This is a "preproof" accepted article for *Psychometrika*.
This version may be subject to change during the production process.
DOI: 10.1017/psy.2026.10154

Online Calibration Designs of Multidimensional Computerized Adaptive Testing Using Kullback-Leibler Information

Hengxin Pang[1] and Yinhong He[2]

[1]School of Mathematics and Statistics, Nanjing University of Information Science and Technology

[2]School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing, China; email: roheyinhong@163.com.

Corresponding Author: Yinhong He, School of Mathematics and Statistics, Nanjing University of Information Science and Technology, No. 219, Ningliu Road, Nanjing 210044, China. Email: roheyinhong@163.com, Mobile Phone: 86-13120003060.

Abstract

Fisher information depends heavily on provisional parameter estimates, which limits its efficiency in the early calibration stage. To address this limitation, this study introduces Kullback–Leibler (KL) information into online calibration for multidimensional computerized adaptive testing (MCAT) and proposes four KL-

This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is unaltered and is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use or in order to create a derivative work.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

information-based designs: M-KI, M-PKI, M-SKI1, and M-SKI2, where M-PKI is a posterior-weighted extension of M-KI, and M-SKI1 and M-SKI2 are simplified approximations thereof. Simulation results showed that all four designs outperformed the Fisher-information-based M-D design in the early calibration stage, and that M-PKI further achieved the best final calibration precision, particularly in recovering the discrimination parameters. The simplified M-SKI1 and M-SKI2 substantially reduced computation time while preserving the early-stage advantage. In a semi-real data analysis, a hybrid design (M-SKI1&D), which applies M-SKI1 in the early stage before switching to M-D, yielded the best overall parameter recovery. Together, these findings establish KL-information-based designs as a promising alternative for MCAT online calibration, particularly when early-stage accuracy and the precise recovery of discrimination parameters are the primary concerns.

Keywords: multidimensional computerized adaptive testing, online calibration, KL information

1. Introduction

Grounded in multidimensional item response theory (MIRT), multidimensional computerized adaptive testing (MCAT) combines multi-trait estimation with adaptive testing to precisely estimate examinees' multidimensional latent traits. This capability has long supported large-scale measurement programs such as the Armed Services Vocational Aptitude Battery (Chen et al., 2017), and its range of applications has broadened considerably in recent years. In education, for instance, MCAT has been

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

implemented in K-12 formative assessments such as mathematical competence testing (Sung et al., 2026). In healthcare, it has been adopted for patient-reported outcome measures, including the Functional Assessment in Acute Care Multidimensional Computerized Adaptive Test (FAMCAT; Cheville et al., 2021) and the Multiple Sclerosis Quality of Life-54 (MSQOL-54; Giordano et al., 2023). Relative to unidimensional computerized adaptive testing (UCAT), MCAT not only provides richer diagnostic information but also achieves higher measurement accuracy with a shorter test length (e.g., Segall, 1996), and it has accordingly attracted growing research attention (e.g., Hassan & Miller, 2022, 2024; He & Qi, 2025; Mao et al., 2022; Yuan et al., 2023, 2026).

Maintaining a high-quality item bank is a core prerequisite for MCAT implementation, yet items inevitably become overexposed or outdated with prolonged use, necessitating the continuous development of replacement items (Wainer & Mislevy, 1990). Online calibration addresses this need by embedding new items within operational testing, collecting responses in real time, and estimating item parameters without interrupting the testing process, while also placing the new items on the same scale as the operational item bank (e.g., Chen et al., 2017; Hassan & Miller, 2024; He & Qi, 2025).

The online calibration design is the key factor determining the effectiveness of new item calibration, as it governs how new items are matched with examinees to optimize the precision of item parameter estimation. Compared with random designs,

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

adaptive online calibration designs have gained broader acceptance among researchers, precisely because they leverage incoming response data to guide item assignment. In UCAT, a range of Fisher-information-based adaptive designs has accordingly been developed: for static examinee pools, these include the D-optimal design (Berger, 1994; Jones & Jin, 1994), the sequential D-optimal design (Chang & Lu, 2010), and the optimal restricted design (Hassan & Miller, 2019); for dynamic examinee pools, notable proposals are the two-point D-optimal design (Ren et al., 2017), the Bayesian D-optimal design (van der Linden & Ren, 2015), the excellence degree design (He et al., 2020), the D-c design (He & Chen, 2020), and the compound optimal design (Song & Wang, 2025).

Research on MCAT online calibration designs, however, remains relatively limited. Since information-based designs in MCAT must accommodate the multidimensional ability structure, the core task has been to extend unidimensional information criteria to a multidimensional framework. For static examinee pools, Hassan and Miller (2024) extended the optimal restricted design (Hassan & Miller, 2019) to a multidimensional context. For dynamic examinee pools, Yang et al. (2023) extended the D-optimal and A-optimal designs to MCAT, which were later termed the M-D and M-A designs by He and Qi (2025), who further proposed response-time-based extensions, M-D-ET and M-A-ET. The M-D and M-A designs achieved comparable accuracy, whereas the ET versions improved efficiency by increasing the

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

information gained per unit of time and reducing response time without sacrificing accuracy.

Nevertheless, all existing online calibration designs, in both UCAT and MCAT, rely exclusively on Fisher information (collectively, Fisher information criteria). As a local measure (Chang & Ying, 1996), Fisher information depends heavily on provisional parameter estimates, which often deviate considerably from their true values in the early calibration stage, when only a small number of examinees have responded to the new items. Under these conditions, the Fisher information matrix fails to correctly reflect the information actually provided by examinees, resulting in low selection efficiency. Kullback-Leibler (KL) information, by contrast, is a global measure (Chang & Ying, 1996) that covers the entire parameter space and quantifies an examinee's ability to discriminate between the provisional estimate and any other value in that space. KL information should therefore offer higher selection efficiency than Fisher information precisely when provisional estimates are least stable, that is, in the early calibration stage.

KL information has, in fact, been widely used in item selection strategies. Chang and Ying (1996) first applied it to CAT, confirmed its advantage over Fisher information in the early stage of testing, and proposed the KL information index (KI) by integrating KL information over an appropriate interval. Veldkamp and van der Linden (2002) subsequently extended this index to multidimensional settings and proposed a Bayesian version, and Wang et al. (2011) further introduced two simplified

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

indices, SKI1 and SKI2, which resolved the time-consuming integration problem and substantially improved computational efficiency. Despite this line of work, KL information has not yet been explored in the context of MCAT online calibration designs. The present study addresses this gap by introducing KL information into MCAT online calibration and developing four new KL-information-based designs (collectively, KL information criteria), which are expected to overcome the low selection efficiency of Fisher information criteria in the early calibration stage.

The remainder of this paper is organized as follows. Section 2 briefly introduces the MIRT model, and Section 3 presents the Fisher-information-based M-D design. Section 4 details the construction of the KI indicator and the four proposed KL information criteria (M-KI, M-PKI, M-SKI1, and M-SKI2). Sections 5 and 6 report the designs and results of Simulation Studies 1 and 2, respectively, and Section 7 evaluates the performance of each design using semi-real data. The paper concludes with a discussion and directions for future research.

2. Multidimensional Item Response Theory Model

In the compensatory multidimensional two-parameter logistic model (M2PLM; Reckase, 2009), the probability that examinee *i*, with a *p*-dimensional ability vector $\boldsymbol{\theta}_i = \left(\theta_{i1}, \theta_{i2}, ...\theta_{ip}\right)^T$, answers item *j* correctly is

$$P\left(u_{ij}=1 \mid \boldsymbol{\theta}_i, \boldsymbol{a}_j, b_j\right) = \frac{1}{1+exp\left[-\left(\boldsymbol{a}_j^T\boldsymbol{\theta}_i - b_j\right)\right]}, \tag{1}$$

where u_{ij} is the response of examinee *i* to item *j*, taking values of 0 or 1 for incorrect and correct answers, respectively; $\boldsymbol{a}_j = \left(a_{j1}, a_{j2}, ...a_{jp}\right)^T$ represents the *p*-dimensional

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

discrimination parameter vector of item *j*; and b_j is the intercept term related to item difficulty, with the corresponding conventional difficulty parameter given by $b_j \Big/ \sqrt{a_{j1}^2 + a_{j2}^2 + ... + a_{jp}^2}$. $\boldsymbol{\eta}_j = \left(a_{j1}, a_{j2}, ...a_{jp}, b_j\right)^T$ denotes the parameter vector of item *j*.

This study primarily focuses on scenarios where the ability vector is two-dimensional, allowing the model to be simplified as the two-dimensional two-parameter logistic model (2D-2PLM):

$$P\left(u_{ij}=1 \mid \boldsymbol{\theta}_i, \boldsymbol{a}_j, b_j\right) = \frac{1}{1+exp\left[-\left(a_{j1}\theta_{i1} + a_{j2}\theta_{i2} - b_j\right)\right]}. \tag{2}$$

3. M-D Design Based on Fisher Information

The likelihood function derived from the response data serves as the foundation for computing Fisher information. Specifically, for examinee *i* on item *j*, the likelihood based on the response data can be expressed as:

$$L(u_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\eta}_j) = P^{u_{ij}}(\boldsymbol{\theta}_i; \boldsymbol{\eta}_j) Q^{1-u_{ij}}(\boldsymbol{\theta}_i; \boldsymbol{\eta}_j), \tag{3}$$

where $Q(\boldsymbol{\theta}_i; \boldsymbol{\eta}_j) = 1 - P(\boldsymbol{\theta}_i; \boldsymbol{\eta}_j)$.

Thus, the Fisher information matrix contributed by examinee *i* for new item *j* can be expressed as

$$\boldsymbol{I}(\boldsymbol{\eta}_j; \boldsymbol{\theta}_i) = -E\left[\frac{\partial^2}{\partial \boldsymbol{\eta}_j \partial \boldsymbol{\eta}_j^T} \ln L(u_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\eta}_j)\right]. \tag{4}$$

Within the framework of the 2D-2PLM, Equation (4) can be simplified as

$$\boldsymbol{I}(\boldsymbol{\eta}_j; \boldsymbol{\theta}_i) = P(\boldsymbol{\theta}_i; \boldsymbol{\eta}_j) Q(\boldsymbol{\theta}_i; \boldsymbol{\eta}_j) \begin{bmatrix} \theta_{i1}^2 & \theta_{i1}\theta_{i2} & -\theta_{i1} \\ \theta_{i1}\theta_{i2} & \theta_{i2}^2 & -\theta_{i2} \\ -\theta_{i1} & -\theta_{i2} & 1 \end{bmatrix}. \tag{5}$$

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Building on the traditional D-optimal design, van der Linden and Ren (2015) proposed a reformed D-optimal design for UCAT that, under a dynamic examinee pool, prioritizes items offering the greatest potential information gain. Rather than relying on the absolute information of an item alone, the design evaluates each candidate examinee's marginal contribution to that item's cumulative information, thereby allocating examinees more evenly across new items and improving overall calibration precision. This design has since been generalized to the multidimensional context as the M-D design (He & Qi, 2025; Yang et al., 2023), under which new item j is assigned to examinee i according to

$$j = \arg\max_{j \in H} \left\{ \det\left[\sum_{k=1}^{n_j} \boldsymbol{I}\left(\boldsymbol{\eta}_j; \boldsymbol{\theta}_k\right) + \boldsymbol{I}\left(\boldsymbol{\eta}_j; \boldsymbol{\theta}_i\right) \right] - \det\left[\sum_{k=1}^{n_j} \boldsymbol{I}\left(\boldsymbol{\eta}_j; \boldsymbol{\theta}_k\right) \right] \right\}, \tag{6}$$

where H denotes the set of new items pending calibration; $\boldsymbol{I}\left(\boldsymbol{\eta}_j; \boldsymbol{\theta}_i\right)$ denotes the Fisher information contributed by examinee i to the parameter vector of new item j; n_j denotes the number of examinees who have responded to new item j; and $\sum_{k=1}^{n_j} \boldsymbol{I}\left(\boldsymbol{\eta}_j; \boldsymbol{\theta}_k\right)$ denotes the cumulative Fisher information provided by the n_j examinees. Because the true item parameters $\boldsymbol{\eta}_j$ are unknown during calibration, Equation (6) is evaluated using the current provisional estimates $\hat{\boldsymbol{\eta}}_j = \left(\hat{a}_{j1}, \hat{a}_{j2}, \hat{b}_j\right)^T$ in place of $\boldsymbol{\eta}_j$.

4. Online Calibration Designs Based on KL Information

Chang and Ying (1996) were the first to note that Fisher information is inherently local: during the early stage of testing, examinees' estimated abilities often deviate substantially from their true levels, undermining the effectiveness of maximizing

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Fisher information. An analogous problem arises in online calibration. As more examinees respond to a new item *j*, its estimated parameter vector $\hat{\boldsymbol{\eta}}_j$ gradually converges toward the true value $\boldsymbol{\eta}_j$. Asymptotically, $\hat{\boldsymbol{\eta}}_j$ follows a normal distribution with mean $\boldsymbol{\eta}_j$ and variance $\boldsymbol{I}^{-1}\left(\boldsymbol{\eta}_j\right)$. Consequently, greater Fisher information yields smaller asymptotic variance, bringing $\hat{\boldsymbol{\eta}}_j$ closer to $\boldsymbol{\eta}_j$. In the early calibration stage, however, the small sample size leads to a considerable discrepancy between $\hat{\boldsymbol{\eta}}_j$ and $\boldsymbol{\eta}_j$; the information centered at $\hat{\boldsymbol{\eta}}_j$ therefore contributes little to updating the item parameter estimates, rendering Fisher information criteria inefficient for allocating new items. KL information overcomes this limitation by providing a global measure of divergence that does not depend on the accuracy of a single point estimate.

4.1. KL Information

KL information, also known as KL divergence (Kullback & Leibler, 1951), measures the divergence between two probability distributions, $f\left(x\right)$ and $g\left(x\right)$, defined as

$$\mathrm{KL}\left(g \,\|\, f\right) = E_f\left[\log\frac{f\left(x\right)}{g\left(x\right)}\right], \tag{7}$$

where $f\left(x\right)$ represents the true distribution of the observed data, and $g\left(x\right)$ represents an approximate distribution of $f\left(x\right)$. KL information quantifies the ease of distinguishing between two distributions (Henson & Douglas, 2005). In the context of MCAT online calibration, KL information can be used to measure an examinee's ability to discriminate between any possible parameter value $\boldsymbol{\eta} = \left(a_1, a_2, b\right)^T$ and the true parameter value $\boldsymbol{\eta}_j$ for a new item *j*:

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

$$\mathrm{KL}\left(\boldsymbol{\eta}\,\|\,\boldsymbol{\eta}_j\right)=P\left(\boldsymbol{\eta}_j\right)\log\frac{P\left(\boldsymbol{\eta}_j\right)}{P\left(\boldsymbol{\eta}\right)}+\left(1-P\left(\boldsymbol{\eta}_j\right)\right)\log\frac{\left(1-P\left(\boldsymbol{\eta}_j\right)\right)}{\left(1-P\left(\boldsymbol{\eta}\right)\right)}, \tag{8}$$

Equations (4) and (8) reveal a key distinction between the two information measures. Fisher information $\boldsymbol{I}\left(\boldsymbol{\eta}_j\right)$ is a function of $\boldsymbol{\eta}_j$ alone, reflecting its dependence on a single point. KL information $\mathrm{KL}\left(\boldsymbol{\eta}\,\|\,\boldsymbol{\eta}_j\right)$, by contrast, is a function of both $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_j$, quantifying the ability to discriminate $\boldsymbol{\eta}_j$ from any other value $\boldsymbol{\eta}$ without requiring $\boldsymbol{\eta}$ to be close to $\boldsymbol{\eta}_j$. Given $\boldsymbol{\eta}_j$, Fisher information is thus a local information matrix evaluated at a single point, whereas KL information remains a global function over the entire parameter space.

Despite this distinction, the two information measures are mathematically connected: the Fisher information matrix equals the Hessian of the KL information with respect to the item-parameter argument, evaluated at $\boldsymbol{\eta}=\boldsymbol{\eta}_j$. This relationship has been well established in the literature for both unidimensional (Chang & Ying, 1996) and multidimensional IRT contexts (Wang et al., 2011). Specifically, for the M2PLM,

$$\boldsymbol{I}(\boldsymbol{\eta})=\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T}\mathrm{KL}(\boldsymbol{\eta}\,\|\,\boldsymbol{\eta}_j). \tag{9}$$

Geometrically, KL information, as a function of the item parameter vector $\boldsymbol{\eta}$, forms a hypersurface over the item parameter space. The Fisher information matrix is precisely the Hessian matrix of this hypersurface evaluated at the true parameter point $\boldsymbol{\eta}_j$, characterizing its local curvature in the neighborhood of $\boldsymbol{\eta}_j$. The unidimensional case makes this geometric interpretation especially intuitive: KL information reduces

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

to a curve in the parameter space, and Fisher information is simply the curvature (second derivative) of that curve at the true value.

This connection follows from a second-order Taylor expansion of the KL information around $\boldsymbol{\eta}_j$, in which the coefficient matrix of the second-order term is exactly the Fisher information matrix. Fisher information therefore captures only the local second-order features of KL information at the true parameter point, while entirely disregarding higher-order terms and the global structure of the parameter space. Consequently, although Fisher information can be fully recovered from KL information through second-order differentiation, the reverse does not hold: the global divergence metric encoded in KL information cannot be reconstructed from Fisher information alone.

This asymmetry carries important implications for online calibration. In the early calibration stage, limited calibration samples cause the provisional item parameter estimates to deviate from their true values. Fisher information computed at these estimates reflects only the local curvature around the incorrect location, potentially yielding misleading item selection decisions. KL information, by quantifying divergence across the entire parameter space, preserves the global structure and is therefore more robust to parameter estimation errors, providing more reliable guidance for item selection. This fundamental distinction justifies characterizing KL information as global and Fisher information as local, regardless of whether the information is defined over the ability space or the item parameter space.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

4.2. Average KL Information Index

In the context of item calibration, Equation (8) quantifies the divergence between the provisional estimate of item *j* and other plausible parameter values. After responses from n_j examinees have been collected, the KL information for new item *j* can be expressed as

$$\mathrm{KL}\left(\boldsymbol{\eta} \,\|\, \hat{\boldsymbol{\eta}}_j^n\right) = P\left(\hat{\boldsymbol{\eta}}_j^n\right)\log\frac{P\left(\hat{\boldsymbol{\eta}}_j^n\right)}{P\left(\boldsymbol{\eta}\right)} + \left(1-P\left(\hat{\boldsymbol{\eta}}_j^n\right)\right)\log\frac{\left(1-P\left(\hat{\boldsymbol{\eta}}_j^n\right)\right)}{\left(1-P\left(\boldsymbol{\eta}\right)\right)}, \tag{10}$$

where $\hat{\boldsymbol{\eta}}_j^n = \left(\hat{a}_{j1}^n, \hat{a}_{j2}^n, \hat{b}_j^n\right)^T$ refers to the parameter estimate vector for the new item *j* after responses from n_j examinees. Since the true values of the item parameters are unknown during calibration, the provisional estimates $\hat{\boldsymbol{\eta}}_j^n$ are used in practice.

Following Chang and Ying (1996), a single indicator is constructed by averaging KL information over an appropriate interval, defined as the average KL information index (KI):

$$\mathrm{KI}\left(\hat{\boldsymbol{\eta}}_j^n\right) = \int_{\boldsymbol{\eta}_v} \ldots \int \mathrm{KL}\left(\boldsymbol{\eta} \,\|\, \hat{\boldsymbol{\eta}}_j^n\right) d\boldsymbol{\eta}, \tag{11}$$

where $\boldsymbol{\eta}_v$ signifies the *v*-th element of the item parameter vector $\boldsymbol{\eta}$ (*v*=1,2,3). The integral is subsequently expanded to:

$$\mathrm{KI}\left(\hat{\boldsymbol{\eta}}_j^n\right) = \int_{\hat{a}_{j1}^n-\delta_n}^{\hat{a}_{j1}^n+\delta_n} \int_{\hat{a}_{j2}^n-\delta_n}^{\hat{a}_{j2}^n+\delta_n} \int_{\hat{b}_j^n-\delta_n}^{\hat{b}_j^n+\delta_n} \mathrm{KL}\left(\boldsymbol{\eta} \,\|\, \hat{\boldsymbol{\eta}}_j^n\right) da_1 da_2 db, \tag{12}$$

where $\delta_n = c/\sqrt{n_j}$ denotes the permissible range of variation for the provisional parameter estimate of new item *j*, a range that narrows progressively as calibration proceeds and the parameter estimates stabilize. The constant *c* is conventionally set to 3 (Chang & Ying, 1996; Wang & Chang, 2009, 2011), corresponding to

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

approximately three standard errors under the asymptotic normal approximation. A detailed justification for this choice is provided in Appendix B.

Although both $\mathrm{KI}\left(\hat{\boldsymbol{\eta}}_j^n\right)$ and $\boldsymbol{I}\left(\hat{\boldsymbol{\eta}}_j^n\right)$ are functions of $\hat{\boldsymbol{\eta}}_j^n$ alone, they capture fundamentally different information: $\mathrm{KI}\left(\hat{\boldsymbol{\eta}}_j^n\right)$ is obtained by integrating $\mathrm{KL}\left(\boldsymbol{\eta} \| \hat{\boldsymbol{\eta}}_j^n\right)$ over a neighborhood of $\hat{\boldsymbol{\eta}}_j^n$, capturing the average information within that domain, whereas $\boldsymbol{I}\left(\hat{\boldsymbol{\eta}}_j^n\right)$ is based solely on point-specific information at $\hat{\boldsymbol{\eta}}_j^n$. This interval-based property makes $\mathrm{KI}\left(\hat{\boldsymbol{\eta}}_j^n\right)$ potentially more informative in the early calibration stage, when provisional estimates remain unstable.

4.3. M-KI and M-PKI Designs

4.3.1. Online Calibration Designs

Maximizing KI optimizes the distinguishability between the provisional item parameter estimates and all other possible values. Selecting the new item with the highest KI value implies that the examinee's ability level provides the greatest expected power to distinguish the provisional parameter estimate from other possible values, and hence offers the greatest potential calibration contribution. We refer to this as the M-KI design, where the new item j assigned to the current examinee satisfies:

$$j = \arg\max_{j \in H} \left\{ \mathrm{KI}\left(\hat{\boldsymbol{\eta}}_j^n\right) \right\}. \tag{13}$$

The M-KI design selects the optimal new item using the KI indicator alone, without incorporating prior parameter information. Building on this design, we further exploit the new item's characteristics by weighting the KI with the posterior information of its parameters. This enhanced indicator, termed the posterior weighted KI (PKI), is defined as

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

$$\mathrm{PKI}\left(\hat{\boldsymbol{\eta}}_j^n\right)=\int_{\hat{a}_{j1}^n-\delta_n}^{\hat{a}_{j1}^n+\delta_n}\int_{\hat{a}_{j2}^n-\delta_n}^{\hat{a}_{j2}^n+\delta_n}\int_{\hat{b}_j^n-\delta_n}^{\hat{b}_j^n+\delta_n}\mathrm{KL}\left(\boldsymbol{\eta}\,\|\,\hat{\boldsymbol{\eta}}_j^n\right)f\left(\boldsymbol{\eta}\,|\,\boldsymbol{u}_{n_j}\right)da_1da_2db, \tag{14}$$

where $f\left(\boldsymbol{\eta}\,|\,\boldsymbol{u}_{n_j}\right)=\dfrac{f\left(\boldsymbol{\eta}\right)L\left(\boldsymbol{u}_{n_j};\boldsymbol{\theta}^{(n_j)},\boldsymbol{\eta}\right)}{\int f\left(\boldsymbol{\eta}\right)L\left(\boldsymbol{u}_{n_j};\boldsymbol{\theta}^{(n_j)},\boldsymbol{\eta}\right)d\boldsymbol{\eta}}$ is the posterior distribution of the item parameters of new item j; $\boldsymbol{u}_{n_j}=\left(u_{1j},u_{2j},...,u_{n_jj}\right)^T$ is the response vector of the n_j examinees who have answered new item j; $\boldsymbol{\theta}^{(n_j)}=\left(\boldsymbol{\theta}_{1j},\boldsymbol{\theta}_{2j},...,\boldsymbol{\theta}_{n_jj}\right)^T$ is the $n_j\times 2$-dimensional ability matrix of those examinees; and $f\left(\boldsymbol{\eta}\right)$ is the prior distribution of the new item's parameters. Although the posterior weight naturally localizes the integration as n_j increases, a shrinking neighborhood constraint is retained to prevent erratic selections in the early calibration stage, when the parameter estimates are most unstable.

When PKI serves as the basis for assigning new items, the resulting design is called the M-PKI design. Accordingly, the new item j assigned to the current examinee satisfies:

$$j=\arg\max_{j\in H}\left\{\mathrm{PKI}\left(\hat{\boldsymbol{\eta}}_j^n\right)\right\}. \tag{15}$$

4.3.2. Approximation of Multiple Integral

The multiple integrals in the M-KI and M-PKI designs are computationally demanding, requiring an appropriate numerical integration method. Monte Carlo (MC) integration offers substantially higher computational efficiency than traditional numerical integration methods in high-dimensional settings (Chen et al., 2017) and works by drawing random samples from a specified distribution and approximating the population expectation by the sample mean. Two MC variants are commonly

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

distinguished: non-adaptive Monte Carlo (NAMC) integration, which draws samples from a fixed distribution, and adaptive Monte Carlo (AMC) integration, which draws samples from a continuously updated personalized distribution. Because the two methods yield comparable accuracy while NAMC requires less computation time (Chen et al., 2017), we adopt NAMC integration and use the distribution fitted from the item bank's parameters as the prior distribution for sampling (Chen, 2017; He & Chen, 2020; van der Linden & Ren, 2015; Wainer & Mislevy, 1990; Zheng, 2014).

The numerical approximation proceeds as follows. Equation (11) can be expressed as

$$\begin{aligned} \mathrm{KI}\left(\hat{\boldsymbol{\eta}}_j^n\right) &= \int_{\boldsymbol{\eta}_v} \cdots \int \mathrm{KL}\left(\boldsymbol{\eta} \,\|\, \hat{\boldsymbol{\eta}}_j^n\right) d\boldsymbol{\eta} \\ &= \int_{\boldsymbol{\eta}_v} \cdots \int \frac{\mathrm{KL}\left(\boldsymbol{\eta} \,\|\, \hat{\boldsymbol{\eta}}_j^n\right)}{f(\boldsymbol{\eta})} f(\boldsymbol{\eta}) d\boldsymbol{\eta}. \end{aligned} \tag{16}$$

KI can thus be viewed as the expectation of $\frac{\mathrm{KL}\left(\boldsymbol{\eta} \,\|\, \hat{\boldsymbol{\eta}}_j^n\right)}{f(\boldsymbol{\eta})}$ with respect to $\boldsymbol{\eta}$ under the prior distribution $f(\boldsymbol{\eta})$. By the law of large numbers, this population expectation can be approximated by the corresponding sample mean:

$$\begin{aligned} &\int_{\boldsymbol{\eta}_v} \cdots \int \frac{\mathrm{KL}\left(\boldsymbol{\eta} \,\|\, \hat{\boldsymbol{\eta}}_j^n\right)}{f(\boldsymbol{\eta})} f(\boldsymbol{\eta}) d\boldsymbol{\eta} \\ &= E_f\left[\frac{\mathrm{KL}\left(\boldsymbol{\eta} \,\|\, \hat{\boldsymbol{\eta}}_j^n\right)}{f(\boldsymbol{\eta})}\right] \\ &= \frac{1}{S}\sum_{s=1}^{S} \frac{\mathrm{KL}\left(\boldsymbol{\eta}^{(s)} \,\|\, \hat{\boldsymbol{\eta}}_j^n\right)}{f(\boldsymbol{\eta}^{(s)})}, \end{aligned} \tag{17}$$

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

where $\boldsymbol{\eta}^{(s)} = \left(a_1^{(s)}, a_2^{(s)}, b^{(s)}\right)\left(s = 1, 2, ..., S\right)$ is the *s*-th sample obtained from $f\left(\boldsymbol{\eta}\right)$ over the interval $[\hat{\boldsymbol{\eta}}_{jv}^{n} - \delta_n, \hat{\boldsymbol{\eta}}_{jv}^{n} + \delta_n]$, and $\hat{\boldsymbol{\eta}}_{jv}^{n}$ represents the *v*-th element of $\hat{\boldsymbol{\eta}}_{j}^{n}$. Similarly, $\mathrm{PKI}\left(\hat{\boldsymbol{\eta}}_{j}^{n}\right)$ can be expressed as

$$
\begin{aligned}
\mathrm{PKI}\left(\hat{\boldsymbol{\eta}}_{j}^{n}\right) &= \int_{\boldsymbol{\eta}_v} ... \int \mathrm{KL}\left(\boldsymbol{\eta} \,||\, \hat{\boldsymbol{\eta}}_{j}^{n}\right) f\left(\boldsymbol{\eta} \,|\, \boldsymbol{u}_{n_j}\right) d\boldsymbol{\eta} \\
&= \int_{\boldsymbol{\eta}_v} ... \int \mathrm{KL}\left(\boldsymbol{\eta} \,||\, \hat{\boldsymbol{\eta}}_{j}^{n}\right) \frac{L\left(\boldsymbol{u}_{n_j}; \boldsymbol{\theta}^{(n_j)}, \boldsymbol{\eta}\right) f\left(\boldsymbol{\eta}\right)}{\int L\left(\boldsymbol{u}_{n_j}; \boldsymbol{\theta}^{(n_j)}, \boldsymbol{\eta}\right) f\left(\boldsymbol{\eta}\right) d\boldsymbol{\eta}} d\boldsymbol{\eta} \\
&= \frac{\int_{\boldsymbol{\eta}_v} ... \int \mathrm{KL}\left(\boldsymbol{\eta} \,||\, \hat{\boldsymbol{\eta}}_{j}^{n}\right) L\left(\boldsymbol{u}_{n_j}; \boldsymbol{\theta}^{(n_j)}, \boldsymbol{\eta}\right) f\left(\boldsymbol{\eta}\right) d\boldsymbol{\eta}}{\int_{\boldsymbol{\eta}_v} ... \int \int L\left(\boldsymbol{u}_{n_j}; \boldsymbol{\theta}^{(n_j)}, \boldsymbol{\eta}\right) f\left(\boldsymbol{\eta}\right) d\boldsymbol{\eta} d\boldsymbol{\eta}} \\
&= \frac{E_f\left[\mathrm{KL}\left(\boldsymbol{\eta} \,||\, \hat{\boldsymbol{\eta}}_{j}^{n}\right) L\left(\boldsymbol{u}_{n_j}; \boldsymbol{\theta}^{(n_j)}, \boldsymbol{\eta}\right)\right]}{E_f\left[L\left(\boldsymbol{u}_{n_j}; \boldsymbol{\theta}^{(n_j)}, \boldsymbol{\eta}\right)\right]} \\
&= \frac{\frac{1}{S_1}\sum_{s_1=1}^{S_1} \mathrm{KL}\left(\boldsymbol{\eta}^{(s_1)} \,||\, \hat{\boldsymbol{\eta}}_{j}^{n}\right) L\left(\boldsymbol{u}_n; \boldsymbol{\theta}_n, \boldsymbol{\eta}^{(s_1)}\right)}{\frac{1}{S_2}\sum_{s_2=1}^{S_2} L\left(\boldsymbol{u}_n; \boldsymbol{\theta}_n, \boldsymbol{\eta}^{(s_2)}\right)},
\end{aligned}
\tag{18}
$$

where $\boldsymbol{\eta}^{(s_1)} = \left(a_1^{(s_1)}, a_2^{(s_1)}, b^{(s_1)}\right)\left(s_1 = 1, 2, ..., S_1\right)$ is the s_1-th sample drawn from $f\left(\boldsymbol{\eta}\right)$ over the interval $[\hat{\boldsymbol{\eta}}_{jv}^{n} - \delta_n, \hat{\boldsymbol{\eta}}_{jv}^{n} + \delta_n]$; $\boldsymbol{\eta}^{(s_2)} = \left(a_1^{(s_2)}, a_2^{(s_2)}, b^{(s_2)}\right)\left(s_2 = 1, 2, ..., S_2\right)$ is the s_2-th sample drawn from $f\left(\boldsymbol{\eta}\right)$.

4.3.3. Selection of Samples for MC Integration

To determine the optimal sample size for NAMC integration, we conducted a rigorous convergence and computational-cost analysis for the M-KI and M-PKI integrals, systematically testing 13 sample size gradients ranging from 100 to 100,000 (i.e., 100, 500, 1,000, 2,000 to 10,000 in steps of 2,000, and 20,000 to 100,000 in steps of 20,000).

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Furthermore, because the optimal sample size depends heavily on the geometric shape of the integrand in the multidimensional space, we avoided arbitrary item parameters and instead constructed three representative scenarios to serve as boundary stress tests for the integration algorithm:

Sit. 1 (typical matched scenario, $\boldsymbol{\theta} = (0,0)^T$, $\boldsymbol{\eta} = (1,1,0)^T$): the baseline case, in which the integrand exhibits an ideal, centered, symmetric bell shape.

Sit. 2 (extreme mismatched scenario, $\boldsymbol{\theta} = (2,2)^T$, $\boldsymbol{\eta} = (1.5,1.5,-2)^T$): designed to capture tail-shift behaviors, this mismatch shifts the information volume away from the origin, to verify whether the integration grid is broad enough to prevent boundary truncation errors.

Sit. 3 (structural imbalance scenario, $\boldsymbol{\theta} = (1.5,1.5)^T$, $\boldsymbol{\eta} = (2.5,0.2,0)^T$): a severe geometric distortion test in which disproportionate discrimination parameters compress the integrand into a narrow "ridge", requiring sufficiently high sample density to capture sharp peaks.

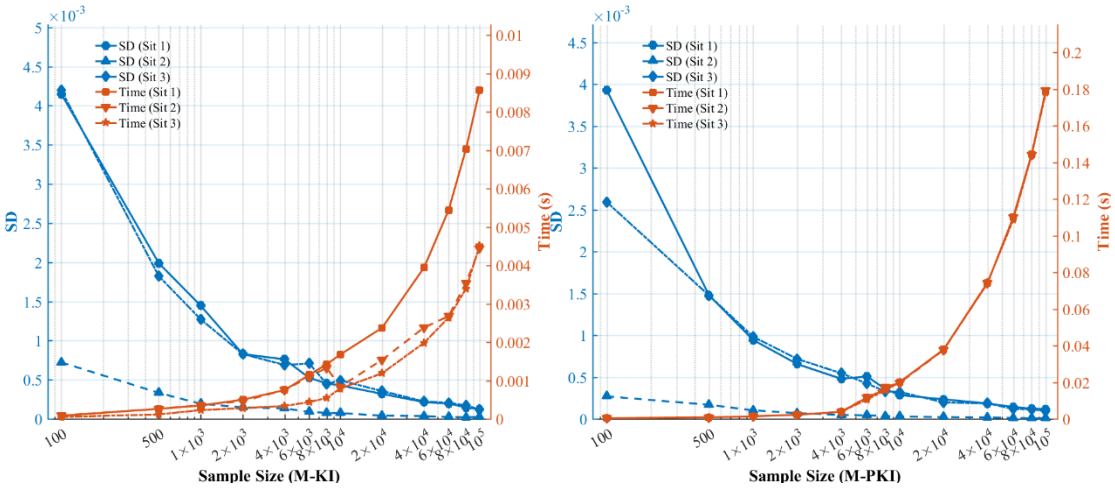
By evaluating the convergence stability across these three scenarios, we identified a sample size that offers a reasonable trade-off between numerical stability and computational cost. Approximate integral values for both KI and PKI were computed across all 39 conditions (13 sample sizes $\times$ 3 scenarios), with each specific condition independently replicated 50 times. Finally, the mean approximate integral, the standard deviation of the estimates, and the average running time per calibration were meticulously recorded and analyzed for each condition.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Table 1. Integral values of KI and PKI

Sample	KI			PKI		
	Sit.1	Sit.2	Sit.3	Sit.1	Sit.2	Sit.3
100	0.044162	0.004789	0.030937	0.019823	0.000843	0.009417
500	0.045068	0.004757	0.031085	0.018878	0.000882	0.010506
1000	0.045132	0.004845	0.031393	0.018906	0.000858	0.010470
2000	0.045086	0.004818	0.031259	0.018952	0.000861	0.010196
10000	0.045170	0.004810	0.031436	0.018864	0.000849	0.010241
100000	0.045181	0.004818	0.031457	0.018905	0.000854	0.010262

Table 1 shows that with a sample size of 1,000, both indicators had already converged well, exhibiting minimal further variation. Figure 1 further reveals that as sample size increased, the standard deviation of the integral decreased monotonically while running time increased monotonically. For the KI indicator, the SD curve exhibited a clear elbow at a sample size of 2,000, where the integral error decreased substantially and leveled off, whereas the running time increased markedly beyond 4,000. These results empirically suggest that a sample size of 2,000 offers a practical trade-off between calibration precision and computational cost, consistent with the sample size used by Chen and Wang (2016) and Chen et al. (2017). We therefore set the MC sample size for KI at $S = 2{,}000$.



Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Figure 1. Standard deviation (SD) and average computation time (Time) across Monte Carlo sample sizes for the M-KI and M-PKI designs.

For the PKI indicator, the added computational burden of posterior weighting made its running time roughly 20 times that of KI and remained nearly constant across the three scenarios. The incorporation of posterior information, however, accelerated convergence within the effective region; with a sample size of 1,000, the SD of PKI had already declined substantially and stabilized. Given the stringent real-time demands of CAT, further increasing the sample size for PKI (e.g., to 2,000 or higher) would introduce non-negligible computational delays without a meaningful gain in calibration precision. We therefore set the integration sample size for PKI at 1,000 (i.e., S_1=S_2=1,000), consistent with He and Chen (2020).

4.4. M-SKI1 and M-SKI2 Designs

Both KI-based designs, namely the M-KI and M-PKI designs, involve multiple integrations, which in practical applications are computationally intensive and time-consuming. To overcome this limitation, we approximated the KI indicator and constructed a simplified indicator. Following Wang et al. (2011), who related the ability-space KI to item parameters to achieve simplification, we analogously simplified the item-parameter-space KI by deriving its dependence on examinee ability, with respect to both the discrimination (a) and intercept (b) parameters. The detailed derivation is provided in Appendix A.

The analysis of the relationship between KI and ability parameters yields two findings: (1) the component of KI associated with the discrimination parameter, denoted KI_a, is proportional to $\theta_1^2+\theta_2^2$; (2) the component of KI associated with the intercept parameter, denoted KI_b, reaches its maximum when $\boldsymbol{a}^T\boldsymbol{\theta}-b=0$.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Maximizing KI_a and KI_b simultaneously is equivalent to maximizing KI. This insight allows us to develop a new indicator $\frac{1}{|\boldsymbol{a}^T\boldsymbol{\theta}-b|}\left(\boldsymbol{\theta}^T\boldsymbol{\theta}\right)$, termed SKI1. For the current examinee answering new item j, SKI1 is

$$\mathrm{SKI1}_j = \frac{1}{|\boldsymbol{a}_j^T\boldsymbol{\theta}-b_j|}\left(\boldsymbol{\theta}^T\boldsymbol{\theta}\right), \tag{19}$$

where $\boldsymbol{\theta} = \left(\theta_1, \theta_2\right)^T$ represents the current examinee's ability.

The new item with the highest SKI1 value is selected for the examinee; we refer to this as the M-SKI1 design, in which new item j is assigned to the current examinee:

$$j = \arg\max_{j\in H}\{\mathrm{SKI1}_j\}. \tag{20}$$

Furthermore, simulation studies by Chen et al. (2017) and Chen (2017) found that, under MCAT, calibration precision decreases as the positive correlation among ability dimensions increases. From a multivariate statistical perspective, this phenomenon can be attributed to multicollinearity: high collinearity among ability estimates causes the information provided by the ability vectors to become linearly dependent and redundant, thereby weakening the identifiability of the multidimensional item parameters. The determinant of the correlation matrix across ability dimensions offers a natural way to quantify this non-redundancy of information, paralleling the logic of D-optimal experimental design, which maximizes the determinant of the information matrix to favor configurations that provide non-redundant information. For two ability dimensions with correlation ρ, this determinant reduces to $1-\rho^2$: it attains its maximum of 1 when the two dimensions are orthogonal $\left(\rho = 0\right)$, reflecting maximal non-redundant information, and approaches 0 as collinearity becomes severe $\left(|\rho| \to 1\right)$, reflecting increasingly

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

redundant information. Building on this idea, we augment the SKI1 criterion with a multiplicative weight based on the determinant of the sample correlation matrix among ability dimensions, yielding a new indicator, SKI2, for item *j*:

$$\text{SKI2}_j = \frac{1-\rho^2}{|\boldsymbol{a}_j^T\boldsymbol{\theta} - b_j|}\left(\boldsymbol{\theta}^T\boldsymbol{\theta}\right), \tag{21}$$

where $\rho = \rho\left(\boldsymbol{\theta}^{(n_j+1)}\right)$ denotes the sample correlation between the two ability dimensions, computed from the $\left(n_j+1\right)\times 2$ matrix $\boldsymbol{\theta}^{(n_j+1)}$ of the n_j examinees who responded to new item *j* along with the current examinee.

The item with the highest SKI2 value is selected for the current examinee. This design is referred to as the M-SKI2 design, where new item *j* assigned to the current examinee satisfies:

$$j = \arg\max_{j\in H}\{\text{SKI2}_j\}. \tag{22}$$

5. Simulation Study 1

A simulation study was conducted to evaluate the calibration performance of the proposed KL information criteria (i.e., the M-KI, M-PKI, M-SKI1, and M-SKI2 designs) and to assess whether they outperform the Fisher-information-based M-D design, particularly in the early calibration stage. Moreover, because item parameters at extreme levels are inherently difficult to estimate, the limitations of Fisher information criteria are likely to be more pronounced for such items. Therefore, the calibration performance of each design for items with extreme parameters was also evaluated.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

5.1. Examinee and Item Generation

Examinee. The true ability values of the examinees were generated from a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (0,0)^T$ and a covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \operatorname{cov}(\boldsymbol{\theta}) \\ & 1 \end{pmatrix}$. Since the variances of the two-dimensional examinee abilities are 1, their covariance equals the correlation ρ_{θ_1,θ_2}. To make the simulated ability vectors more realistic, following Chen and Wang (2016), the correlation between abilities was generated from $U(0.4, 0.8)$, indicating a moderate to high correlation between the two dimensions.

Item bank. In the simulation study, 1,000 operational items were generated to construct the item bank, where the intercept parameters *b* were generated from a standard normal distribution, and the discrimination parameters were randomly drawn from the truncated multivariate normal distribution $\left(a_i \in (lb_a, ub_a), i = 1,2\right)$, satisfying:

$$\boldsymbol{\mu}_a = \left(\mu_{a_1}, \mu_{a_2}\right), \boldsymbol{\Sigma}_a = \begin{pmatrix} Var(a_1) & Cov(a_1, a_2) \\ & Var(a_2) \end{pmatrix}. \tag{23}$$

The parameters of this distribution were determined following the conventional approach in UIRT, where the discrimination parameter *a* is typically assumed to follow a log-normal distribution truncated to a finite interval. According to Lien (1985), the *r*-th moment of the truncated log-normal distribution is

$$E(a^r \mid lb_a < a < ub_a) = E(a^r) \frac{F\left[r\sigma - \dfrac{\log(lb_a) - u}{\sigma}\right] - F\left[r\sigma - \dfrac{\log(ub_a) - u}{\sigma}\right]}{F\left[\dfrac{\log(ub_a) - u}{\sigma}\right] - F\left[\dfrac{\log(lb_a) - u}{\sigma}\right]}, \tag{24}$$

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

where $E(a^r) = \exp(ru + (r^2\sigma^2/2))$, and F is the cumulative distribution function of the standard normal distribution. Setting $u = 0,\ \sigma = 1,\ lb_a = 0.1,\ ub_a = 2.3$, we obtained $E\left(a \mid 0.1 < a < 2.3\right) = 0.908,\ \ \mathrm{Var}\left(a \mid 0.1 < a < 2.3\right) = 0.318$ (Chen & Wang, 2016; Chen, 2017). Thus, we set $\mu_{a_1} = \mu_{a_2} = 0.908$, and $\mathrm{Var}\left(a_1\right) = \mathrm{Var}\left(a_2\right) = 0.318$. The two-dimensional discrimination parameters were randomly generated as follows:

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \sim \left(\begin{pmatrix} 0.908 \\ 0.908 \end{pmatrix},\ \begin{pmatrix} 0.318 & Cov(a_1, a_2) \\ Cov(a_2, a_1) & 0.318 \end{pmatrix} \right), \tag{25}$$

where $Cov(a_1, a_2) = \rho_{a_1,a_2} \times \sqrt{\mathrm{var}\left(a_1\right)} \times \sqrt{\mathrm{var}\left(a_2\right)}$, and R_{a_1,a_2} was generated from $U\left(0.2, 0.3\right)$.

New item. Two sets of new items were generated: the typical set and the extreme set. The typical set comprised 20 items simulated using the same method as the item bank. The extreme set comprised another 20 items generated by truncating the parameter distributions based on the same method, with discrimination and intercept parameters drawn from the extreme tails of their respective distributions. Specifically, discrimination parameters were drawn from [0.1, 0.5] (representing flawed or highly ambiguous items with near-zero Fisher information) and [1.5, 2.3] (representing hypersensitive items with narrow, ridge-like information functions). Intercept parameters were drawn from [-3.0, -1.2] and [1.2, 3.0], representing items whose response probabilities were located in the extreme tails of the ability distribution. This extreme item set was used to evaluate the calibration performance of the different designs under extreme conditions.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

5.2. MCAT Simulation and Online Calibration Procedure

During the adaptive testing process, the initial ability estimate vector of each examinee was set to the zero vector $(0,0)^T$. The D-optimality item selection criterion (Segall, 1996, 2001) was employed as a classical method to select operational items from the item bank for examinees. For each item administered to an examinee, the probability of a correct response was first computed using the true item parameters and the examinee's true ability vector via Equation (1), and then this probability was compared to a pseudo-random uniform number to simulate the observed response. Based on the examinee's responses to the operational items, the ability was then estimated. Specifically, a combination of expected a posteriori (EAP) and maximum likelihood estimation (MLE) was used (He & Chen, 2020; He et al., 2020). When the number of operational items answered by an examinee was fewer than five, the ability vector was updated using EAP; otherwise, MLE was applied. The EAP estimate was computed using the two-dimensional Gauss-Hermite quadrature (Stroud & Secrest, 1966), while the MLE estimate was obtained by transforming the system of equations into an optimization problem solved via particle swarm optimization (PSO; Kennedy & Eberhart, 1995).

Each examinee *i* responded to a total of $(20+d_i)$ items in the test, consisting of 20 operational items and d_i new items. The value of d_i was determined by the number of remaining uncalibrated new items: if more than five new items remained, $d_i=5$; otherwise, equaled the remaining number. The seeding location of new items in the test is critical, as ability estimates must be obtained from the operational items before new items are administered. To ensure reasonably accurate ability estimates for new item calibration while remaining unobtrusive to examinees, following the

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

recommendation of van der Linden and Ren (2015) and He and Chen (2020), the d_i new items for examinee *i* were randomly seeded into positions selected from the last (d_i+3) positions of the $(20+d_i)$ items.

The entire online calibration was conducted in a dynamic examinee pool. Before a new item was assigned to an examinee using an adaptive design, initial estimates of the item parameters needed to be obtained. To ensure a fair comparison of the performance across different designs, these initial estimates were obtained by adding normally distributed errors to the true item parameters, i.e., $\hat{\boldsymbol{\eta}}_{jv}^{(0)} = \boldsymbol{\eta}_{jv} + \varepsilon$, where $\varepsilon \sim N\left(0, \sigma^2\right)$. Specifically, three error levels were considered: $\sigma = 0, 0.1, \text{and } 0.5$. While this approach ensures fairness across designs, it introduces a practical issue for the M-SKI2 design. When assigning examinees to items with very few or no prior responses, the sample correlation between ability dimensions is estimated from an extremely small sample, making it either undefined for a single response or highly unstable for very small samples. To avoid such anomalies at the very beginning of calibration, we adopted a cold-start strategy for the M-SKI2 design: for each new item, the first five examinees were assigned using the M-SKI1 design, and the M-SKI2 design was activated thereafter. Additionally, a small regularization constant λ=0.01 was introduced into the SKI2 criterion (i.e., replacing $1-\rho^2$ with $1-\rho^2+\lambda$) to ensure numerical stability when the estimated correlation is close to ±1. Together, the cold-start strategy and the regularization constant ensure that the M-SKI2 design operates reliably throughout the calibration process.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

The cold-start strategy is specific to the simulation setting, where initial item parameter estimates were obtained directly under artificially imposed error levels. In practice, most online calibration studies adopt a two-stage procedure: new items are first randomly assigned to a small number of examinees to obtain initial parameter estimates, after which adaptive designs are applied (e.g., He & Qi, 2025; He et al., 2021; Song & Wang, 2025). Under such a procedure, the cold-start strategy becomes unnecessary, as each new item already has a sufficient number of responses before the M-SKI2 design is activated.

After collecting examinees' responses to the new items, M-Method A was used to update the item parameter vectors every time each new item received 30 additional responses. M-Method A is a simple and efficient approach that avoids the computational instability caused by high-dimensional integration and multiple EM cycles (Chen et al., 2017). Specifically, M-Method A first obtained ability estimates from responses to the operational items using MLE and treated them as the true values. Then, based on these ability estimates and the examinees' responses to the new items, it applied MLE again to obtain the item parameter estimates.

Furthermore, each new item was retired from the calibration process upon receiving 100 responses, which marked its calibration completion. The entire calibration process ended when all 20 new items had been calibrated. To reduce random experimental error, all experimental conditions were replicated 100 times.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

5.3. Evaluation Criteria

5.3.1. Calibration Precision

For the online calibration designs under different experimental conditions, bias and RMSE were used to evaluate calibration precision. The equations are as follows:

$$\text{Bias} = \frac{1}{rep}\frac{1}{m}\sum_{r=1}^{rep}\sum_{j=1}^{m}\left(\boldsymbol{\eta}_{rjv} - \hat{\boldsymbol{\eta}}_{rjv}\right) \tag{26}$$

$$\text{RMSE} = \frac{1}{rep}\sum_{r=1}^{rep}\sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(\boldsymbol{\eta}_{rjv} - \hat{\boldsymbol{\eta}}_{rjv}\right)^2} \tag{27}$$

where $\boldsymbol{\eta}_{jv}$ is the *v*-th element of $\boldsymbol{\eta}_j$. *rep* denotes the number of simulation replications (*rep*=100), while *m* represents the number of new items (*m*=20). Values of Bias and RMSE closer to zero indicate higher estimation accuracy of the item parameters, meaning better performance of the online calibration designs. In addition, the standard deviation of RMSE across the 100 replications was calculated to reflect the stability of the estimates.

To assess the linear relationship between the true and estimated item parameter vectors, the Pearson correlation coefficient was used. A larger correlation indicates more accurate item parameter recovery.

Another metric, the weighted mean squared error (WMSE), was computed to reflect the overall calibration precision. WMSE is defined as the weighted mean squared error between the estimated item characteristic curve (ICC) and the true ICC, given by:

$$\text{WMSE} = \frac{1}{rep}\frac{1}{m}\sum_{r=1}^{rep}\sum_{j=1}^{m}\int\left[P(\boldsymbol{\theta};\boldsymbol{\eta}_{jr}) - P(\boldsymbol{\theta};\hat{\boldsymbol{\eta}}_{jr})\right]^2 \omega(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{28}$$

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

where $\omega(\boldsymbol{\theta})$ is a weight function based on the prior distribution of $\boldsymbol{\theta}$, and denotes the bivariate normal density used to generate examinee abilities. The WMSE was also calculated using NAMC integration. A lower WMSE indicates better overall parameter recovery.

The recovery of the correlation between discrimination parameters was also evaluated using quantile-quantile (Q-Q) plots. A Q-Q plot compares two probability distributions by plotting their quantiles against each other; if the distributions are identical, the points approximately lie on the line $y=x$, whereas deviations indicate differences in location, scale, or shape. In this study, Q-Q plots were used to visually assess the recovery of the discrimination correlation across quantile levels, complementing scalar summary indices by showing where recovery deteriorated.

5.3.2. Calibration Information

To further evaluate the performance of the various designs, the average calibration Fisher information (INF) was additionally used to evaluate design performance. Unlike calibration precision, which is tied to the estimation method, INF reflects the information gain from the assigned examinees based solely on the selection design. It thus serves as a complementary metric independent of the subsequent estimation routine, providing a more intuitive reflection of the contribution of examinees to new item calibration under different designs. The INF is defined as:

$$\mathrm{INF} = \frac{1}{rep}\frac{1}{m}\sum_{r=1}^{rep}\sum_{j=1}^{m} det[\sum_{i=1}^{n_j} \boldsymbol{I}(\boldsymbol{\eta}_{jr};\boldsymbol{\theta}_{ir})] \,. \tag{29}$$

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

5.3.3. Running Time

To compare the improvement in computational efficiency of the two simplified designs (M-SKI1 and M-SKI2), the average running time T_{run} spent in assigning each new item to each examinee was recorded. The equation is as follows:

$$T_{run} = \frac{1}{rep}\frac{1}{n}\frac{1}{m}\sum_{r=1}^{rep}\sum_{i=1}^{n}\sum_{j=1}^{m}T_{ijr}, \tag{30}$$

where n is the total number of examinees, and T_{ij} is the running time required to assign new item j to examinee i.

5.4. Results

5.4.1. Results of Typical Set

Calibration precision and efficiency. Table 2 presents the parameter recovery accuracy and computational cost of five calibration designs (M-D, M-KI, M-PKI, M-SKI1, and M-SKI2) under three initial item parameter estimation errors. Overall, as the error increased, the calibration precision of each design showed a slight decline, reflected in a minor increase in WMSE. Moreover, all designs were approximately unbiased under the simulated conditions, with bias values remaining within ±0.05.

In terms of global recovery measured by WMSE, the M-PKI design performed best under all conditions. The WMSE values of the M-SKI1 and M-SKI2 designs were generally higher than those of the other designs, indicating weaker global recovery ability. Regarding specific parameter types, the M-KI and M-PKI designs held an advantage in recovering the discrimination parameter a, while the M-D design outperformed the other designs in recovering the intercept parameter b. In terms of

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

numerical stability, the standard deviation of RMSE across 100 replications (SD_R) showed that the M-KI and M-PKI designs had the smallest fluctuations, while the M-SKI1 design exhibited the poorest stability. Regarding computational cost, the running time of all designs was less than 0.05 seconds per run; the M-D design was the most efficient, and the M-SKI1 and M-SKI2 designs achieved substantial computational savings compared with the M-KI and M-PKI designs.

Concerning the simplified designs (M-SKI1 and M-SKI2), the M-SKI2 design, which explicitly incorporates the correlation between ability dimensions, outperformed the M-SKI1 design on nearly all parameter recovery metrics, confirming the positive role of prior correlation in multidimensional approximation. Compared with the M-KI and M-PKI designs, the M-SKI1 and M-SKI2 designs compromised somewhat in precision but substantially reduced computational

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Table 2. Calibration precision of five designs under different initial item parameter estimation errors

Error	Design	Bias		RMSE		Corr(η)		SD_R		WMSE	T_{run}(s)
		ave_a	*b*	*ave_a*	*b*	*ave_a*	*b*	*ave_a*	*b*		
$\sigma = 0$	M-D	-0.0412	0.0117	0.3686	**0.3122**	0.7214	**0.9551**	0.0785	0.0653	0.00533	**0.00094**
	M-KI	-0.0199	0.0066	0.3565	0.3260	0.7341	0.9494	**0.0666**	0.0655	0.00483	0.01023
	M-PKI	-0.0265	0.0126	**0.3550**	0.3219	**0.7424**	0.9542	0.0698	**0.0584**	**0.00477**	0.04048
	M-SKI1	0.0101	0.0345	0.4030	0.3725	0.6983	0.9369	0.0853	0.0821	0.00556	0.00195
	M-SKI2	-0.0101	0.0043	0.3772	0.3267	0.7039	0.9507	0.0767	0.0658	0.00535	0.00249
$\sigma = 0.1$	M-D	-0.0451	-0.0096	0.3588	**0.3019**	0.7327	**0.9590**	0.0685	0.0644	0.00497	**0.00083**
	M-KI	-0.0120	0.0171	0.3601	0.3268	0.7353	0.9521	0.0636	0.0733	0.00488	0.00935
	M-PKI	-0.0363	0.0163	**0.3449**	0.3185	**0.7472**	0.9538	**0.0605**	**0.0623**	**0.00484**	0.04314
	M-SKI1	0.0099	0.0353	0.4036	0.3767	0.6869	0.9352	0.0836	0.0846	0.00555	0.00179
	M-SKI2	-0.0149	0.0055	0.3836	0.3442	0.7041	0.9447	0.0808	0.0735	0.00522	0.00235
$\sigma = 0.5$	M-D	-0.0502	-0.0030	0.3623	**0.2968**	0.7223	**0.9569**	0.0666	**0.0551**	0.00513	**0.00084**
	M-KI	-0.0276	-0.0096	**0.3617**	0.3312	0.7155	0.9496	0.0664	0.0614	0.00506	0.00976
	M-PKI	-0.0337	0.0181	0.3639	0.3124	**0.7270**	0.9546	**0.0658**	0.0665	**0.00504**	0.03800
	M-SKI1	0.0021	0.0220	0.3974	0.3551	0.6988	0.9414	0.0795	0.0816	0.00548	0.00179
	M-SKI2	-0.0115	-0.0056	0.3762	0.3326	0.7196	0.9499	0.0780	0.0707	0.00511	0.00250

Notes. (1) Corr(η) represents the Pearson correlation coefficient between the estimated item parameters and the true values, and SD_R denotes the standard deviation of RMSE across 100 replications; (2) *ave_a* represents the average of the Bias, RMSE, Corr(η) or SD_R values on two discrimination parameters; (3) because the values of WMSE and T_{run} are relatively small, we report five significant digits rather than four significant digits to reflect differences, and omitting five decimal places will not affect the conclusion.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

complexity. Thus, they still offer practical value in scenarios with limited computational resources or when fast and reliable estimates are needed. In summary, if precise estimation of discrimination parameters is the primary concern, the M-KI or M-PKI design is recommended; if the recovery of intercept parameters is of greater concern, the M-D design is the best choice; and the M-SKI1 and M-SKI2 designs offer an alternative trade-off between precision and efficiency.

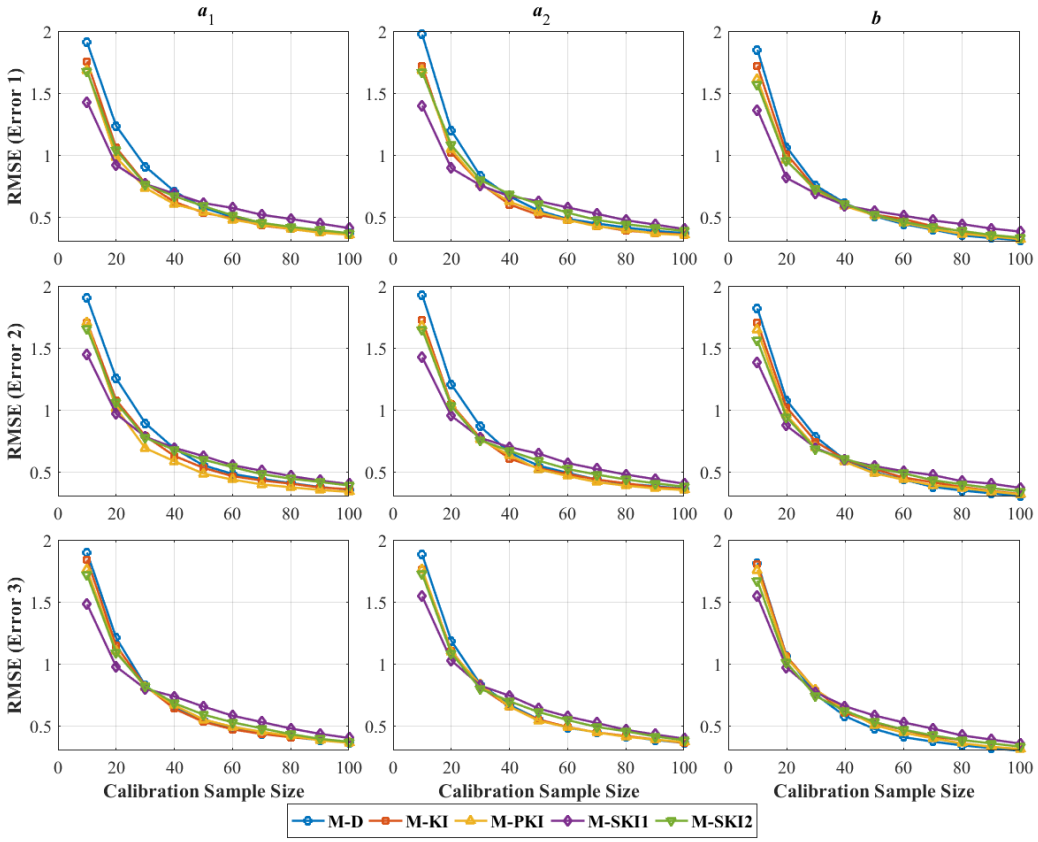


Figure 2. Trend of RMSE. Note. Error 1, Error 2, and Error 3 represent the initial item parameter estimation errors of 0, 0.1, and 0.5, respectively.

Figures 2 and 3 illustrate the trends of RMSE and Bias, respectively, as functions of the calibration sample size under three levels of initial item parameter estimation error. Overall, as the calibration sample size increased, the RMSE and Bias values for all designs gradually decreased across all conditions, while the three error levels had little effect on either the initial or final calibration performance.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Under all conditions, the RMSE values for both the discrimination and intercept parameters of the M-D design were higher than those of the four KL information criteria in the early stage of calibration (especially when the calibration sample size was less than 40). When the sample size reached 40, the curves gradually converged, and the RMSE values across different designs became quite close. Notably, the M-SKI1 design exhibited the lowest initial RMSE values under almost all conditions; however, once the sample size exceeded 30, its RMSE declined much more slowly than that of the other designs.

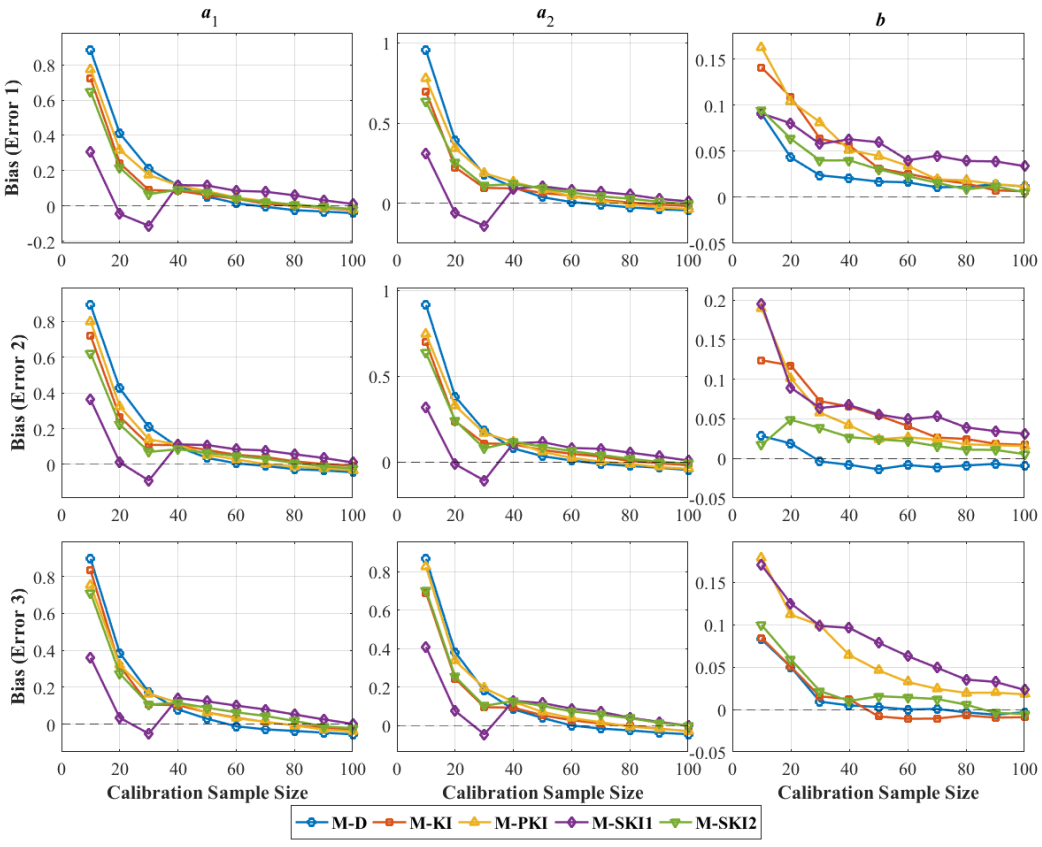


Figure 3. Trend of Bias. Note. Error 1, Error 2, and Error 3 represent the initial item parameter estimation errors of 0, 0.1 and 0.5, respectively.

The bias of the discrimination parameter *a* generally decreased and approached zero, whereas the bias of the intercept parameter *b* showed no clear trend but also gradually approached zero. The bias of the discrimination parameter under the M-D

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

design remained noticeably higher than that of the other designs in the early calibration stage. The M-SKI1 design maintained an absolute advantage in recovering the discrimination parameter but performed the worst on the intercept parameter.

Taken together, the results in Figures 2 and 3 confirm the advantage of the KL information criteria in the early stage of calibration—an advantage that is retained by the M-SKI1 and M-SKI2 designs while substantially reducing running time.

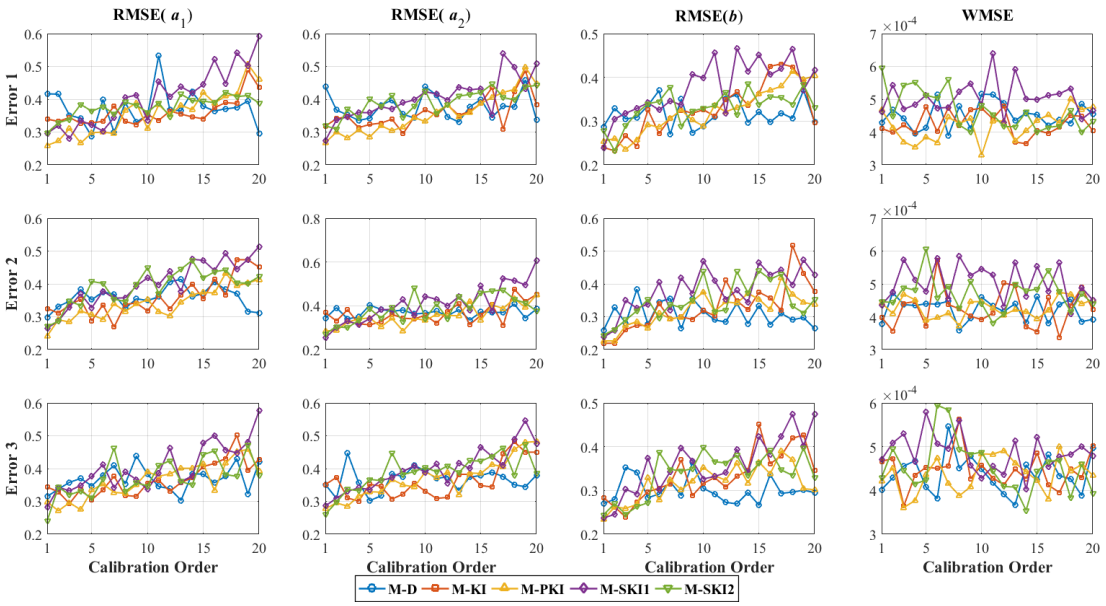


Figure 4. RMSE and WMSE values averaged across 100 replications as a function of item completion order for each new item under three initial item parameter estimation errors. Note. Error 1, Error 2, and Error 3 represent the initial item parameter estimation errors of 0, 0.1, and 0.5, respectively.

In addition, calibration precision can be influenced by multiple factors. On the one hand, item characteristics play a key role. Given a limited number of examinees, items at extreme parameter levels tend to have fewer well-matched examinees, thereby receiving less information and consequently achieving slightly poorer final calibration results. On the other hand, the sequential nature of the dynamic examinee pool also plays a role. Due to the influence of examinee response order, some items may fail to encounter suitable examinees in a timely manner and are thus delayed

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

until later stages of calibration, where they must be completed by the remaining examinees. This delayed completion is associated with slightly poorer parameter recovery. As shown in Figure 4, almost all designs exhibited an upward trend in RMSE for individual parameters as the calibration order progressed, indicating a decrease in calibration precision for items calibrated later. This upward trend was more pronounced for the M-KI and M-PKI designs. However, for the overall recovery precision of item parameters (WMSE), no clear trend was observed, suggesting that calibration order has a limited impact on overall parameter recovery.

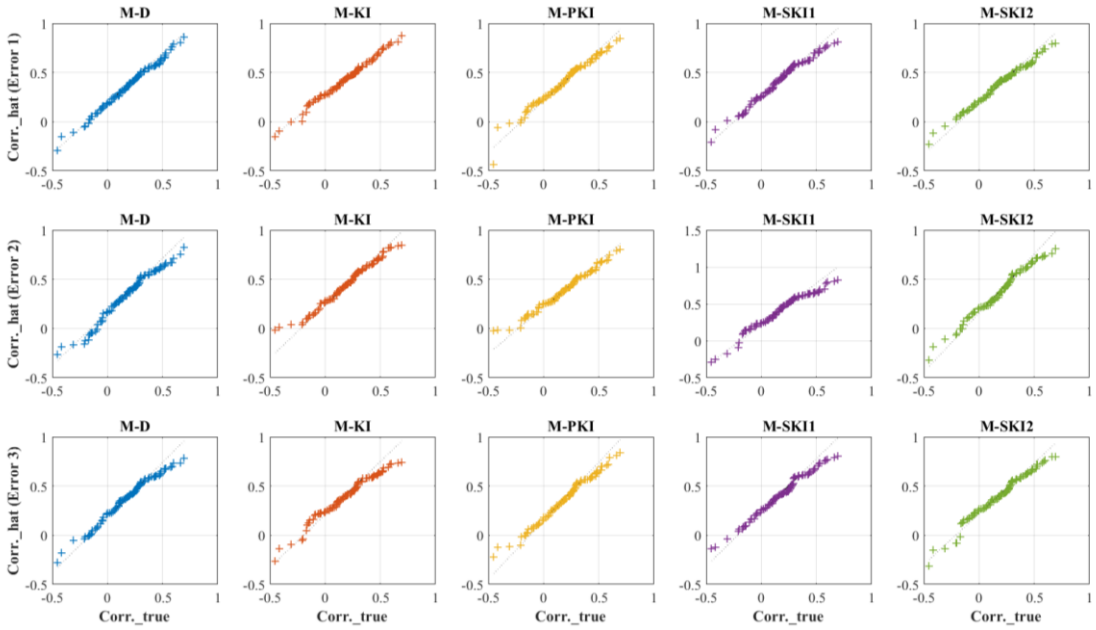


Figure 5. True discrimination correlation and estimated discrimination correlation under three initial item parameter estimation errors. Notes. (1) Error 1, Error 2, and Error 3 represent the initial item parameter estimation errors of 0, 0.1, and 0.5, respectively. (2) *Corr._true* represents true discrimination correlation, and *Corr._hat* represents estimated discrimination correlation.

To assess the similarity between the true and estimated correlations of the two-dimensional discrimination parameters, Q-Q plots were constructed under three levels of initial item parameter estimation error (Figure 5). Across all designs, the recovery of extreme correlations (low or high) was considerably poorer than that of moderate

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

correlations, as shown by larger deviations at both ends of the scatterplot. Among the designs, the M-PKI design achieved the best overall recovery, followed by the M-D design, which performed worse than M-PKI in recovering high correlations.

Calibration information. Figures 6-8 present the average information obtained by new items as a function of the number of responding examinees, under three levels of initial item parameter estimation error. Overall, the KL information criteria consistently outperformed the M-D design in terms of information gain. Among the KL information criteria, the M-PKI design yielded the highest amount of information, followed by M-KI design, whereas the M-D design produced the least. The amount of information directly reflects the effective contribution of the selected examinees in the matching process, serving as a useful indicator of examinee resource utilization efficiency.

The retirement pattern of individual new items provides a complementary perspective. Figure 9 presents a scatterplot of the information (INF) obtained for each new item at its retirement against the cumulative total number of examinees (i.e., the total number of examinees who had been tested up to the retirement of that item) under three initial item parameter estimation errors. Each point represents one new item.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

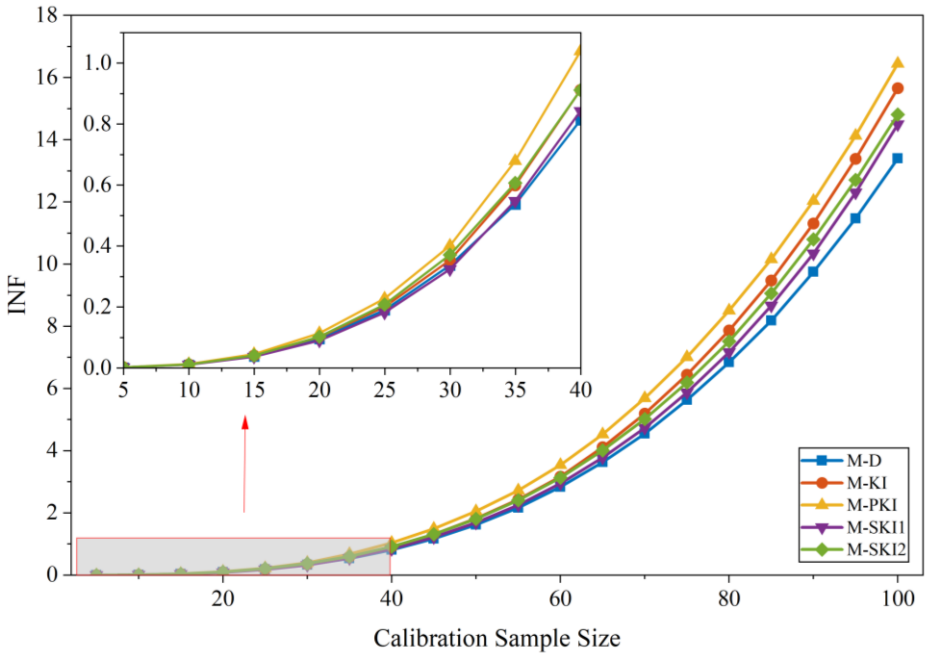


Figure 6. Calibration information as a function of the number of responding examinees under an initial item parameter estimation error of 0

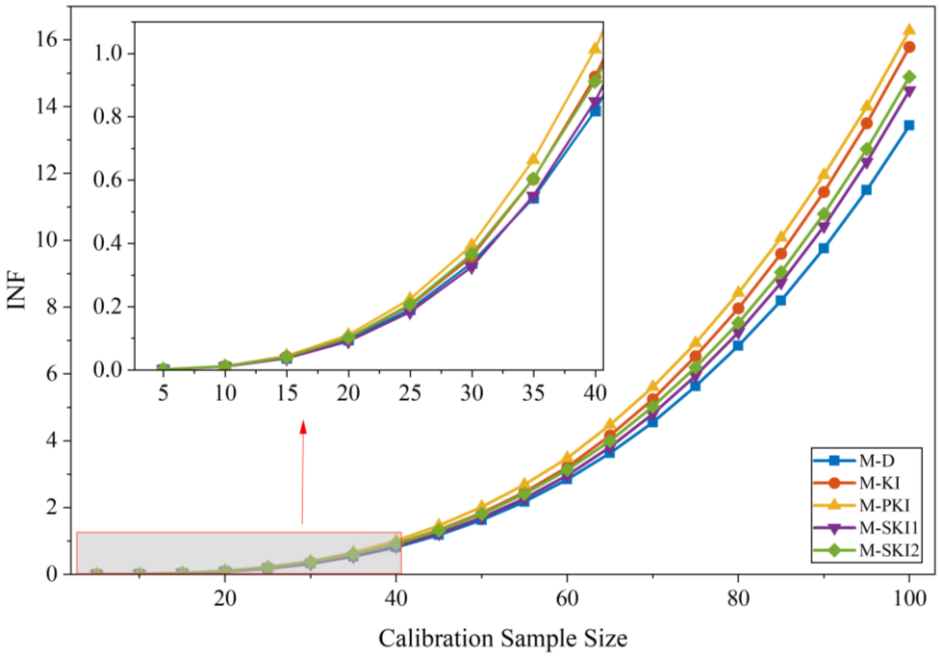


Figure 7. Calibration information as a function of the number of responding examinees under an initial item parameter estimation error of 0.1

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

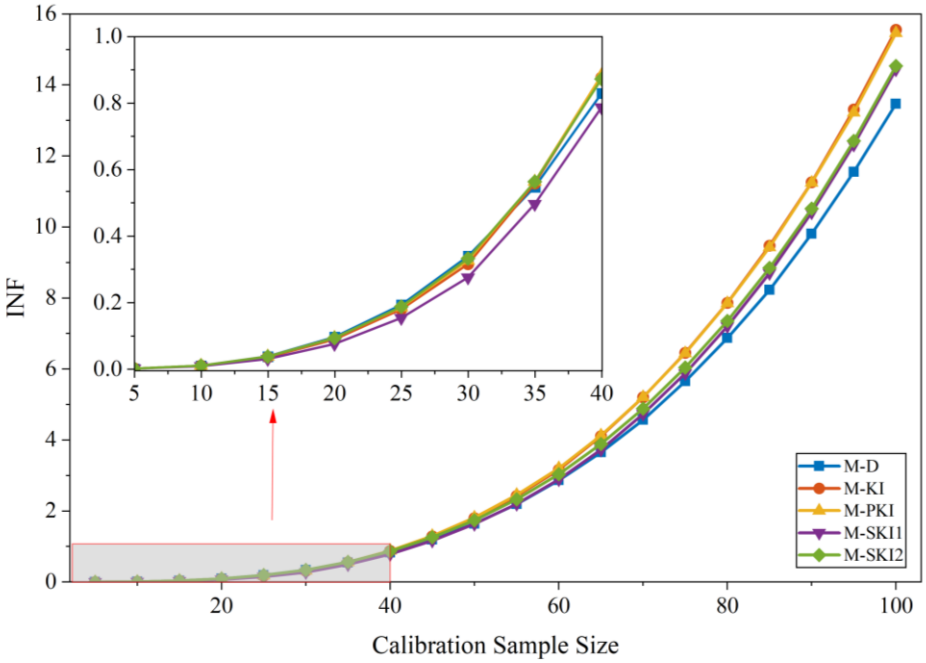


Figure 8. Calibration information as a function of the number of responding examinees under an initial item parameter estimation error of 0.5

First, the calibration of new items under the M-D design exhibited a distinct pattern: new items were calibrated in batches. In this study, the calibration of 20 new items was divided into four successive batches. This indicates that the M-D design tends to complete the calibration of some new items prematurely and preferentially selects items that have already been administered to a certain number of examinees. This phenomenon aligns with the characteristics of the reformed D-optimal design in unidimensional contexts, as noted by He and Chen (2020), van der Linden and Ren (2015), and Zheng (2014).

This batch-wise retirement pattern implies that if the calibration process is terminated prematurely, a large number of unretired items would have little or no valid response data. In contrast, the continuous retirement pattern exhibited by the KL information criteria allows each item to accumulate a certain amount of response data

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

even in the event of premature termination, thereby significantly enhancing the reliability and robustness of the calibration procedure.

Furthermore, examining the information at the time of retirement revealed the following patterns. Under the M-D design, the amount of information was similar across batches. Under the M-KI design, new items retiring with high information were more concentrated in the middle stage of calibration. For the M-PKI, M-SKI1, and M-SKI2 designs, new items retiring with high information were generally concentrated in the early stage, allowing items with higher information to be calibrated first.

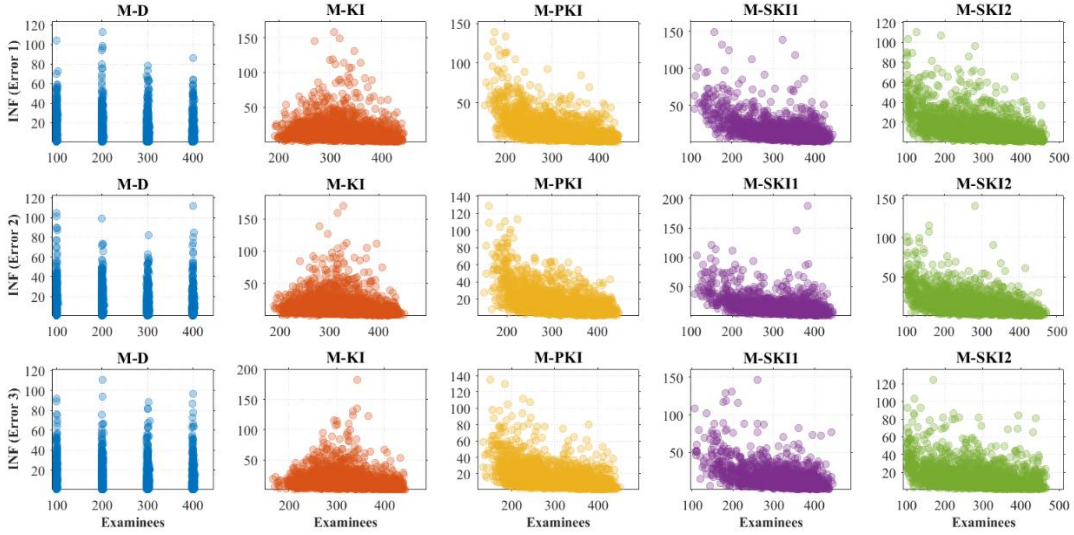


Figure 9. Scatterplot of information (INF) obtained for each new item at retirement, plotted against the cumulative total number of examinees under three initial item parameter estimation errors. Note. Error 1, Error 2, and Error 3 represent the initial item parameter estimation errors of 0, 0.1, and 0.5, respectively.

5.4.2. Results of Extreme Set

To test the robustness of the proposed designs under more challenging conditions, their performance on extreme items under the initial item parameter estimation error of 0.1 was further analyzed. As presented in Table 3, all KL information criteria outperformed the M-D design in overall parameter recovery (WMSE), with the M-PKI design performing best. In terms of the recovery of the

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

individual discrimination and intercept parameters, the M-PKI design again achieved the best results, whereas the M-SKI1 design performed worse than the M-D design. The difficulty of estimating extreme item parameters appears to magnify the limitations of the M-D design while amplifying the advantages of the KL information criteria.

The corresponding RMSE and Bias trends are shown in Figure 10. The KL information criteria also outperformed the M-D design in the early calibration stage, and this advantage was even more pronounced for extreme items. Notably, the M-SKI1 design maintained a clear advantage in the initial stage, with its estimation errors substantially smaller than those of the other designs.

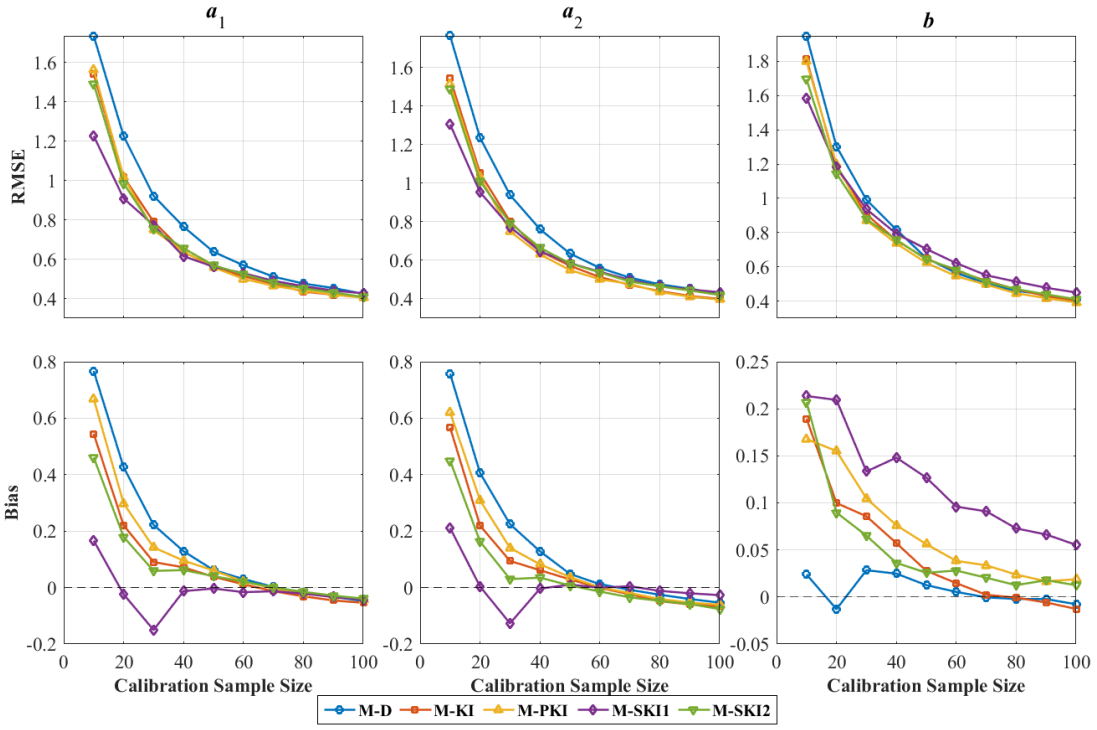


Figure 10. Trend of RMSE and Bias for extreme items

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Table 3. Calibration precision of five designs with extreme items

Design	Bias		RMSE		Corr(η)		SD_R		WMSE
	ave_a	*b*	*ave_a*	*b*	*ave_a*	*b*	*ave_a*	*b*	
M-D	-0.0506	-0.0086	0.4255	0.4090	0.8291	0.9741	0.0916	0.0993	0.00449
M-KI	-0.0601	-0.0062	0.4006	0.4005	0.8471	0.9757	**0.0815**	0.0888	0.00398
M-PKI	-0.0527	0.0216	**0.4003**	**0.3917**	**0.8481**	**0.9769**	0.0838	**0.0734**	**0.00387**
M-SKI1	-0.0347	0.0643	0.4327	0.4660	0.8215	0.9677	0.0876	0.1196	0.00405
M-SKI2	-0.0591	0.0116	0.4136	0.4053	0.8363	0.9747	0.0834	0.0836	0.00404

Notes. (1) Corr(η) represents the Pearson correlation coefficient between the estimated item parameters and the true values, and SD_R denotes the standard deviation of RMSE across 100 replications; (2) *ave_a* represents the average of the Bias, RMSE, Corr(η) or SD_R values on two discrimination parameters; (3) because the values of WMSE are relatively small, we report five significant digits rather than four significant digits to reflect differences, and omitting five decimal places will not affect the conclusion.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

6. Simulation Study 2

Simulation Study 1 was conducted in the context of within-item multidimensionality. To further examine the performance of the different designs across diverse multidimensional structures, we designed Simulation Study 2 under both within-item and between-item multidimensionality with an initial item parameter estimation error of 0.1.

6.1. General Setup

To ensure the fairness of the two item banks, before generating the two structured banks, an item bank in which all items measured all dimensions was first constructed, and the number and generation method of items consistent with those in Study 1. This bank was then split into a within-item multidimensional bank (Item bank 1) and a between-item multidimensional bank (Item bank 2). Item bank 1 consists of 500 items measuring only a single dimension and 500 items measuring both dimensions; Item bank 2 consists of 500 items measuring only the first dimension and 500 items measuring only the second dimension. Table 4 presents the structural characteristics of the simulated item banks.

New items were also constructed using the same method to form two types: a within-item multidimensional structure (New item set 1) and a between-item multidimensional structure (New item set 2), and the number and generation method of items consistent with those in Study 1. New item set 1 consists of 10 items measuring only one dimension and 10 items measuring both dimensions; New item

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

set 2 consists of 10 items measuring only the first dimension and 10 items measuring only the second dimension. The structural characteristics are presented in Table 5.

Table 4. Structural description of two simulated item banks

Item no.	Item bank			Item bank 1 (within-item)			Item bank 2 (between-item)		
	a_1	a_2	b	a_1	a_2	b	a_1	a_2	b
1	$a_{1,1}$	$a_{1,2}$	b_1	$a_{1,1}$	0	b_1	$a_{1,1}$	0	b_1
2	$a_{2,1}$	$a_{2,2}$	b_2	0	$a_{2,2}$	b_2	$a_{2,1}$	0	b_2
…	…	…	…	…	…	…	…	…	…
249	$a_{249,1}$	$a_{249,2}$	b_{249}	$a_{249,1}$	0	b_{249}	$a_{249,1}$	0	b_{249}
250	$a_{250,1}$	$a_{250,2}$	b_{250}	0	$a_{250,2}$	b_{250}	$a_{250,1}$	0	b_{250}
251	$a_{251,1}$	$a_{251,2}$	b_{251}	$a_{251,1}$	0	b_{251}	$a_{251,1}$	0	b_{251}
252	$a_{252,1}$	$a_{252,2}$	b_{252}	0	$a_{252,2}$	b_{252}	$a_{252,1}$	0	b_{252}
…	…	…	…	…	…	…	…	…	…
499	$a_{499,1}$	$a_{499,2}$	b_{499}	$a_{499,1}$	0	b_{499}	$a_{499,1}$	0	b_{499}
500	$a_{500,1}$	$a_{500,2}$	b_{500}	0	$a_{500,2}$	b_{500}	$a_{500,1}$	0	b_{500}
501	$a_{501,1}$	$a_{501,2}$	b_{501}	$a_{501,1}$	$a_{501,2}$	b_{501}	0	$a_{501,2}$	b_{501}
502	$a_{502,1}$	$a_{502,2}$	b_{502}	$a_{502,1}$	$a_{502,2}$	b_{502}	0	$a_{502,2}$	b_{502}
…	…	…	…	…	…	…	…	…	…
999	$a_{999,1}$	$a_{999,2}$	b_{999}	$a_{999,1}$	$a_{999,2}$	b_{999}	0	$a_{999,2}$	b_{999}
1000	$a_{1000,1}$	$a_{1000,2}$	b_{1000}	$a_{1000,1}$	$a_{1000,2}$	b_{1000}	0	$a_{1000,2}$	b_{1000}

Table 5. Structural description of two new item sets

New item no.	New item set 1 (within-item design)			New item set 2 (between-item design)		
	a_1	a_2	b	a_1	a_2	b
1	$a_{1,1}$	0	b_1	$a_{1,1}$	0	b_1
2	0	$a_{2,2}$	b_2	$a_{2,1}$	0	b_2
…	…	…	…	…	…	…
9	$a_{9,1}$	0	b_9	$a_{9,1}$	0	b_9
10	0	$a_{10,2}$	b_{10}	$a_{10,1}$	0	b_{10}
11	$a_{11,1}$	$a_{11,2}$	b_{11}	0	$a_{11,2}$	b_{11}
12	$a_{12,1}$	$a_{12,2}$	b_{12}	0	$a_{12,2}$	b_{12}
…	…	…	…	…	…	…
19	$a_{19,1}$	$a_{19,2}$	b_{19}	0	$a_{19,2}$	b_{19}
20	$a_{20,1}$	$a_{20,2}$	b_{20}	0	$a_{20,2}$	b_{20}

6.2. Results

Table 6 presents the calibration precision and average information of each online calibration design under two item bank structures, namely within-item multidimensionality and between-item multidimensionality. Overall, the calibration

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

performance of all designs under the between-item condition was considerably better than that under the within-item condition. The performance pattern differed across the two item-bank structures. This difference can be attributed to the fact that, in the between-item structure, information is more concentrated along each dimension, making the optimization of information-based criteria more effective. In contrast, in the within-item design, selected items tend to provide information on multiple dimensions simultaneously, but the information for each dimension is relatively dispersed. This dispersion, combined with the correlations among multidimensional parameters, further impairs calibration efficiency, whereas the between-item structure naturally avoids this problem.

A further comparison of the performance of each design across the two frameworks revealed that the relative order of superiority remained consistent. The M-PKI design, grounded in the KL information criteria, demonstrated a clear advantage in recovering the discrimination parameter *a*, whereas the M-D design excelled in recovering the intercept parameter *b*. In contrast, the overall calibration precision of the M-SKI1 and M-SKI2 designs remained relatively low. Nevertheless, the M-SKI2 design performed significantly better than the M-SKI1 design.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Table 6. Calibration precision of five designs under two multidimensional structures

Structure	Design	Bias		RMSE		Corr($\boldsymbol{\eta}$)		SD_R		WMSE	INF
		ave_a	*b*	*ave_a*	*b*	*ave_a*	*b*	*ave_a*	*b*		
Within	M-D	-0.0390	0.0065	0.3511	**0.2764**	0.8113	**0.9627**	0.0688	**0.0582**	0.00612	22.0423
	M-KI	-0.0225	0.0103	0.3476	0.2969	0.8153	0.9607	0.0665	0.0774	0.00613	24.7927
	M-PKI	-0.0261	0.0085	**0.3307**	0.3126	**0.8348**	0.9556	**0.0656**	0.0602	**0.00576**	**25.9982**
	M-SKI1	0.0253	0.0166	0.3898	0.3822	0.7894	0.9398	0.0806	0.0910	0.00707	23.1945
	M-SKI2	-0.0028	0.0073	0.3703	0.3392	0.8030	0.9514	0.0767	0.0772	0.00666	24.0258
Between	M-D	-0.0285	-0.0034	0.3113	**0.2557**	0.8474	**0.9684**	0.0670	0.0480	0.00474	30.8271
	M-KI	0.0078	0.0021	0.3223	0.2989	0.8418	0.9641	0.0674	0.0775	0.00513	29.9997
	M-PKI	-0.0212	0.0084	**0.3056**	0.2709	**0.8558**	0.9649	**0.0655**	**0.0471**	**0.00456**	**32.7139**
	M-SKI1	0.0343	0.0085	0.3544	0.3622	0.8208	0.9467	0.0822	0.0835	0.00623	29.1602
	M-SKI2	-0.0024	0.0012	0.3279	0.3095	0.8371	0.9565	0.0707	0.0638	0.00535	31.6146

Notes. (1) Corr($\boldsymbol{\eta}$) represents the Pearson correlation coefficient between the estimated item parameters and the true values, and SD_R denotes the standard deviation of RMSE across 100 replications; (2) *ave_a* represents the average of the Bias, RMSE, Corr($\boldsymbol{\eta}$) or SD_R values on two discrimination parameters; (3) because the values of WMSE are relatively small, we report five significant digits rather than four significant digits to reflect differences, and omitting five decimal places will not affect the conclusion.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

7. Semi-real Data Analysis

In this part of the study, semi-real data from the International Association for the Evaluation of Educational Achievement (IEA) Trends in International Mathematics and Science Study (TIMSS) were used to validate the effectiveness of the proposed designs. TIMSS 2023 was the first cycle in which the fourth and eighth grade assessments were fully digitized with a group-adaptive assessment design to better measure the ability levels of participating students. The fourth-grade mathematics achievement was selected as the research focus.

7.1. Raw Data Illustration

The mathematics items of TIMSS 2023 covered three content domains: *Data*, *Number*, and *Measurement* and *Geometry*. To focus on two specific content domains, *Data* and *Number* were selected as the measurement dimensions. Hence, only the items that participated in the digital assessment were retained, and the Problem-Solving and Inquiry (PSI) tasks along with polytomously scored items were excluded. This resulted in 34 items measuring *Data* and 74 items measuring *Number*, yielding a total of 108 unidimensional items.

7.2. Construction of Item Bank and New Item

TIMSS 2023 provided the unidimensional IRT item parameters for each item. To preserve data authenticity, these parameters were used directly to represent the between-item multidimensional structure. Specifically, the 108 original unidimensional items were treated as between-item multidimensional items, with *Data* serving as the first dimension and *Number* as the second.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

To further simulate a more realistic multidimensional testing scenario, an additional 108 within-item multidimensional items were generated from the original items, such that each item simultaneously measures both dimensions (*Data* as the first, *Number* as the second). The detailed construction procedure is outlined below:

Discrimination Parameters. First, the means $\mu_{\tilde{a}_1}$ and $\mu_{\tilde{a}_2}$, and the variances $Var(\tilde{a}_1)$ and $Var(\tilde{a}_2)$ of the discrimination parameters for the dimensions measured by the real items were calculated to serve as the basis for generating the two-dimensional discrimination parameters. Here, $\tilde{a}_1$ and $\tilde{a}_2$ denote the discrimination parameters for the first and second dimensions, respectively.

However, due to the unequal numbers of items measuring the two dimensions, the correlation between dimensions could not be directly calculated. To construct a semi-synthetic within-item bank, we estimated a plausible cross-dimensional discrimination correlation from matched subsets. Specifically, we randomly selected 34 items from the 74 items measuring the second dimension and paired them with the 34 items measuring the first dimension, which yielded a correlation of 0.4751. Given that the two ability dimensions, *Data* and *Number*, are cognitively correlated, we randomly generated the correlation coefficient ρ_{a_1,a_2} between the two-dimensional discrimination parameters from a uniform distribution $U(0.4, 0.6)$. Let $Var(a_1) = Var(\tilde{a}_1)$ and $Var(a_2) = Var(\tilde{a}_2)$, and the covariance $Cov(a_1, a_2) = \rho_{a_1,a_2} \times \sqrt{Var(a_1)} \times \sqrt{Var(a_2)}$ was then derived accordingly.

Finally, the truncation interval was set between the minimum (i.e., 0.3981) and maximum (i.e., 1.5591) true discrimination values of the two dimensions. The

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

discrimination parameters for the 108 items measuring both dimensions were generated from a truncated multivariate normal distribution with mean vector of $\boldsymbol{\mu}_a = \left(\mu_{\tilde{a}_1}, \mu_{\tilde{a}_2}\right)^T$ and covariance matrix $\boldsymbol{\Sigma}_a = \begin{pmatrix} Var(a_1) & Cov(a_1, a_2) \\ & Var(a_2) \end{pmatrix}$.

Intercept parameter. First, the mean $\mu_{\tilde{b}}$ and variance $Var\left(\tilde{b}\right)$ of the intercept parameters for all real items were calculated, where $\tilde{b}$ denotes the difficulty parameter of the real items. Next, the difficulty parameters for the 108 items were generated from a truncated normal distribution with mean $\mu_{\tilde{b}}$ and variance $Var\left(\tilde{b}\right)$, with the truncation range also set between the minimum (i.e., -2.9194) and maximum (i.e., 1.7971) difficulty values of all the original items. Since the difficulty parameter equals $b/\sqrt{a_1^2 + a_2^2}$, the intercept parameter value *b* was obtained through transformation.

In sum, a total of 216 items were obtained, comprising 108 items with between-item multidimensionality and 108 items with within-item multidimensionality. In addition, for the items with between-item multidimensionality, their difficulty parameters were also transformed into intercept parameters using the same transformation procedure. Finally, 20 items were randomly selected from these 216 items as new items to be calibrated, and the remaining items served as the item bank.

7.3. MCAT Online Calibration Procedure

Unlike the simulation studies, each new item was first randomly administered to 20 examinees to obtain initial parameter estimates before adaptive assignment. With these initial responses already accumulated, the cold-start strategy used in the simulation study was unnecessary, and the M-SKI2 design could be directly applied

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

for item selection. In addition, given the substantial advantage of the M-SKI1 design in the early stage of calibration, we combined the M-SKI1 design with the M-D design. Each item was administered adaptively to 100 examinees, with the first 30 examinees matched using the M-SKI1 design and the remaining 70 using the M-D design. For ease of reference, this design is referred to as the M-SKI1&D design. The generation of examinees and all other procedures were kept consistent with those in the simulation study, and both the selection of new items and the MCAT process were replicated 100 times.

7.4. Results

Table 7. Calibration precision of the six designs on the semi-real data

Design	Bias		RMSE		SD_R		WMSE
	ave _ a	*b*	*ave _ a*	*b*	*ave _ a*	*b*	
M-D	-0.1950	0.0637	0.3655	0.3503	0.0517	**0.0656**	0.00847
M-KI	-0.1846	0.0725	0.3623	0.3757	0.0468	0.0698	0.00833
M-PKI	-0.1781	0.0754	0.3549	0.3487	0.0484	0.0714	0.00784
M-SKI1	-0.1723	0.0706	0.3679	0.3936	0.0500	0.0731	0.00903
M-SKI2	-0.1658	0.0531	0.3656	0.3872	**0.0461**	0.0743	0.00887
M-SKI1&D	-0.1807	0.0720	**0.3536**	**0.3379**	0.0497	**0.0656**	**0.00777**

Notes. (1) SD_R denotes the standard deviation of RMSE across 100 replications; (2) *ave_a* represents the average of the Bias RMSE, or SD_R values on two discrimination parameters; (3) because the values of WMSE are relatively small, we report five significant digits rather than four significant digits to reflect differences, and omitting five decimal places will not affect the conclusion; (4) M-SKI1&D design is a hybrid design where M-SKI1 design is applied in the early stage of calibration for each new item, and M-D design is applied thereafter.

Table 7 presents the parameter recovery results of the semi-real data under different designs. The calibration performance of each design was generally consistent with the results from simulation studies. In terms of both overall and individual parameter recovery, the M-SKI1&D design achieved the best performance. By compensating for the early-stage disadvantage of the M-D design, this hybrid approach significantly improved calibration precision. The M-PKI design ranked

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

second overall, also demonstrating strong performance across both the discrimination and intercept parameters. In terms of estimation stability, the M-SKI2 design achieved the smallest SD_R for the discrimination parameter, while the M-SKI1&D design achieved the smallest SD_R for the intercept parameter. The empirical results further verified the advantage of the M-PKI design over the M-D design in parameter recovery, and confirmed the feasibility of combining M-SKI1 with the M-D design as well as the resulting improvement in calibration precision.

8. Discussion and Future Direction

Online calibration design, a key technique for matching new items with examinees, is essential to the calibration framework and determines final precision. However, existing online calibration designs are almost exclusively based on Fisher information, which overlook its limitations in the early stage of calibration. To address this gap, this study leveraged the advantage of KL information over Fisher information when item parameter estimates are relatively imprecise (i.e., the early stage of calibration), and proposed a series of novel online calibration designs, namely the M-KI, M-PKI, M-SKI1, and M-SKI2 designs. KI is a scalar indicator obtained by integrating KL information. Compared with Fisher information, KI captures not only the information from the current provisional estimates of item parameters but also the information within the neighborhood of these estimates. The M-KI design aims to maximize KI, while the M-PKI design further incorporates prior information. Motivated by the need to reduce computational cost, we further proposed two

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

simplified designs, the M-SKI1 and M-SKI2, which approximate the KI indicator via local Taylor expansions of the KL information.

Simulation results indicated that the KL information criteria outperformed the M-D design in the early stage of calibration. In terms of final calibration precision, the M-KI and M-PKI designs showed superior performance to the M-D design under various conditions, particularly in the recovery of discrimination parameters. Compared with M-KI and M-PKI, the M-SKI1 and M-SKI2 designs sacrificed some calibration precision but achieved substantial savings in computation time while retaining the advantage of KL information in the early calibration stage. In the empirical study, the M-SKI1 design was combined with the M-D design, leading to further improvement in the calibration performance of the M-D design. For practical applications, the M-SKI1&D design is recommended, as it achieved higher calibration precision while maintaining computational efficiency.

Several directions merit further investigation. In real online calibration settings, the total number of examinees, the number of responses per new item, and the method for obtaining initial item parameter estimates may all influence calibration outcomes. Future research should evaluate the stability of the proposed KL information criteria under a wider range of experimental conditions. Additionally, the degree of consistency between the parameter distributions of new items and those of operational items warrants further investigation, as systematic discrepancies may exist in practice.

Compared with paper-and-pencil (P&P) tests, a key advantage of CAT is the ease of obtaining response times, which have been widely utilized in CAT research,

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

including operational item selection (e.g., Fan et al., 2012; Cheng et al., 2017; Choe et al., 2018; He & Qi, 2023) and online calibration (e.g., Kang et al., 2020; He et al., 2021; He & Qi, 2025). Future work could integrate response times into KL information criteria. For instance, new items could be assigned to maximize KI and PKI per unit of time, thereby preventing the calibration process from being prolonged by assigning examinees to items that are highly informative but overly time-consuming, particularly under time-limited conditions.

In the sequential calibration process, two issues may arise: (1) items introduced later in the calibration process may receive poorer calibration samples due to the reduced availability of examinees; and (2) items with suboptimal measurement properties (e.g., lower information) may be less likely to be selected by adaptive designs. These issues can lead to uneven calibration quality across items. To address these concerns, Zheng (2014) and He et al. (2020) proposed the ordered informative range priority index (OIRPI) and the ED criterion, both based on Fisher information. However, like most Fisher information criteria, KL information criteria belong to the category of optimal designs based on item information standards. Therefore, future research could integrate the proposed methods into the OIRPI framework or construct analogous indicators inspired by the ED criteria to balance the sample sizes across items.

Although all designs in this study were developed and evaluated in a two-dimensional context, they can be readily generalized to higher dimensions (i.e., $p \geqslant 3$). However, for the M-SKI2 design, the multiplicative weight $1-\rho^2$ in the two-

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

dimensional case generalizes to the determinant of the $p \times p$ ability correlation matrix, which varies with the dimensionality and the correlation structure among dimensions. In addition, the proposed KL information criteria are not restricted to the M2PLM; they are highly adaptable and can be extended to the multidimensional three-parameter logistic model (M3PLM), which includes a guessing parameter, as well as to polytomously scored IRT models, such as the multidimensional graded response model (MGRM). Moreover, non-compensatory MIRT models also represent a promising direction for future research, as their parameter interpretation differs fundamentally from that of compensatory models.

nCompeting interests. The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Financial support. This study was supported by ① Humanities and Social Science Fund of Ministry of Education of China (Grant No. 25YJCZH073); and ② The National Natural Science Foundation of China (Grant No. 32000778).

References

Berger, M. P. F. (1994). D-optimal sequential sampling designs for item response theory models. *Journal of Educational Statistics*, *19*, 43–56.

Chang, H. H., & Ying, Z. L. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213–229.

Chang, Y. C. I., & Lu, H. Y. (2010). Online calibration via variable length computerized adaptive testing. *Psychometrika*, *75*, 140–157.

Chen, P. (2017). A comparative study of online item calibration methods in multidimensional computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *42*, 559–590.

Chen, P., & Wang, C. (2016). A new online calibration method for multidimensional computerized adaptive testing. *Psychometrika*, *81*, 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*, 81–117.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Cheng, Y., Diao, Q., & Behrens, J. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behavior Research Methods*, *49*, 502–512.

Cheville, A. L., Wang, C., Yost, K. J., Teresi, J. A., Ramirez, M., Ocepek-Welikson, K., ... & Weiss, D. J. (2021). Improving the delivery of function-directed care during acute hospitalizations: methods to develop and validate the Functional Assessment in Acute Care Multidimensional Computerized Adaptive Test (FAMCAT). *Archives of Rehabilitation Research and Clinical Translation*, *3*, 100112.

Choe, E., Kern, J. L., & Chang, H. H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *43*, 135–158.

Fan, Z., Wang, C., Chang, H. H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*, 655–670.

Giordano, A., Testa, S., Bassi, M., Cilia, S., Bertolotto, A., Quartuccio, M. E., ... & Rosato, R. (2023). Applying multidimensional computerized adaptive testing to the MSQOL-54: a simulation study. *Health and Quality of Life Outcomes*, *21*, 61.

Hassan, M. U., & Miller, F. (2019). Optimal item calibration for computerized achievement tests. *Psychometrika*, *84*, 1101–1128.

Hassan, M. U., & Miller, F. (2022). Discrimination with unidimensional and multidimensional item response theory models for educational data. *Communications in Statistics–Simulation and Computation*, *51*, 2992–3012.

Hassan, M. U., & Miller, F. (2024). Optimal calibration of items for multidimensional achievement tests. *Journal of Educational Measurement*, *61*, 274–302.

He, Y., & Chen, P. (2020). Optimal online calibration designs for item replenishment in adaptive testing. *Psychometrika*, *85*, 35–55.

He, Y., Chen, P., & Li, Y. (2020). New efficient and practicable adaptive designs for calibrating items online. *Applied Psychological Measurement*, *44*, 3–16.

He, Y., Chen, P., & Li, Y. (2021). Maximum information per time unit designs for continuous online item calibration. *British Journal of Mathematical and Statistical Psychology*, *74*, 24–51.

He, Y., & Qi, Y. (2023). Using response time in multidimensional computerized adaptive testing. *Journal of Educational Measurement*, *60*, 697–738.

He, Y., & Qi, Y. (2025). Maximum information per time unit designs for calibrating multidimensional items online. *Journal of Educational and Behavioral Statistics*, 10769986251348678.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*, 262–277.

Jones, D. H., & Jin, Z. Y. (1994). Optimal sequential designs for online item estimation. *Psychometrika*, *59*, 59–75.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Kang, H.-A., Zheng, Y., & Chang, H. H. (2020). Online calibration of a joint model of item responses and response times in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *45*, 175–208.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks* (Vol. 4, pp. 1942–1948). IEEE.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation*. New York, NY: Springer New York.

Lien, D.-H. D. (1985). Moments of truncated bivariate log-normal distributions. *Economics Letters*, *19*, 243–247.

Mao, X., Zhang, J., & Xin, T. (2022). The optimal design of bifactor multidimensional computerized adaptive testing with mixed-format items. *Applied Psychological Measurement*, *46*, 605–621.

Reckase, M. D. (2009). Multidimensional item response theory. New York: Springer.

Ren, H., van der Linden, W. J., & Diao, Q. (2017). Continuous online item calibration: Parameter recovery and item utilization. *Psychometrika*, *82*, 498–522.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354.

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, *66*, 79–97.

Song, L., & Wang, W. (2025). Compound optimal design for online item calibration under the two-parameter logistic model. *Applied Psychological Measurement*, *49*, 177–198.

Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas* (Vol. 374). Englewood Cliffs, NJ: Prentice-Hall.

Sung, Y. T., Lien, Y. C., Chou, Y. T., & Yang, W. H. (2026). Constructing a Computerized Adaptive Test for Multidimensional Mathematical Competence: Development, Simulation, and Validation. *International Journal of Science and Mathematics Education*, *24*, 73.

van der Linden, W. J., & Ren, H. (2015). Optimal Bayesian adaptive design for test-item calibration. *Psychometrika*, *80*, 263–288.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575–588.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 65–102). Hillsdale, NJ: Lawrence Erlbaum.

Wang, C., & Chang, H. H. (2009). Kullback-Leibler information in multidimensional adaptive testing: Theory and application. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. Retrieved from http://www.psych.umn.edu/psylabs/CATCentral/

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Wang, C., & Chang, H. H. (2011). Item selection in multidimensional computerized adaptive testing-gaining information from different angles. *Psychometrika*, *76*, 363–384.

Wang, C., Chang, H. H., & Boughton, K. A. (2011). Kullback-Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika*, *76*, 13–39.

Yang, S., He, Y., & Qi, Y. (2023). Research on multidimensional online calibration design based on D-optimal and A-optimal designs. *Advances in Applied Mathematics*, *12*, 81–95.

Yuan, L., Huang, Y., & Chen, P. (2026). Online calibration for multidimensional CAT with polytomously scored items: A neural network–based approach. *Journal of Educational and Behavioral Statistics*, 10769986251315531.

Yuan, L., Huang, Y., Li, S., & Chen, P. (2023). Online calibration in multidimensional computerized adaptive testing with polytomously scored items. *Journal of Educational Measurement*, *60*, 476–500.

Zheng, Y. (2014). New methods of online calibration for item bank replenishment [Doctoral dissertation, University of Illinois at Urbana-Champaign].

Appendix A. Relationship between KI and Ability Parameters

This appendix derives the relationship between the KI and examinees' ability $\boldsymbol{\theta}$ parameters for the discrimination parameter $\boldsymbol{a}$ and the intercept parameter b, respectively.

A1. Relationship between the KI of $\boldsymbol{a}$ and $\boldsymbol{\theta}$

In a two-dimensional scenario, assuming the intercept parameter is fixed at its true value ($b = b_j$), the KL information between the true discrimination parameters $\boldsymbol{a}_j = \left(a_{j1}, a_{j2}\right)^T$ and any given discrimination parameters $\boldsymbol{a} = \left(a_1, a_2\right)^T$ under the 2D-2PLM is:

$$\mathrm{KL}_a(\boldsymbol{a} \,\|\, \boldsymbol{a}_j) = P(\boldsymbol{a}_j)\log\frac{P(\boldsymbol{a}_j)}{P(\boldsymbol{a})} + [1-P(\boldsymbol{a}_j)]\log\frac{1-P(\boldsymbol{a}_j)}{1-P(\boldsymbol{a})}. \tag{A1}$$

To evaluate this locally, we perform a second-order Taylor expansion of KL_a at $\boldsymbol{a} = \boldsymbol{a}_j$. Note that $\mathrm{KL}_a(\boldsymbol{a}_j \,\|\, \boldsymbol{a}_j) = 0$ and its first derivative at $\boldsymbol{a}_j$ is zero. The second derivative (Hessian matrix) evaluated at $\boldsymbol{a}_j$ is the Fisher information matrix for the

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

discrimination parameters, denoted as $\boldsymbol{I}(\boldsymbol{a}_j)$. Under the 2D-2PLM, $\boldsymbol{I}(\boldsymbol{a}_j) = P(\boldsymbol{a}_j)Q(\boldsymbol{a}_j)\boldsymbol{\theta}\boldsymbol{\theta}^T$, where $Q(\boldsymbol{a}_j) = 1 - P(\boldsymbol{a}_j)$.

Substituting this into the Taylor expansion yields:

$$\mathrm{KL}_a(\boldsymbol{a} \,||\, \boldsymbol{a}_j) \approx \frac{1}{2}(\boldsymbol{a} - \boldsymbol{a}_j)^T \left[P(\boldsymbol{a}_j)Q(\boldsymbol{a}_j)\boldsymbol{\theta}\boldsymbol{\theta}^T \right](\boldsymbol{a} - \boldsymbol{a}_j) + O(\| \boldsymbol{a} - \boldsymbol{a}_j \|^3). \quad \text{(A2)}$$

To compute the KI_a, we define the integration region as a circular disk Ω with a small radius *R* centered at $\boldsymbol{a}_j$ (Wang & Chang, 2009; Wang et al., 2011). By transforming from rectangular to cylindrical coordinates, we set:

$$\begin{cases} a_1 - a_{j1} = r\cos\alpha, \\ a_2 - a_{j2} = r\sin\alpha, \end{cases} \text{ with } 0 \le r \le R, 0 \le \alpha \le 2\pi. \quad \text{(A3)}$$

The quadratic form can be expanded as:

$$(\boldsymbol{a} - \boldsymbol{a}_j)^T \boldsymbol{\theta} = \theta_1(r\cos\alpha) + \theta_2(r\sin\alpha). \quad \text{(A4)}$$

Ignoring the higher-order error terms $o(r^3)$, the KI_a over the circular region Ω is:

$$\mathrm{KI}_a = \int_0^{2\pi}\int_0^R \frac{1}{2} P(\boldsymbol{a}_j)Q(\boldsymbol{a}_j)\left[r(\theta_1\cos\alpha + \theta_2\sin\alpha)\right]^2 r dr d\alpha. \quad \text{(A5)}$$

Separating the integrals for *r* and α:

$$\mathrm{KI}_a = \frac{1}{2}P(\boldsymbol{a}_j)Q(\boldsymbol{a}_j)\left(\int_0^R r^3 dr\right)\left[\int_0^{2\pi}(\theta_1^2\cos^2\alpha + \theta_2^2\sin^2\alpha + 2\theta_1\theta_2\sin\alpha\cos\alpha)d\alpha\right]. \quad \text{(A6)}$$

Applying the standard definite integrals, we obtain:

$$\mathrm{KI}_a = \frac{1}{2}P(\boldsymbol{a}_j)Q(\boldsymbol{a}_j)\left(\frac{R^4}{4}\right)\left(\theta_1^2\pi + \theta_2^2\pi\right) = \frac{\pi R^4}{8}P(\boldsymbol{a}_j)Q(\boldsymbol{a}_j)(\theta_1^2 + \theta_2^2). \quad \text{(A7)}$$

This mathematically proves that as the integration region approaches zero, the information provided for calibrating item discrimination is directly proportional to the squared length of the examinee's ability vector:

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

$$\mathrm{KI}_a \propto \theta_1^2 + \theta_2^2. \tag{A8}$$

For a fixed item (fixed a_1, a_2, b), when KI is computed by integrating KL information over a small neighborhood in the discrimination parameter space (centered at the current estimates), the resulting KI value is proportional to $\theta_1^2 + \theta_2^2$. This means that examinees with higher ability provide more information for calibrating the item parameters.

A2. Relationship between the KI of *b* and $\boldsymbol{\theta}$

Next, we fix the discrimination vector $\boldsymbol{a}$ at its true value $\boldsymbol{a}_j$ and evaluate the information for calibrating the intercept parameter *b*. The KL information as a function of any given intercept *b* around the true intercept b_j is:

$$\mathrm{KL}_b(b \,||\, b_j) = P\left(b_j\right)\log\frac{P\left(b_j\right)}{P\left(b\right)} + \left(1-P\left(b_j\right)\right)\log\frac{1-P\left(b_j\right)}{1-P\left(b\right)}. \tag{A9}$$

We perform a second-order Taylor expansion around $b = b_j$. Note that $\mathrm{KL}_b(b_j \,||\, b_j) = 0$ and its first derivative at b_j is zero. The second derivative of KL_b at b_j equals the Fisher information for the intercept parameter, which is $P\left(b_j\right)Q\left(b_j\right)$. Thus, the expansion yields:

$$\mathrm{KL}_b(b \,||\, b_j) \approx \frac{1}{2} P(b_j)Q(b_j)(b-b_j)^2 + O(|\, b-b_j\,|^3). \tag{A10}$$

To evaluate the KI for the intercept parameter (KI_b), we integrate this local divergence over a small interval $[t_j - \delta, t_j + \delta]$:

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

$$\begin{aligned} \mathrm{KI}_b &= \int_{b_j-\delta}^{b_j+\delta} \frac{1}{2} P(b_j)Q(b_j)(b-b_j)^2\,db \\ &= \frac{1}{2} P(b_j)Q(b_j)\left[\frac{(b-b_j)^3}{3}\right]_{b_j-\delta}^{b_j+\delta} \\ &= \frac{\delta^3}{3} P(b_j)Q(b_j). \end{aligned} \tag{A11}$$

Thus, KI_b is directly proportional to $P\left(b_j\right)Q\left(b_j\right)$. Since $P(b_j)=\dfrac{1}{1+\exp(-(\boldsymbol{a}_j^T\boldsymbol{\theta}-b_j))}$, the parabolic function $P\left(b_j\right)Q\left(b_j\right)$ reaches its global maximum of 0.25 when $\boldsymbol{a}_j^T\boldsymbol{\theta}-b_j=0$. This algebraically proves that KI_b is maximized when the examinee's composite ability exactly matches the item's true intercept $\left(\boldsymbol{a}_j^T\boldsymbol{\theta}=b_j\right)$.

Remark on higher dimensions. Although the derivation above is presented for the two-dimensional case for geometric clarity, the same conclusions hold for any $p\geq 2$. For a general p-dimensional ability vector $\boldsymbol{\theta}=\left(\theta_1,\theta_2,\ldots\theta_p\right)^T$ and discrimination vector $\boldsymbol{a}=\left(a_1,a_2,\ldots a_p\right)^T$, the same algebraic steps (Taylor expansion followed by integration over a small hyper spherical region) yield: $\mathrm{KI}_a \propto \sum_{k=1}^{p}\theta_k^2$, and KI_b is maximized when $\boldsymbol{a}^T\boldsymbol{\theta}-b=0$.

Appendix B. Derivation of the Integration Constant *c*=3

This appendix provides a rigorous theoretical justification for the choice of the constant c=3 in the integration interval used in the KL information index (Equation (12) in the main text). We demonstrate that this formulation is grounded in the asymptotic normality of the MLE estimator for item parameters.

B1. Asymptotic Distribution of the Item Parameter MLE

Consider a new item j that has been answered by n_j examinees. Under the 2D-2PLM, the likelihood function for the item parameter vector $\boldsymbol{\eta}_j$ given the response

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

vector $\boldsymbol{u}_{n_j j} = \left(u_{1j}, u_{1j}..., u_{n_j j}\right)^T$ and the examinee ability estimates $\hat{\boldsymbol{\theta}}^{(n_j)} = \left(\hat{\boldsymbol{\theta}}_{1j}, \hat{\boldsymbol{\theta}}_{2j}, ..., \hat{\boldsymbol{\theta}}_{n_j j}\right)^T$ is

$$L(\boldsymbol{\eta}_j; \boldsymbol{u}_{n_j j}, \hat{\boldsymbol{\theta}}^{(n_j)}) = \prod_{i=1}^{n_j} P^{u_{ij}}(\hat{\boldsymbol{\theta}}_{ij}) Q^{1-u_{ij}}(\hat{\boldsymbol{\theta}}_{ij}), \tag{B1}$$

where $Q(\hat{\boldsymbol{\theta}}_{ij}) = 1 - P(\hat{\boldsymbol{\theta}}_{ij})$. The MLE estimate $\hat{\boldsymbol{\eta}}_j$ is obtained by solving the score equations $\partial \log L / \partial \boldsymbol{\eta}_j = 0$.

Under standard regularity conditions (Lehmann & Casella, 1998), $\hat{\boldsymbol{\eta}}_j$ is consistent and asymptotically normal as the calibration sample size n_j increases. Specifically, the total Fisher information matrix contributed by all n_j examinees is $I_{total}\left(\boldsymbol{\eta}_j\right) = \sum_{k=1}^{n_j} \boldsymbol{I}\left(\boldsymbol{\eta}_j; \boldsymbol{\theta}_k\right)$, which is of order n_j. By the asymptotic theory of maximum likelihood estimation, this implies that

$$\hat{\boldsymbol{\eta}}_j \overset{\sim}{} N\left(\boldsymbol{\eta}_j, I_{total}^{-1}(\boldsymbol{\eta}_j)\right), n_j \to \infty. \tag{B2}$$

B2. Derivation of the Integration Interval Bounds

For the *v*-th component of the item parameter vector $\boldsymbol{\eta}_{jv}$, the asymptotic variance is the corresponding diagonal element of the inverse information matrix, which is of order n_j^{-1}. Consequently, the standard error $\mathrm{SE}\left(\hat{\boldsymbol{\eta}}_{jv}\right)$ is of order $n_j^{-1/2}$.

To ensure that the integration interval $\left(\hat{\boldsymbol{\eta}}_{jv} - \delta_n, \hat{\boldsymbol{\eta}}_{jv} + \delta_n\right)$ contains the true parameter $\boldsymbol{\eta}_{jv}$ with high probability as n_j increases, the radius δ_n must shrink at the same rate as the MLE's standard error, that is, $\delta_n = O\left(n_j^{-1/2}\right)$. If δ_n were to shrink faster (e.g., $O\left(n_j^{-1}\right)$), the interval would eventually fail to cover the true parameter; if it shrank slower, the interval would remain unnecessarily wide and fail to achieve the desired transition from global to local information as calibration progresses.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

Therefore, since the total Fisher information is of order n_j, following Chang and Ying (1996), the asymptotic confidence interval $\hat{\boldsymbol{\eta}}_{jv} \pm z_{\alpha/2} \cdot \mathrm{SE}\left(\hat{\boldsymbol{\eta}}_{jv}\right)$ can be expressed in the simplified form:

$$\hat{\boldsymbol{\eta}}_{jv} \pm \frac{c}{\sqrt{n_j}}. \tag{B3}$$

B3. The Statistical Meaning of *c*=3

In the standard normal distribution, $\Phi(3) - \Phi(-3) \approx 0.9973$ meaning that ± 3 standard errors define an approximate 99.73% confidence interval. Setting c=3 thus ensures that the integration interval covers the plausible parameter space with high probability under the asymptotic normal approximation. This mirrors the classic application of c=3 in Chang and Ying (1996).

B4. Implications for Online Calibration

Although Chang and Ying (1996) originally applied this logic to ability estimation, the same asymptotic theory applies directly to item parameter estimation in online calibration. This substitution is theoretically justified because the MLE of item parameters also obeys the same $O\left(n_j^{-1/2}\right)$ convergence rate.

Moreover, this derivation reveals the adaptive nature of the integration interval: When n_j is small (early calibration stage), the interval width $2c/\sqrt{n_j}$ is large, and the KL information integrates over a wide region of the discrimination parameter space, capturing robust global information and preventing premature convergence. As n_j increases (later calibration stage), the interval width shrinks rapidly. The KL index asymptotically exhibits local behavior, naturally converging to maximize Fisher

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.

information evaluated at the true parameter. This dynamic property ensures a theoretically sound transition from global to local information, which is the central motivation for using the KL-based design in the early stages of online calibration.

Downloaded from https://www.cambridge.org/core. 16 Nov 2026 at 11:46:58, subject to the Cambridge Core terms of use.