represents the interaction effect of all attributes. This paper uses δ_j to denote the item parameter vector for item j.

2.2 The SIE Method

Existing simultaneous calibration methods for new item Q-matrices and item parameters in CD-CAT mainly include JEA (Chen & Xin, 2011b), SIE (Chen et al., 2015), IGEOCM (Tan et al., 2021), and the Gini-based method (Tan et al., 2022). Among these, the JEA method achieves high item calibration accuracy when item quality is high and sample size is large, but its accuracy needs further improvement when item quality is low. However, real test banks

may contain both high-quality and low-quality items. For example, in the large-scale English Level 2 test bank developed by Liu et al. (2013), the slipping parameters (the probability of incorrectly answering an item despite mastering all measured attributes) ranged from 0.001 to 0.5. When new item quality is low, using the JEA method may result in low calibration accuracy, affecting the quality of the entire item bank and test. Additionally, while IGEOCM and the Gini-based method can theoretically be applied to cognitive diagnosis models beyond DINA, these methods are affected by the number of examinee classes. The DINA model classifies examinees into two classes for each item, whereas the G-DINA model classifies examinees into $2^{K_j^*}$ classes (where K_j^* is the number of attributes measured by item j). Their performance under G-DINA and similar models may not be ideal. For example, under G-DINA and similar models, the number of examinee classes increases with the number of attributes measured by an item, while the information gain index of entropy increases with the number of classes (Li, 2012). Therefore, using the IGEOCM method to calibrate q-vectors for new items under G-DINA and similar models may result in over-specification of attributes. Based on this analysis, this paper only details the SIE method and compares it with the new method.

The SIE method, proposed based on the DINA model, considers estimation error in examinee attribute mastery patterns when calibrating new items, fully utilizing the posterior distribution of examinee attribute mastery patterns when calibrating new item Q-matrices and item parameters (Chen et al., 2015). The SIE method comprises two parts: Q-matrix calibration and item parameter calibration. For Q-matrix calibration of new items, the posterior distribution of attribute mastery patterns is first calculated for examinees who responded to new item j based on their responses to old items. Then, using the posterior distribution of examinee attribute mastery patterns and the probability of correct response for each attribute mastery pattern to new item j with q-vector $\mathbf{q}_j$, the posterior predictive distribution for examinee i with a specific response Y_{ij} is calculated:

$$f(Y_{ij}|\mathbf{q}_j, \delta_j) = \sum_{c=1}^{2^K} P(Y_{ij}|\alpha_c, \mathbf{q}_j, \delta_j) P(\alpha_c|\mathbf{Y}_{i,old}), \quad (2)$$

where δ_j denotes the item parameter vector. Under the DINA model, this includes the slipping parameter s_j and guessing parameter g_j. $P(Y_{ij} = 1|\alpha_c, \mathbf{q}_j, \delta_j)$ represents the probability of correct response to new item j for examinees with attribute mastery pattern α_c. $P(\alpha_c|\mathbf{Y}_{i,old})$ represents the posterior probability that examinee i's attribute mastery pattern is α_c based on their responses to O old items ($\mathbf{Y}_{i,old}$), calculated as:

$$P(\alpha_c|\mathbf{Y}_{i,old}) = \frac{\pi_c \prod_{o=1}^{O} P(Y_{io}|\alpha_c, \mathbf{q}_o, \delta_o)}{\sum_{c=1}^{2^K} \pi_c \prod_{o=1}^{O} P(Y_{io}|\alpha_c, \mathbf{q}_o, \delta_o)}, \quad (3)$$

where π_c denotes the prior probability of attribute mastery pattern α_c, and $P(Y_{io}|\alpha_c, \mathbf{q}_o, \delta_o)$ represents the probability of correct response to old item o for examinees with attribute mastery pattern α_c.

Finally, combining the examinee posterior predictive distribution and their responses to new item j, the likelihood is constructed and maximized to estimate the q-vector for the new item, expressed as:

$$\hat{\mathbf{q}}_j = \arg\max_{\mathbf{q}j \in \mathbf{Q}_j} \prod_{i=1}^{n_j} f(Y_{ij}|\mathbf{q}_j, \delta_j), \quad (4)$$

where $\mathbf{Q}_j$ represents the set of all 2^K possible q-vectors for new item j. Additionally, the SIE method uses the EM algorithm to estimate item parameters for new items.

It should be noted that when using the SIE method to calibrate new items under the DINA model, for any given item parameter estimate, all possible q-vectors for the new item must be substituted into the likelihood function to calculate the likelihood values corresponding to all possible q-vectors, based on which the q-vector and item parameters for the new item are calibrated. This is feasible under the DINA model because the number of item parameters does not change with the number of attributes measured by the item; all q-vectors correspond to exactly two item parameters (the slipping and guessing parameters). However, this is difficult to implement under the G-DINA model because the number of item parameters varies with the number of attributes measured by the item, and different q-vectors may correspond to different numbers of item parameters. For example, when an item measures two attributes, there are four item parameters; when it measures three attributes, there are eight item parameters. Therefore, when extending the SIE method from the DINA model to the G-DINA model, for item parameter estimates based on a particular q-vector, only one likelihood value is calculated by combining these parameter estimates with their corresponding q-vector. For instance, if the estimated item parameters are $\hat{\delta}_j(\mathbf{q}_j^{(1)})$ based on q-vector $\mathbf{q}_j^{(1)}$, only the likelihood value combining $\hat{\delta}_j(\mathbf{q}_j^{(1)})$ with $\mathbf{q}_j^{(1)}$ is calculated, not with other possible q-vectors such as $\mathbf{q}_j^{(2)}$, $\mathbf{q}_j^{(3)}$, etc. For all possible q-vectors of new item j and their corresponding item parameter estimates, a likelihood value can be calculated. If there are eight possible q-vectors for the new item, eight likelihood values can be computed, and the q-vector and item parameters corresponding to the maximum likelihood value are selected as the estimates. All other steps for using the SIE method to calibrate new items under the G-DINA model remain consistent with those under the DINA model.

3.1 Basic Idea of SCADOCM Development

Currently, most methods in data mining revolve around regularization methods, which are a type of coefficient shrinkage method that achieves feature selection

by compressing feature coefficients and has become a mainstream feature selection approach. Regularization methods are based on the idea of penalty, adding a penalty term to the objective function to minimize the new objective function and select important features. SCAD penalty is a regularization method that demonstrates good performance in feature selection (Fan & Li, 2001). For simplified expression, SCAD penalty is referred to as SCAD. The log-likelihood function based on SCAD can be expressed as:

$$\ell(\beta) = \ell_0(\beta) - \sum_{w=1}^{W} p_\lambda(|\beta_w|), \quad (5)$$

where $\ell_0(\beta)$ represents the log-likelihood function based on features constructing a regression equation. If logistic regression is constructed based on features, its log-likelihood function can be expressed as:

$$\ell_0(\beta) = \sum_{i=1}^{n} \{R_i \log p_i + (1 - R_i) \log(1 - p_i)\}, \quad (6)$$

where n represents the number of examinees, R_i represents examinee i's response to the dependent variable R, $\mathbf{D}_i$ represents the transpose of examinee i's response vector to the independent variable set $\mathbf{D}$, and β represents the vector of regression coefficients.

$\sum_{w=1}^{W} p_\lambda(|\beta_w|)$ is the penalty term for the log-likelihood function, where W is the dimension of the independent variable vector $\mathbf{D}$, and $p_\lambda(\cdot)$ is the penalty function constructed as follows:

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta| & \text{if } |\beta| \leq \lambda, \\ -\frac{(|\beta|^2 - 2a\lambda|\beta| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |\beta| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta| > a\lambda, \end{cases} \quad (7)$$

where a and λ are two parameters that need to be defined in the SCAD function. Fan and Li (2001) recommended $a = 3.7$, which has shown good performance in various feature selection problems. λ is a tuning parameter that substantially influences the performance of the SCAD method (Fan & Li, 2001; Fan & Lv, 2010; Fan & Tang, 2013; Zhang et al., 2010). Fan and Li (2001) suggested using cross-validation to select the λ parameter, and researchers have also proposed different λ parameter selection methods, such as GCV, AIC, and BIC criteria. The BIC criterion is a commonly used method for λ parameter selection (Wang et al., 2007; Zhang et al., 2010).

The first term of the SCAD log-likelihood function represents model fit, where smaller values indicate better fit; the second term penalizes the number of independent variables included in the model (model complexity), effectively balancing model fit and complexity.

The SCAD-based likelihood function $\ell(\beta)$ can be estimated using local quadratic approximations (LQA; Fan & Li, 2001). The LQA algorithm characteristically estimates regression coefficients that converge to zero as exactly zero, thereby simplifying the model and improving computational efficiency.

Estimating the q-vector for new item j can be viewed as a feature selection problem, where all attributes measured by the test are treated as candidate features, and important attributes are selected as the measured attributes for new item j to construct the q-vector (attributes measured by new item j are marked as 1, others as 0). If item j measures certain attributes, examinees with higher mastery probabilities for these attributes are more likely to answer item j correctly, while those with lower mastery probabilities are less likely to answer correctly. Therefore, the greater the influence of mastery probability for a particular attribute on the probability of correct response, the more important that attribute is for the item. Conversely, if the influence is negligible, the item likely does not measure that attribute. Treating examinees' responses to new item j as the dependent variable and examinees' mastery status of each test attribute as independent variables (candidate features), the SCAD log-likelihood function is constructed and minimized to select the measured attributes for new item j and construct its q-vector. Based on this idea, this study proposes the SCAD-based online calibration method (SCADOCM), which uses SCAD to calibrate the Q-matrix of new items and then employs the EM algorithm to calibrate item parameters. The calculation formulas and procedures for simultaneously calibrating Q-matrices and item parameters using SCADOCM are detailed below.

3.2 Algorithm Design for Q-Matrix and Item Parameter Calibration in SCADOCM

This section details how SCADOCM is used to estimate the q-vector and item parameters for new items. For estimating the q-vector of a new item, the estimation is first treated as a feature selection problem, and then an effective and feasible estimator is constructed through SCAD. In cognitive diagnosis, examinees' responses to new item j depend on their mastery of attributes. Generally, examinees who have mastered the attributes measured by new item j have a higher probability of answering it correctly. Conversely, if examinees who have mastered attribute k have a higher probability of correctly answering new item j, then new item j likely measures attribute k. How can we select attributes that significantly affect examinees' correct responses from all attributes measured by the test? SCAD, as a feature selection method with many excellent properties, offers a viable solution.

To calibrate the Q-matrix of new items using SCAD based on test attributes and examinee responses, a regression model between attributes and examinee responses must first be constructed. The key step is finding an appropriate indicator to describe examinees' mastery of attributes. Examinees' marginal mastery probabilities for test attributes can be estimated based on their re-

sponses to old items during the CD-CAT process, providing a good reflection of their attribute mastery. Higher marginal mastery probability for an attribute indicates a greater likelihood of mastering that attribute. Additionally, examinees' responses to new item j follow a Bernoulli distribution. Therefore, for new item j, the following logistic regression model can be constructed based on examinees' marginal mastery probabilities for measured attributes and their item responses:

$$\text{logit}(P(Y_{ij}=1)) = \beta_{j0} + \sum_{k=1}^{K} \beta_{jk} P(\alpha_{ik}=1|\mathbf{Y}_{i,old}), \quad (8)$$

where $\mathbf{D}$ represents the $n_j \times K$ matrix of examinee attribute marginal mastery probabilities, and β_j represents the $K \times 1$ vector of attribute regression coefficients.

The log-likelihood function can then be constructed based on this regression equation, expressed as:

$$\ell_0(\beta_j) = \sum_{i=1}^{n_j} \left\{ Y_{ij} \log p_{ij} + (1-Y_{ij}) \log(1-p_{ij}) \right\}, \quad (9)$$

where Y_{ij} represents examinee i's response to new item j. Adding SCAD to equation (9) yields the SCAD-based log-likelihood function:

$$\ell(\beta_j) = \ell_0(\beta_j) - \sum_{k=1}^{K} p_\lambda(|\beta_{jk}|), \quad (10)$$

As shown in equation (7), this study adopts the recommended $a = 3.7$ (Fan & Li, 2001) and uses the BIC criterion to select the λ parameter. For a given λ value, the BIC index can be calculated as:

$$\text{BIC}(\lambda) = -2\ell_0(\widehat{\beta}_j(\lambda)) + |\mathcal{A}_\lambda| \log(n_j), \quad (11)$$

where $\mathcal{A}_\lambda = \{k : \widehat{\beta}_{jk}(\lambda) \neq 0, k = 1, \ldots, K\}$ represents the active set excluding the intercept term, and $|\mathcal{A}_\lambda|$ denotes the size of this active set.

Finally, based on the λ parameter selected by the BIC criterion, minimizing equation (10) yields the estimate of β_j, expressed as:

$$\widehat{\beta}_j = \arg\min_{\beta j} \ell(\beta_j). \quad (12)$$

If $\widehat{\beta}_{jk} \neq 0$, then new item j measures attribute k. For example, if $K = 5$ and the first and fourth elements of $\widehat{\beta}_j$ are non-zero coefficients, then the q-vector

for new item j is $\mathbf{q}_j = [1, 0, 0, 1, 0]$. If for the λ parameter selected by the BIC criterion, $|\mathcal{A}_\lambda| = 0$, then the attribute corresponding to the largest regression coefficient among the estimates obtained when λ takes its minimum value is selected as the measured attribute for new item j to ensure that new item j measures at least one attribute. The range of λ values refers to the method proposed in the study by Breheny and Huang (2011).

In SCADOCM, after using SCAD to calibrate the q-vector of a new item, item parameters are estimated based on this q-vector, specifically by employing the EM algorithm (Chen et al., 2015). In the E-step, the posterior distribution for each examinee is first calculated based on examinee i's response to new item j:

$$P(\alpha_c|\mathbf{Y}_{i,old}, Y_{ij}) = \frac{P(Y_{ij}|\alpha_c, \mathbf{q}_j, \delta_j)P(\alpha_c|\mathbf{Y}_{i,old})}{\sum_{c=1}^{2^K} P(Y_{ij}|\alpha_c, \mathbf{q}_j, \delta_j)P(\alpha_c|\mathbf{Y}_{i,old})}. \quad (13)$$

Then, based on the response vector $\mathbf{Y}_j$ of n_j examinees to new item j and each examinee's posterior distribution of attribute mastery patterns, assuming independence of responses from n_j examinees to new item j, the log-marginal likelihood function can be constructed as:

$$\log L(\delta_j|\mathbf{Y}_j, \mathbf{q}_j) = \sum_{i=1}^{n_j} \log \left[\sum_{c=1}^{2^K} P(Y_{ij}|\alpha_c, \mathbf{q}_j, \delta_j)P(\alpha_c|\mathbf{Y}_{i,old}) \right]. \quad (14)$$

The M-step maximizes equation (14) to estimate the item parameters δ_j for new item j. The EM algorithm iterates between the E-step and M-step until a pre-set convergence criterion is met.

3.3 Basic Steps for Simultaneous Q-Matrix and Item Parameter Calibration in SCADOCM

The specific steps for simultaneous calibration of Q-matrices and item parameters for new items using SCADOCM are as follows:

Step 1: New item q-vector estimation. For new item j, based on the marginal mastery probabilities of examinees who responded to new item j for each attribute and their response data to new item j, construct the SCAD-based log-likelihood function to obtain the estimated q-vector for new item j.

Step 2: New item parameter estimation. Treat the estimated q-vector from Step 1 as the true q-vector for new item j. Based on the posterior distributions of attribute mastery patterns of examinees who responded to new item j and their responses to new item j, use the item parameter estimation method in SCADOCM to estimate the item parameters. New item j is now calibrated.

Step 3: Repeat Steps 1 and 2 for all other new items to be calibrated to obtain estimates of their Q-matrices and item parameters. The process terminates when all new items have been calibrated.

4 Study 1: Performance Validation of SCADOCM Under Simulated Item Banks and Comparison with the SIE Method

Study 1 aims to examine the performance of SCADOCM in calibrating new items under different calibration sample sizes (50, 100, 500, 1000, 2000), attribute mastery pattern distributions (uniform, higher-order, multivariate normal), and item qualities (high quality: $P_j(0)$ and $1-P_j(1)$ randomly drawn from $U[0.05, 0.15]$; low quality: $P_j(0)$ and $1-P_j(1)$ randomly drawn from $U[0.1, 0.3]$), and to compare it with the SIE method. The calibration sample refers to the number of examinees who responded to new item j. This study adopts the settings from Chen and Xin (2011b) and Chen et al. (2015), where $n_j = (N \times Z)/m$, with N being the total number of examinees participating in CD-CAT, Z being the number of new items each examinee answers, and m being the number of new items to be calibrated. This study includes $5 \times 3 \times 2 = 30$ simulation conditions, with each condition replicated 100 times to reduce random error.

4.1.1 Generation of Examinee Attribute Mastery Patterns and Item Banks

The calibration sample size has five levels: $n_j = 50, 100, 500, 1000$, and 2000. Examinee attribute mastery patterns were generated from uniform, higher-order, and multivariate normal distributions MVN(0, Σ). Under the uniform distribution, examinee attribute mastery patterns were generated with equal probability from all possible patterns. Under the higher-order distribution, whether examinee i masters attribute k is related to their general latent ability θ_i. The probability of examinee i with ability θ_i mastering attribute k is:

$$P(\alpha_{ik} = 1|\theta_i, \lambda_{0k}, \lambda_{1k}) = \frac{\exp(\lambda_{1k}\theta_i + \lambda_{0k})}{1 + \exp(\lambda_{1k}\theta_i + \lambda_{0k})}, \quad (15)$$

where λ_{0k} and λ_{1k} are structural parameters. In this study, $K = 5$, $\lambda_0 = (-1, -0.5, 0, 0.5, 1)$, and $\lambda_{1k} = 1.5$ for all attributes k. Examinee ability values were generated from $N(0, 1)$ (de la Torre & Chiu, 2016). A random number between 0 and 1 was generated and compared with the probability calculated using equation (15). If the probability exceeded the random number, examinee i mastered attribute k ($\alpha_{ik} = 1$); otherwise, examinee i did not master attribute k ($\alpha_{ik} = 0$) (Ma & de la Torre, 2020). Under the multivariate normal distribution, the correlation between attributes was set to 0.5 (J. Chen, 2017; Chiu, 2013). Assuming examinee i's ability vector is θ_i, examinee i's attribute mastery pattern α_i can be obtained using (Chiu, 2013):

$$\alpha_{ik} = \begin{cases} 1 & \text{if } \Phi(\theta_{ik}) > u_{ik}, \\ 0 & \text{otherwise,} \end{cases} \quad (16)$$

where $\Phi(\cdot)$ is the inverse of the normal distribution probability density function.

Item bank generation includes Q-matrix generation and item parameter generation. The item bank contains 300 items, each measuring at most three attributes, with 100 items measuring one attribute, 100 measuring two attributes, and 100 measuring three attributes. The test measures $K = 5$ attributes, resulting in 31 possible item q-vectors: five q-vectors measuring one attribute, ten measuring two attributes, and ten measuring three attributes. The five one-attribute q-vectors were repeated 20 times, the ten two-attribute q-vectors were repeated 10 times, and the ten three-attribute q-vectors were repeated 10 times to form a temporary 300×5 test Q-matrix.

Item parameters were generated as follows: The intercept parameter δ_{j0} was randomly drawn from $U[0.05, 0.15]$, and main effect parameters δ_{jk} were randomly drawn from $U[0.1, 0.3]$. Correct response probabilities for other attribute mastery patterns on new item j were randomly generated from $U[0.05, 0.95]$ while satisfying the monotonicity condition: examinees mastering more attributes have higher probabilities of correctly answering item j than those mastering fewer attributes (de la Torre & Chiu, 2016).

4.1.2 New Item Generation

New item generation includes Q-matrix and item parameter generation. The number of new items to be calibrated was set to $m = 20$, with the new item Q-matrix being a 20×5 matrix. Twenty rows were randomly selected from the simulated Q-matrix to construct the new item Q-matrix. New item parameters were generated using the same procedure as for bank items. After generating true examinee attribute mastery patterns and true item parameters, the probability of correct response for each examinee on each new item was calculated based on the specified cognitive diagnosis model. This probability was compared with a random number between 0 and 1; if the probability exceeded the random number, the item was answered correctly, otherwise incorrectly.

4.2 CD-CAT Process and New Item Calibration

A fixed-length termination rule was used, with each examinee answering 20 old items and 5 new items ($Z = 5$). The CD-CAT simulation proceeded as follows:

At the beginning of the test, nothing is known about the examinee, so (1) one item was randomly selected from the bank as the initial item; (2) the examinee's response to the current item was simulated, and then based on responses to administered items, the Shannon entropy (SHE; Cheng, 2009) item selection strategy was used to select the most suitable item from the remaining bank

as the next item. This process was repeated until the test length reached the predetermined standard. The SHE item selection strategy has a solid theoretical foundation and high estimation accuracy. Previous studies on simultaneous calibration of Q-matrices and item parameters have also shown that online calibration methods perform well under the SHE strategy (Chen et al., 2015; Tan et al., 2022; Zheng & Chang, 2016; Tan et al., 2021; Zhang, 2010). Therefore, this study selected SHE as the item selection strategy; (3) Maximum likelihood estimation (MLE) was used to estimate examinee attribute mastery patterns.

During the CD-CAT simulation, five new items were randomly selected from the 20 new items to be calibrated and placed at random positions in each examinee's test. After CD-CAT administration, SCADOCM and SIE methods were used to calibrate new items based on examinee attribute marginal mastery probabilities, posterior distributions of attribute mastery patterns, and examinee responses to new items.

4.3 Evaluation Criteria

Calibration efficiency: Average running time (ART) was used to evaluate the calibration efficiency of online calibration methods, calculated as:

$$\text{ART} = \frac{1}{R}\sum_{r=1}^{R} t_r, \quad (17)$$

where t_r represents the time used by each online calibration method to calibrate new items in the rth replication. Smaller ART values indicate higher calibration efficiency. All experiments were run on a computer with an Intel Core i5-8400 2.81GHz processor and 20GB memory to ensure comparable estimation efficiency across methods.

Attribute vector correct estimation rate (AVCER): AVCER was used to evaluate Q-matrix estimation accuracy for new items, calculated as:

$$\text{AVCER} = \frac{1}{R}\sum_{r=1}^{R} I(\hat{\mathbf{q}}_{jr} = \mathbf{q}_{jr}), \quad (18)$$

where r represents the rth replication of 100 simulation experiments, $\hat{\mathbf{q}}_{jr}$ represents the estimated q-vector for new item j in the rth replication, and $\mathbf{q}_{jr}$ represents the true q-vector for new item j in the rth replication. $I(\cdot)$ is an indicator function evaluating whether $\hat{\mathbf{q}}_{jr}$ equals $\mathbf{q}_{jr}$ in the rth replication. Higher AVCER values indicate higher Q-matrix estimation accuracy.

Root mean squared error (RMSE): RMSE was used to evaluate item parameter estimation accuracy for new items, expressed as:

$$\text{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\frac{1}{2^{K_j^*}}\sum_{c=1}^{2^{K_j^*}}\left(\hat{P}_{jr}(\alpha_c) - P_{jr}(\alpha_c)\right)^2}, \quad (19)$$

where $\hat{P}_{jr}(\alpha_c)$ and $P_{jr}(\alpha_c)$ represent the estimated and true probabilities of correct response for examinees with attribute mastery pattern α_c on new item j in the rth replication, respectively. Smaller RMSE values indicate higher item parameter estimation accuracy. Additionally, the RMSE calculation formulas for $P(0)$ and $1 - P(1)$ parameters differ slightly from equation (19) and are as follows:

For $P(0)$:

$$\text{RMSE}_{P(0)} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{P}_{jr}(0) - P_{jr}(0)\right)^2}, \quad (20)$$

For $1 - P(1)$:

$$\text{RMSE}_{1-P(1)} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left((1 - \hat{P}_{jr}(1)) - (1 - P_{jr}(1))\right)^2}. \quad (21)$$

4.4 Results of Study 1

Figures 1 [Figure 1: see original paper] through 3 [Figure 3: see original paper] and Table 1 present the item calibration efficiency and accuracy results for SCADOCM and SIE methods under the simulated item bank. Across all simulation conditions, SCADOCM achieved mean ART, AVCER, and RMSE values of 5.231s, 66.4%, and 0.101, respectively, while the corresponding values for the SIE method were 99.893s, 0.0%, and 0.242. It should be noted that the AVCER values for the SIE method were close to 0.0%, possibly because the MLE method used in SIE for estimating new item q-vectors under the G-DINA model tends to select the q-vector measuring all attributes (Wang et al., 2020; Chen et al., 2013). Overall, SCADOCM demonstrates good estimation efficiency and item calibration accuracy, outperforming the SIE method.

Figure 1 shows the average running time (in seconds) for estimating 20 new items using SCADOCM and SIE methods. The SIE method has lower estimation efficiency than SCADOCM, with an average ART value across all conditions approximately 19.095 times that of SCADOCM. The average ART values for SCADOCM and SIE were 5.231s and 99.893s, respectively. Regarding the effect of calibration sample size on method efficiency, both SCADOCM and SIE showed increased average running time with larger calibration samples. When the calibration sample was 50, the average ART values for SCADOCM and SIE were 1.216s and 25.554s, respectively, increasing to 12.643s and 222.052s

when the calibration sample was 2000. Item quality had minimal impact on the calibration efficiency of both SCADOCM and SIE. When item parameter ranges were $U(0.05, 0.15)$ and $U(0.1, 0.3)$, SCADOCM's average ART values were 6.543s and 3.920s, respectively, while SIE's average ART values were 81.624s and 118.162s. SCADOCM's calibration efficiency was less affected by attribute mastery pattern distribution, while SIE's efficiency under uniform and higher-order distributions was slightly better than under normal distribution. The average ART values for SCADOCM and SIE were 4.304s and 58.204s under uniform distribution, 4.615s and 65.781s under higher-order distribution, and 6.776s and 175.695s under normal distribution.

Figure 2 [Figure 2: see original paper] shows that SCADOCM achieves higher Q-matrix estimation accuracy than the SIE method. Calibration sample size, item quality, and attribute mastery pattern distribution all affect SCADOCM's Q-matrix estimation accuracy, while having negligible effects on the SIE method. The AVCER values for the SIE method were close to 0 across all simulation conditions. SCADOCM's Q-matrix estimation accuracy improved with increasing calibration sample size. The mean AVCER values for SCADOCM across calibration sample sizes of 50, 100, 500, 1000, and 2000 were 38.3%, 48.9%, 74.5%, 82.3%, and 88.3%, respectively. After reaching a certain sample size, the impact of sample size on SCADOCM's Q-matrix estimation accuracy gradually diminished. When calibration sample size increased from 50 to 100, the AVCER difference was 10.6%; from 100 to 500, the difference was 25.6%, with an average increase of 3.2% per 50 examinees; from 1000 to 2000, the difference was only 6.0%, with an average increase of 0.3% per 50 examinees. Higher item quality led to higher Q-matrix estimation accuracy for SCADOCM. When item parameter ranges changed from $U(0.05, 0.15)$ to $U(0.1, 0.3)$, AVCER values decreased monotonically under fixed calibration sample size and attribute mastery pattern distribution. With item parameter range $U(0.05, 0.15)$, SCADOCM's AVCER values ranged from 40.4% to 96.0%; with range $U(0.1, 0.3)$, values ranged from 30.2% to 89.4%. Regarding the effect of attribute mastery pattern distribution on Q-matrix calibration accuracy, SCADOCM's Q-matrix estimation accuracy was best under uniform distribution, followed by higher-order distribution, and worst under normal distribution in most conditions. The possible reason is that under uniform distribution, the number of examinees for each attribute mastery pattern is relatively balanced, whereas under higher-order and normal distributions, some attribute mastery patterns have very few examinees, especially under normal distribution where certain patterns have even fewer examinees. This is not conducive to identifying the correct q-vector (Chiu, 2013; Wang et al., 2018), resulting in lower Q-matrix estimation accuracy under higher-order and normal distributions. SCADOCM's Q-matrix estimation accuracy ranges were 35.2%–96.0% under uniform distribution, 33.7%–93.4% under higher-order distribution, and 30.2%–86.0% under normal distribution. However, when item parameter range was $U(0.05, 0.15)$ and calibration sample size was 100, SCADOCM's AVCER value was larger under higher-order distribution than uniform distribution, with values of 59.9% and 58.4%, respectively.

Figure 3 presents the item parameter calibration results for SCADOCM and SIE. SCADOCM's item parameter calibration accuracy is superior to that of the SIE method, with both methods affected by calibration sample size, item quality, and attribute mastery pattern distribution. Item parameter calibration accuracy for both SCADOCM and SIE improved with increasing calibration sample size. Across calibration sample sizes, SCADOCM's mean RMSE values were 0.188, 0.145, 0.076, 0.057, and 0.042, while SIE's mean RMSE values were 0.400, 0.337, 0.200, 0.156, and 0.120. The RMSE differences between calibration sample sizes of 50 and 2000 were 0.146 for SCADOCM and 0.280 for SIE, indicating that calibration sample size had a slightly greater impact on SIE than on SCADOCM. Item parameter calibration accuracy for both SCADOCM and SIE increased slightly with item quality in some conditions but decreased slightly in others. Overall, SCADOCM's mean RMSE values were 0.101 (range: 0.020–0.231) and 0.102 (range: 0.025–0.220) under the two item parameter ranges ($U(0.05, 0.15)$ and $U(0.1, 0.3)$), showing a slight increase in mean RMSE. SIE's mean RMSE values were 0.235 (range: 0.046–0.448) and 0.250 (range: 0.058–0.429), also showing increased mean RMSE. Under normal distribution, SCADOCM had larger RMSE values when item parameter range was $U(0.05, 0.15)$, with a maximum RMSE difference of 0.013 between the two parameter ranges. Under normal distribution with calibration sample sizes of 50 and 100, SIE had larger RMSE values when item parameter range was $U(0.05, 0.15)$, with an RMSE difference of 0.019 at sample size 50. This may result from the interaction between calibration sample size and attribute mastery pattern distribution. Item parameter calibration accuracy is lower with small calibration samples, and when sample size is small and attribute mastery pattern distribution is normal, there may be cases where some attribute mastery patterns have many examinees while others are missing. The combined effect may lead to slightly higher RMSE values for high-quality items than for low-quality items, but this difference is small and can be reversed by increasing sample size or changing the attribute mastery pattern distribution. Regarding the effect of attribute mastery pattern distribution on item parameter calibration accuracy, both SCADOCM and SIE performed best under uniform distribution, followed by higher-order distribution, and worst under normal distribution. Under uniform, higher-order, and normal distributions, SCADOCM's RMSE ranges were 0.020–0.154, 0.028–0.185, and 0.070–0.231, respectively, while SIE's ranges were 0.046–0.378, 0.079–0.403, and 0.221–0.448.

Table 1 presents the $P(0)$ and $1 - P(1)$ parameter calibration results for SIE and SCADOCM. The results show that SCADOCM achieves good calibration accuracy for $P(0)$ and $1 - P(1)$ parameters, outperforming the SIE method, especially with small calibration samples. Both SIE and SCADOCM are affected by calibration sample size, item quality, and attribute mastery pattern distribution. Calibration accuracy for $P(0)$ and $1 - P(1)$ parameters improved with increasing calibration sample size for both methods. For the $P(0)$ parameter, SIE's mean RMSE values across calibration sample sizes were 0.223, 0.155, 0.066, 0.046, and 0.032, while SCADOCM's corresponding values were

0.155, 0.120, 0.048, 0.032, and 0.022. For the $1 - P(1)$ parameter, SIE's mean RMSE values were 0.235, 0.163, 0.067, 0.046, and 0.033, while SCADOCM's values were 0.118, 0.087, 0.037, 0.026, and 0.018. Calibration accuracy for $P(0)$ and $1 - P(1)$ parameters increased with item quality for both methods, except when calibration sample size was 50. At sample size 50, SCADOCM's calibration accuracy was higher for low-quality items than high-quality items, but the RMSE difference was small, with a maximum difference of 0.022. Regarding the effect of attribute mastery pattern distribution on $P(0)$ and $1 - P(1)$ parameter calibration accuracy, both SIE and SCADOCM performed slightly better under higher-order distribution than under uniform and normal distributions. For the $P(0)$ parameter, SIE's RMSE ranges were 0.038–0.362, 0.019–0.180, and 0.023–0.229 under uniform, higher-order, and normal distributions, respectively, while SCADOCM's ranges were 0.018–0.184, 0.014–0.133, and 0.019–0.161. For the $1 - P(1)$ parameter, SIE's RMSE ranges were 0.039–0.356, 0.019–0.186, and 0.023–0.232, while SCADOCM's ranges were 0.015–0.122, 0.013–0.107, and 0.017–0.134.

5 Study 2: Performance Validation of SCADOCM Under Real Item Banks

Based on Study 1's results and considering that the SIE method exhibited low Q-matrix calibration accuracy across all experimental conditions and is not suitable for G-DINA and similar models, Study 2 only examines the performance of SCADOCM under a real item bank across different calibration sample sizes (50, 100, 500, 1000, 2000) and attribute mastery pattern distributions (uniform, higher-order, multivariate normal). This study includes $5 \times 3 = 15$ simulation conditions, with each condition replicated 100 times to reduce random error.

5.1 Real Item Bank and New Item Specification

Real item bank: Due to its unique advantage of providing comprehensive and detailed symptom profiles for patients, cognitive diagnosis is increasingly applied in psychological disorder assessment and diagnosis. Researchers have applied cognitive diagnosis to the assessment and diagnosis of pathological gambling, schizotypal personality, borderline personality, anxiety, depression, and Internet addiction (de la Torre et al., 2018; Peng et al., 2019; Templin & Henson, 2006; Tu et al., 2017; Xi et al., 2020; Shi, 2017). Shi (2017) constructed an Internet addiction item bank based on the symptom criteria for Internet addiction defined in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), and validated the bank's reliability and validity as meeting psychometric requirements in practice. This study uses this Internet addiction item bank as the real item bank, which contains 263 dichotomously scored items, each measuring at most three attributes (symptom criteria), with nine attributes measured in total (as shown in Table 2). According to DSM-5 diagnostic criteria, examinees meeting five or more of the nine symptom criteria can be diagnosed with Internet addiction. This study uses the original Q-matrix

from Shi (2017) as the true Q-matrix and estimates item parameters for the bank using the G-DINA model based on this Q-matrix and responses from 1,558 real examinees. Descriptive statistics for item parameters are presented in Table 3 , and parameter results for all bank items are shown in Appendix Table 1. The G-DINA model was selected for analysis primarily because it allows for both compensatory and non-compensatory relationships between attributes, making it suitable for analyzing Internet addiction tests. Model-data fit tests (Table 4) revealed that the G-DINA model fit the Internet addiction data better than other constrained cognitive diagnosis models such as DINA.

New item specification: Twenty items were randomly selected from the Internet addiction bank as new items to be calibrated for Q-matrix and item parameters.

The generation of examinee attribute mastery patterns, CD-CAT process, new item calibration, and evaluation criteria in Study 2 were consistent with Study 1. It should be noted that the "true" item parameters in Study 2 were estimated based on the expert-defined Q-matrix and all examinees' real response data from previous research. The RMSE index calculated based on these "true" values reflects consistency between item parameter estimation results.

5.2 Results of Study 2

Table 3 presents descriptive statistics for item parameters in the Internet addiction bank. Compared with item quality in the simulated bank of Study 1 ($P(0)/(1-P(1)) \sim U[0.05, 0.15]$), item quality in the Internet addiction bank is lower. Further validation of SCADOCM's performance under this real bank can examine its applicability range and robustness in practical applications.

Table 5 presents SCADOCM's item calibration efficiency, Q-matrix estimation accuracy, and item parameter calibration consistency results under the real bank. The results show that SCADOCM maintains good estimation efficiency, Q-matrix estimation accuracy, and item parameter calibration consistency under the real bank. Specifically, across all simulation conditions, SCADOCM's mean ART, AVCER, and RMSE values were 37.612s, 79.8%, and 0.101, respectively.

The average running time (in seconds) for estimating 20 new items using SCADOCM is shown in Table 5. SCADOCM's mean ART value was 37.612s. Regarding the effect of calibration sample size on SCADOCM's efficiency, average running time increased with calibration sample size. When calibration sample size was 50, SCADOCM's mean ART was 4.507s, increasing to 101.849s when sample size was 2000. Differences in SCADOCM's calibration efficiency across attribute mastery pattern distributions were small, with mean ART values of 37.567s, 38.060s, and 37.209s under uniform, higher-order, and normal distributions, respectively.

Table 5 shows that calibration sample size and attribute mastery pattern distri-

bution both affect SCADOCM's Q-matrix estimation accuracy. SCADOCM's Q-matrix estimation accuracy improved with increasing calibration sample size. Mean AVCER values across calibration sample sizes of 50, 100, 500, 1000, and 2000 were 57.0%, 69.8%, 88.0%, 91.2%, and 92.8%, respectively. Consistent with the simulated bank, after reaching a certain sample size, the impact of sample size on Q-matrix estimation accuracy gradually diminished. When calibration sample size increased from 50 to 100, the AVCER difference was 12.8%; from 100 to 500, the difference was 18.2%, with an average increase of 2.3% per 50 examinees; from 1000 to 2000, the difference was only 1.6%, with an average increase of 0.1% per 50 examinees. Regarding the effect of attribute mastery pattern distribution on Q-matrix calibration accuracy, SCADOCM performed best under uniform distribution, followed by higher-order distribution, and worst under normal distribution. Q-matrix estimation accuracy ranges were 69.7%–97.8% under uniform distribution, 56.0%–94.5% under higher-order distribution, and 45.4%–86.3% under normal distribution.

Consistent with the simulated bank, SCADOCM's item parameter calibration consistency was affected by calibration sample size and attribute mastery pattern distribution. Item parameter calibration consistency improved with increasing calibration sample size. Mean RMSE values across calibration sample sizes were 0.192, 0.135, 0.069, 0.050, and 0.041. SCADOCM's item parameter calibration consistency was best under uniform distribution, followed by higher-order distribution, and worst under normal distribution, with RMSE ranges of 0.019–0.142, 0.032–0.189, and 0.105–0.244, respectively.

6 Discussion and Future Directions

How can a developed CD-CAT remain effective in practical testing over the long term, efficiently providing accurate and detailed diagnostic results to test users? Effective item bank maintenance or updating methods are essential. Item replenishment plays a crucial role in bank maintenance, and online calibration is an effective item replenishment method. However, research on simultaneous online calibration methods for Q-matrices and item parameters in CD-CAT is limited, and existing methods are basically developed based on the DINA model. Research on simultaneous online calibration methods for Q-matrices and item parameters under the G-DINA model is almost non-existent, which to some extent hinders the further promotion of CD-CAT in practical testing.

This study developed a new online calibration method, SCADOCM, applicable to models such as G-DINA, based on the idea of feature selection using regularization methods, aiming to provide new methodological support for item replenishment in CD-CAT banks. The new method SCADOCM uses regularization methods to calibrate Q-matrices for new items. Compared with the optimal subset approach used in existing online calibration methods, this can effectively save time for new item calibration and provides a new perspective for research on simultaneous online calibration methods for Q-matrices and item parameters in CD-CAT. Monte Carlo simulation studies under both simulated and

real banks were conducted to test the feasibility and rationality of SCADOCM, examine the effects of factors such as calibration sample size, item quality, and attribute mastery pattern distribution on its performance, and compare it with the traditional SIE method. The results show that the new method SCADOCM achieves satisfactory calibration efficiency and accuracy across all simulation conditions and outperforms the SIE method. For example, under the simulated bank, SIE's mean ART value was 19.096 times that of SCADOCM, indicating that SCADOCM has higher calibration efficiency. SCADOCM's mean AVCER value was 66.4% higher than SIE's, and its mean RMSE value was 0.141 lower than SIE's, demonstrating better calibration accuracy. Additionally, the results show that SIE's Q-matrix estimation accuracy was close to 0 across all conditions. The possible reason is that the AVCER index used in this study evaluates the consistency between the entire estimated q-vector and the true q-vector, i.e., the pattern estimation accuracy of the q-vector. The SIE method uses MLE to estimate new item q-vectors, and under the G-DINA model, MLE tends to select the q-vector measuring all attributes (i.e., all elements equal to 1) as the estimated q-vector (Wang et al., 2020; Chen et al., 2013). For example, when the test measures $K = 5$ attributes, the SIE method selects q-vector $\mathbf{q} = [1, 1, 1, 1, 1]$ as the estimated q-vector, and experimental investigations confirmed this. In the simulation, the test measured 5 attributes, with each item (old and new) measuring at most 3 attributes. Using SIE to calibrate new item Q-matrices tended to specify that each item measured all 5 attributes, resulting in attribute vector estimation accuracy lower than random assignment probability and AVCER values around 0. Assuming all 20 new items measured 3 attributes, the true new item Q-matrix would have 60 elements equal to 1 and 40 equal to 0, giving SIE a maximum possible attribute estimation accuracy of 60%. In this study, the 20 new item q-vectors were randomly selected from 300 old items (with 100 items each measuring 1, 2, and 3 attributes), so in most cases the number of elements equal to 1 in the true new item Q-matrix was less than 50, making SIE's attribute estimation accuracy lower than 50%. In Study 1, SIE's mean attribute estimation accuracy across conditions was 39.8%, greater than 0 but less than 50%. Study 1 retained the SIE method as a comparison benchmark despite its extremely low AVCER values, primarily to provide reference for other researchers and practitioners, helping them avoid selecting this method as a comparison benchmark in future online calibration method research under G-DINA and similar saturated models. Additionally, the SIE method does not consider model complexity when calibrating new item Q-matrices and may not be suitable for saturated models such as G-DINA. Improvements could be made by penalizing model complexity. Specifically, when using SIE to calibrate new item Q-matrices, the likelihood could be penalized based on model complexity to construct a BIC index, selecting the q-vector that minimizes the BIC value as the estimated q-vector. Preliminary experiments showed that the improved SIE method achieved better item calibration accuracy than SIE. Under conditions where $P(0)$ and $1 - P(1)$ ranged from $U(0.1, 0.3)$, attribute mastery pattern distribution was normal, and calibration sample size was 500, the improved SIE method achieved mean ART, AVCER, RMSE, $P(0)$

RMSE, and $1 - P(1)$ RMSE values of 153.758s, 54.9%, 0.104, 0.058, and 0.048, respectively. Its Q-matrix calibration accuracy was much better than SIE's but still inferior to the new SCADOCM method (SCADOCM's AVCER was 61.7% under these conditions).

Although this study addresses the technical challenges of item replenishment in CD-CAT bank development and maintenance, it is closely related to psychological issues. Psychometrics is a tool for studying psychology, and the assessment and measurement of psychological problems (e.g., depression, anxiety) cannot be separated from psychometrics. As a new test form, CD-CAT can more efficiently and accurately screen patients with psychological problems, alleviating the burden of completing lengthy questionnaires for patients (e.g., those with depression or mania). More importantly, CD-CAT can help test users understand patients' performance on various symptoms of a psychological problem, obtain diagnostic results more quickly, and develop targeted treatment plans based on these results. The application of CD-CAT in psychological assessment is significant for both patients and test users. This study is committed to solving a major challenge in the continuous application of CD-CAT in practical testing, namely the technical difficulties encountered during item replenishment in CD-CAT bank construction and maintenance, promoting the application and promotion of CD-CAT in psychological assessment practice, with the goal of helping test users obtain more detailed diagnostic results and develop corresponding treatment plans. This is closely related to psychological issues.

Although this study enriches research on online calibration methods in CD-CAT, many aspects remain to be further improved and investigated. Specifically:

First, the performance of the new method SCADOCM using SCAD to calibrate new item Q-matrices is affected by the λ parameter. An appropriate and optimal λ value can improve SCADOCM's Q-matrix calibration accuracy and consequently enhance the method's item calibration accuracy (Fan & Li, 2001; Fan & Lv, 2010; Fan & Tang, 2013; Zhang et al., 2010). This study used the BIC criterion, which is commonly used and performs well in data mining, to select the λ value (Wang et al., 2007; Zhang et al., 2010). Although the results show that SCADOCM achieves satisfactory item calibration accuracy when using this criterion to select λ, whether better λ parameter selection criteria exist for simultaneous online calibration of Q-matrices and item parameters remains an issue worth exploring. Future research could systematically compare existing λ parameter selection criteria to provide recommendations and references for λ selection in SCADOCM.

Second, this study only considered fixed-length CD-CAT termination rules, whereas variable-length termination rules better reflect the adaptive nature of CD-CAT. How to calibrate new items under variable-length termination rules is a topic for future research. For example, how should new items be assigned to examinees under variable-length termination rules, and does the assignment method affect final item calibration accuracy? Additionally, this study's design focused on examining online calibration method performance and the effects of

related factors, without exploring measurement invariance issues. Unlike previous studies with complete examinee response matrices and known, correct item Q-matrices (Bradshaw & Madison, 2015; de la Torre & Lee, 2010; Madison & Bradshaw, 2018), in CD-CAT simultaneous calibration of Q-matrices and item parameters, the examinee response matrix is a sparse matrix lacking many responses—each item is answered by only some examinees, and each examinee answers only a few items (if examinees need to answer too many new items to be calibrated, CD-CAT test length may increase substantially, adding to examinee burden), and item Q-matrices are unknown. Even with large calibration samples (e.g., 1,000 examinees), item parameter calibration accuracy is low, and measurement invariance cannot be guaranteed. Bradshaw and Madison (2015) noted that strong measurement invariance is difficult to observe when parameter estimation accuracy is low, and they also mentioned that violations of model-data fit assumptions in other forms (e.g., Q-matrix misspecification, Bradshaw & Madison, 2015) may affect classification consistency. Therefore, whether measurement invariance can still be observed when the examinee response matrix is sparse and Q-matrices are unknown or misspecified, and under what conditions measurement invariance can be observed, are directions for future research.

Third, existing simultaneous online calibration methods for Q-matrices and item parameters in CD-CAT focus on examinee response data while ignoring process data that can be conveniently obtained in computerized testing, such as response times (RTs). Previous studies have shown that response time data can provide valuable information about examinees' cognitive processes and improve item parameter estimation accuracy (Kang et al., 2020; Klein Entink et al., 2009; van der Linden et al., 2010). Future research could consider calibrating new items within a joint framework of response and response time data to examine whether response time data helps improve the calibration accuracy of online calibration methods.

Fourth, this study assumes that the number of attributes measured by the CD-CAT bank is fixed and known. However, new attributes may be added to the bank periodically during continuous CD-CAT use. Undoubtedly, the performance of various online calibration methods will fluctuate with the addition of new attributes. How to improve the performance of existing simultaneous online calibration methods for Q-matrices and item parameters in CD-CAT when the number of measured attributes changes over time is a major challenge for researchers. Additionally, this study assumes that test attributes are independent. The performance of various online calibration methods when attribute hierarchies exist (e.g., linear, branching, convergent) remains to be explored.

Fifth, this study not only examined method performance under simulated banks but also further validated SCADOCM's performance under a real bank, ensuring ecological validity. The results show that SCADOCM's calibration performance is satisfactory under both simulated and real banks, demonstrating good generalizability and providing guidance for practical applications. However, consistent with previous domestic and international research on simultaneous online

calibration methods (Chen et al., 2015; Tan et al., 2022; Chen & Xin, 2011b; Tan et al., 2021), this study always used Monte Carlo simulation methods without applying them in empirical research settings to evaluate performance. The main reason is that validating online calibration method performance in real testing situations requires constructing a real CD-CAT testing platform that can be used for actual testing, which demands substantial time and effort and is currently difficult to obtain. This is a limitation of this study and current CD-CAT online calibration research, and also a direction for future research. In summary, research on simultaneous online calibration methods for Q-matrices and item parameters in CD-CAT remains to be further deepened.

7 Conclusions

The main conclusions of this study are:

1. SCADOCM demonstrates good item calibration performance and outperforms the SIE method. Additionally, the SIE method's Q-matrix estimation accuracy is close to 0 across all conditions, making it unsuitable for saturated models such as G-DINA.
2. Overall, SCADOCM and SIE achieve higher item calibration accuracy with large calibration samples, high item quality, and uniform/higher-order attribute mastery pattern distributions than with small samples, low item quality, and normal distributions.
3. SCADOCM has higher calibration efficiency with small calibration samples, and its efficiency is less affected by item quality and attribute mastery pattern distribution. The SIE method has higher calibration efficiency with small samples than with large samples, and its efficiency is higher under uniform and higher-order distributions than under normal distribution, with minimal impact from item quality.

Appendix: Item Parameter Values for the Internet Addiction Item Bank

Item parameter values for the Internet addiction bank are presented in Appendix Table 1, where $P(0)$, $P(1)$, $P(00)$, $P(10)$, $P(01)$, $P(11)$, $P(000)$, $P(100)$, $P(010)$, $P(001)$, $P(110)$, $P(101)$, $P(011)$, and $P(111)$ represent the probabilities of correct response for examinees with reduced attribute mastery patterns (if an item measures the first two of nine attributes, the reduced attribute mastery pattern is $\alpha_{ci}^{*} = (\alpha_{c1}, \alpha_{c2})$). For example, $P(0)$ and $P(1)$ represent the probabilities of correct response for examinees who have not mastered and have mastered, respectively, a single attribute measured by an item; $P(10)$ represents the probability of correct response for examinees who have mastered the first but not the second of two attributes measured by an item; $P(011)$ represents the probability of correct response for examinees who have mastered the second and third but not the first of three attributes measured by an item.