

Article

Evaluating Person-Parameter Estimation Methods in Cognitive-Diagnostic Models with Testlet Effects

Mei You, Jingjing Yang and Xin Xu * 

School of Science, Minzu University of China, Beijing 100081, China

* Correspondence: xux636@muc.edu.cn

Abstract

Testlet-based cognitive-diagnostic models involve both discrete attribute profiles and continuous testlet effects, making person-parameter estimation more challenging than in standard cognitive-diagnostic models. This study compares maximum likelihood estimation (MLE), maximum a posteriori (MAP), and expected a posteriori (EAP) under the IT-DINA model. A simulation study was conducted to evaluate the recovery of testlet effects and the accuracy of attribute classification under varying testlet lengths, numbers of testlets, and parameter settings. Furthermore, an empirical analysis using TIMSS 2023 mathematics data was carried out. Results showed that, under the studied model specifications and simulation conditions, Bayesian methods provided more stable recovery of continuous testlet effects than MLE. For attribute-pattern classification, Bayesian methods achieved higher accuracy in the baseline and moderately misspecified conditions, whereas their advantage weakened when the assumed attribute-profile prior was inconsistent with the true population distribution. In the empirical analysis, the IT-DINA model showed better relative fit than the DINA model and produced different person-level diagnostic classifications. These findings support considering Bayesian estimation in testlet-based cognitive diagnosis and highlight the importance of accounting for testlet effects in practice.

Keywords: cognitive-diagnostic models (CDMs); testlet effect; person parameter estimation

1. Introduction

Cognitive-diagnostic models (CDMs) are a class of psychometric models designed to classify examinees according to the mastery or non-mastery of multiple latent attributes (Junker & Sijtsma, 2001; Macready & Dayton, 1977; Tatsuoaka, 1983). Unlike traditional latent trait models that summarize performance using a single continuous ability parameter, CDMs provide fine-grained diagnostic feedback regarding specific knowledge structures and cognitive skills. Because of this feature, CDMs have been increasingly applied in educational assessment, intelligent tutoring systems, and personalized learning environments (Leighton & Gierl, 2007; Ravand & Baghaei, 2020). More recent developments, including the G-DINA framework and related generalized diagnostic models, have further expanded the flexibility and applicability of cognitive diagnosis by allowing more complex relationships between attributes and item responses (de la Torre, 2011; de la Torre & Chiu, 2016; von Davier, 2008). Recent applied and methodological work has also emphasized that diagnostic classifications are used in practical score reporting and implementation contexts, making the stability of person-level mastery estimates consequential for interpretation (Gane et al., 2025; Ma & de la Torre, 2020; Shi et al., 2021; Xu et al., 2024b).



Academic Editors: Yi-Hsin Chen and Xin Qiao

Received: 11 May 2026

Revised: 10 July 2026

Accepted: 13 July 2026

Published: 15 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Although CDMs are useful for fine-grained diagnosis, most standard cognitive-diagnostic models rely on the assumption of local item independence. This assumption is frequently violated in modern educational assessments that adopt testlet structures. Testlets refer to clusters of items associated with a common stimulus material, such as a reading passage, data table, scientific scenario, or contextual task (Wainer & Kiely, 1987). Because items within a testlet share contextual information, responses often exhibit local item dependence (LID), meaning that correlations among items remain even after conditioning on the measured attributes. Ignoring such dependence may distort parameter estimation, inflate measurement precision, and bias diagnostic classification (Bradlow et al., 1999; Cai, 2010; Yen, 1993). Consequently, modeling testlet effects has become an important methodological issue in large-scale assessments and cognitive-diagnostic applications (Ackerman, 1987; Fang et al., 2021; Wainer et al., 2007). Recent measurement research continues to regard testlet effects and local dependence as substantive methodological issues, because unmodeled dependence may affect not only model fit but also person-level inference. Accordingly, recent studies have examined local dependence through testlet response theory, aggregated and directional testlet effects, and broader frameworks for locally dependent test data (Itamiya, 2025; Jiao & Liao, 2021; Jin & Siu, 2025; Lim, 2025; Xu et al., 2025).

To address local item dependence in cognitive diagnosis, a series of testlet-based CDM extensions has been developed. Earlier work incorporated testlet-specific random effects into the DINA framework and related higher-order diagnostic models (Hansen, 2013; Zhan et al., 2015). Building on this line of research, the IT-DINA model further allows the testlet effect to interact with the ideal response pattern, thereby representing both the shared dependence induced by item bundles and its potential variation across latent attribute mastery states (Xu et al., 2024a). The IT-DINA model, therefore, provides a representative testlet-based DINA framework for studying person parameter estimation under local dependence. Recent methodological developments have extended testlet-based CDMs to more complex assessment settings, including attribute hierarchies, multiple-choice item structures, visual analog scaling responses, and more flexible local dependence specifications (L. Guo et al., 2024; Liu, 2026; Ma et al., 2023).

A central issue in person-level diagnostic inference is the choice of person parameter estimation method. In standard CDMs, person parameter estimation is commonly conducted using maximum likelihood estimation (MLE), maximum a posteriori estimation (MAP), and expected a posteriori estimation (EAP) (de la Torre, 2009; Templin & Hoffman, 2013; Warm, 1989). Previous studies have shown that these methods can yield different classification behavior under the DINA and related CDM frameworks, indicating that the choice of person parameter estimator is not merely a computational detail but a substantive measurement decision (Huebner & Wang, 2011; Zhou et al., 2024). The practical relevance of this issue is reflected in current CDM software, such as the GDINA R package (version 2.9.12), where MLE, MAP, and EAP are routinely implemented as alternative approaches for estimating examinee attribute profiles (Ma & de la Torre, 2020; Shi et al., 2021). At the same time, recent CDM estimation research suggests that Bayesian estimation methods may provide greater stability under sparse data conditions or certain complex latent structures because prior distributions regularize extreme estimates (Culpepper, 2015; Ma & Jiang, 2021; Sorrel et al., 2023).

In testlet-based CDMs, however, person parameters involve both discrete attribute profiles and continuous person-specific testlet effects, yielding a mixed latent structure that is substantially more complex than that of conventional CDMs (Xu et al., 2024a). Existing studies on testlet-based CDMs have primarily focused on model construction, item-parameter estimation procedures, identifiability, and model fit (L. Guo et al., 2024; Liu, 2026; Ma et al., 2023). Comparatively less attention has been devoted to the problem of

person parameter estimation, particularly the relative performance of different estimation methods for recovering continuous testlet effects and classifying attribute profiles under mixed discrete and continuous latent structures.

Evidence from conventional CDMs does not directly resolve this issue. Previous comparisons of MLE, MAP, and EAP have largely been conducted under models with purely discrete latent attribute structures (Huebner & Wang, 2011; Zhou et al., 2024). However, recent developments in testlet-based diagnostic modeling have moved toward more complex structures, including models that incorporate continuous testlet effects or other forms of local dependence (L. Guo et al., 2024; Lim, 2025; Liu, 2026; Ma et al., 2023). In these more complex testlet-based CDM settings, person-level inference may involve both discrete attribute profiles and continuous person-specific testlet effects. The IT-DINA model provides a representative framework for studying this mixed discrete–continuous estimation problem because it explicitly combines attribute profiles with individual testlet-effect parameters (Xu et al., 2024a). Similar to recent work on examinee scoring under complex latent trait models, this additional layer of person-level latent variation makes person parameter estimation a substantive measurement issue rather than a routine post-estimation step (J. Guo et al., 2023). To our knowledge, systematic evidence remains limited on how MLE, MAP, and EAP perform in jointly recovering attribute profiles and testlet effects under such testlet-based CDM structures.

The present study contributes to the literature by systematically implementing and comparing MLE, MAP, and EAP for person parameter estimation under the IT-DINA framework. The comparison addresses two complementary aspects of person-level inference in testlet-based CDMs, including continuous testlet-effect recovery and discrete attribute-pattern classification. Through simulation analyses, we examine how estimator performance varies across practically relevant conditions, including testlet length, number of testlets, parameter settings, and model misspecification. In addition, TIMSS 2023 eighth-grade mathematics data are analyzed to illustrate the application of the proposed comparison in an operational assessment context and to examine the practical consequences of ignoring testlet effects in cognitive diagnosis.

The remainder of this paper is structured as follows. Section 2 presents the testlet-based CDMs considered, including DINA, T-DINA, and IT-DINA, and describes the three person-parameter estimation methods, namely MLE, MAP, and EAP, with detailed implementation in the IT-DINA framework. Section 3 describes the simulation study design, evaluation criteria, and results across multiple simulation conditions. Section 4 provides an empirical illustration using TIMSS 2023 mathematics data. Section 5 concludes with a discussion of findings, implications, and limitations.

2. Method

In this section, we first introduce the cognitive-diagnostic models considered in this study, which incorporate testlet effects. We then describe the three person-parameter estimation methods evaluated in the subsequent analyses. Finally, we provide a detailed exposition of the estimation implementation under the IT-DINA framework, which involves a mixed latent structure comprising discrete attribute profiles and continuous testlet effects.

2.1. Testlet-Based Cognitive-Diagnostic Models Considered in This Study

We begin by establishing the general notation used throughout this paper. Suppose a test is administered to N examinees. The test consists of S testlets, where testlet s contains J_s items, yielding a total of $J = \sum_{s=1}^S J_s$ items. These J items are designed to measure K binary attributes. For examinee i , the latent attribute profile is denoted by a K -dimensional binary vector $\alpha_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iK})$, where $\alpha_{ik} = 1$ indicates mastery of attribute k and

$\alpha_{ik} = 0$ indicates non-mastery. The relationship between items and attributes is specified by a $J \times K$ Q-matrix, where element $q_{jk} = 1$ if attribute k is required to answer item j correctly, and $q_{jk} = 0$ otherwise.

The ideal response for examinee i on item j , denoted η_{ij} , indicates whether the examinee possesses all attributes necessary for the item. Under the conjunctive assumption, it is defined as $\eta_{ij} = \prod_{k=1}^K (\alpha_{ik})^{q_{jk}}$. Let Y_{ij} denote the observed response of examinee i to item j , with $Y_{ij} = 1$ for a correct response and $Y_{ij} = 0$ otherwise. The item parameters s_j and g_j represent the slipping and guessing probabilities, respectively.

2.1.1. DINA Model

The deterministic inputs, noisy “and” gate (DINA) model (Junker & Sijtsma, 2001) is a foundational and widely used cognitive-diagnostic model. It assumes a conjunctive condensation rule, where an examinee must master all required attributes to have a high probability of success. The item response function (IRF) is given by

$$P(Y_{ij} = 1 \mid \alpha_i) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}. \quad (1)$$

Let $\mathbf{y}_i = (y_{i1}, \dots, y_{iJ})$ denote the vector of observed responses for examinee i . Assuming conditional independence of items given the attribute profile, the likelihood of the response vector is

$$L(\mathbf{y}_i \mid \alpha_i) = \prod_{j=1}^J P(Y_{ij} = y_{ij} \mid \alpha_i). \quad (2)$$

2.1.2. Testlet Effect DINA (T-DINA) Model

The standard DINA model, like many cognitive-diagnostic models, assumes local item independence. This assumption is often violated in testlet-based assessments, where items are grouped around a common stimulus. To account for the resulting local dependence, the T-DINA model was developed within the broader family of testlet-based cognitive-diagnostic models (Hansen, 2013; Zhan et al., 2015). The T-DINA model incorporates a testlet-specific random effect ξ_{is} for examinee i on testlet s , which captures the shared variance among items within the testlet that is not explained by the measured attributes α_i . The random effect is typically assumed to follow a zero-mean normal distribution and to be independent of the attribute profile.

For item j belonging to testlet s , the IRF is expressed using a link function $h(\cdot)$ (e.g., the logistic or probit link):

$$P(Y_{ij} = 1 \mid \alpha_i, \xi_{is}) = h(\delta_{0j} + \delta_{1j}\eta_{ij} + \beta_{js}\xi_{is}), \quad (3)$$

where δ_{0j} and δ_{1j} are item-specific intercept and slope parameters, and β_{js} is the factor loading representing the strength of the testlet effect on item j .

2.1.3. Interactive Testlet DINA (IT-DINA) Model

The T-DINA model assumes that the testlet effect is additive and independent of the attribute profile. However, it is plausible that the magnitude of the testlet effect varies depending on whether the examinee possesses the attributes required by the items within the testlet. The IT-DINA model (Xu et al., 2024a) relaxes this assumption by introducing an interaction term that allows the testlet effect to differ according to the ideal response pattern η_{ij} .

The IT-DINA model can be derived as a special case of a more general formulation. Specifically, by allowing the variance of the testlet effect to depend on the ideal response, the IRF for the IT-DINA model is

$$P = h(\delta_{0j} + \delta_{1j}\eta_{ij} + \beta_{js}\xi_{is1} + \gamma_{js}(2\eta_{ij} - 1)\xi_{is2}) \quad (4)$$

where $\xi_{is} = (\xi_{is1}, \xi_{is2})'$ is a bivariate testlet-effect vector, often assumed to follow a bivariate standard normal distribution. The parameter γ_{js} modulates the interaction effect. In contrast to the T-DINA model, which includes a testlet-specific random effect independent of the attribute profile, the IT-DINA model additionally allows the testlet effect to vary with ideal response patterns through an interaction term, thereby capturing more nuanced local dependence structures.

2.2. Person Parameter Estimation Methods

In this study, we focus on the estimation of person parameters, which in the context of testlet-based CDMs consist of two distinct components: the discrete attribute profile α_i and the continuous testlet-effect vector ξ_i . The former is the target of attribute-pattern classification, while the latter is the focus of testlet-effect estimation. The presence of this mixed latent structure—discrete classes for attributes and continuous factors for testlets—makes person parameter estimation more challenging than in standard CDMs, where only the discrete attribute profile needs to be inferred. We evaluate three widely used estimation methods: Maximum Likelihood Estimation, Maximum a Posteriori estimation, and Expected A Posteriori estimation.

Let $\lambda_i = (\alpha_i, \xi_i)$ denote the complete vector of person parameters for examinee i , where ξ_i collects all testlet effects for that examinee (e.g., $\xi_i = (\xi_{i1}, \dots, \xi_{iS})$ in T-DINA, or a stacked vector in IT-DINA). The observed response vector is $\mathbf{y}_i$.

2.2.1. Maximum Likelihood Estimation (MLE)

The MLE of the person parameters is obtained by maximizing the likelihood function. For examinee i , the log-likelihood is

$$\begin{aligned} \ell(\lambda_i | \mathbf{y}_i) &= \log L(\lambda_i | \mathbf{y}_i) \\ &= \sum_{j=1}^J [y_{ij} \log P_{ij} + (1 - y_{ij}) \log(1 - P_{ij})], \end{aligned} \quad (5)$$

where $P_{ij} = P(Y_{ij} = 1 | \lambda_i)$ is the IRF from the chosen model. The MLE is

$$\hat{\lambda}_i^{\text{MLE}} = \arg \max_{\lambda_i} \ell(\lambda_i | \mathbf{y}_i). \quad (6)$$

2.2.2. Maximum a Posteriori (MAP) Estimation

MAP estimation incorporates prior information about the person parameters. Let $\pi(\lambda_i)$ denote the joint prior density. The MAP estimator maximizes the log-posterior density:

$$\hat{\lambda}_i^{\text{MAP}} = \arg \max_{\lambda_i} \{\ell(\lambda_i | \mathbf{y}_i) + \log \pi(\lambda_i)\}. \quad (7)$$

In this study, we assume the following independent priors: a uniform prior over the 2^K possible attribute profiles for α_i , and a standard normal prior for the continuous testlet effects ξ_i . Thus, $\log \pi(\lambda_i) = \log \pi(\alpha_i) + \log \pi(\xi_i)$, where $\log \pi(\alpha_i)$ is a constant and $\log \pi(\xi_i) \propto -\frac{1}{2}\xi_i'\xi_i$.

2.2.3. Expected A Posteriori (EAP) Estimation

The EAP estimator is defined as the posterior mean of the person parameters:

$$\hat{\lambda}_i^{\text{EAP}} = \mathbb{E}[\lambda_i | \mathbf{y}_i] = \int \lambda_i p(\lambda_i | \mathbf{y}_i) d\lambda_i, \quad (8)$$

where the posterior density is $p(\lambda_i | \mathbf{y}_i) \propto L(\lambda_i | \mathbf{y}_i)\pi(\lambda_i)$. Because λ_i contains both discrete and continuous components, the EAP estimates for α_i and ξ_i are obtained separately.

2.3. Estimation Implementation in the IT-DINA Framework

This subsection details the specific computational procedures for MLE, MAP, and EAP within the IT-DINA model, which features a mixed latent structure of discrete attribute profiles and continuous testlet effects. The implementation follows a two-stage approach: in the first stage, the continuous testlet effects are estimated (or integrated out) conditional on a given attribute profile; in the second stage, the optimal attribute profile is selected by comparing the resulting likelihoods or posteriors.

Let α_c for $c = 1, \dots, C$ denote the $C = 2^K$ possible attribute patterns. For a fixed pattern α_c , we define the conditional log-likelihood as $\ell(\xi_i | \mathbf{y}_i, \alpha_c) = \log P(\mathbf{y}_i | \xi_i, \alpha_c)$.

2.3.1. MLE Implementation

- Step 1: Continuous optimization for testlet effects. For each latent class c , we obtain the MLE of the testlet-effect vector ξ_i by maximizing the conditional log-likelihood:

$$\hat{\xi}_i^{(c)} = \arg \max_{\xi_i} \ell(\xi_i | \mathbf{y}_i, \alpha_c). \quad (9)$$

This is an unconstrained continuous optimization problem. In our implementation, we employ gradient ascent with backtracking line search. At iteration m , the update is

$$\xi_i^{(m+1)} = \xi_i^{(m)} + \gamma^{(m)} \nabla_{\xi} \ell(\xi_i^{(m)} | \mathbf{y}_i, \alpha_c), \quad (10)$$

where the step size $\gamma^{(m)}$ is chosen adaptively to ensure a sufficient increase in the log-likelihood.

- Step 2: Discrete selection of attribute profile. The estimated attribute profile is the one that yields the maximum log-likelihood when paired with its corresponding optimal testlet-effect estimates:

$$\hat{\alpha}_i^{\text{MLE}} = \arg \max_{\alpha_c} \ell(\hat{\xi}_i^{(c)} | \mathbf{y}_i, \alpha_c). \quad (11)$$

The final testlet-effect estimate is then set to $\hat{\xi}_i^{\text{MLE}} = \hat{\xi}_i^{(\hat{e})}$. Details are provided in Algorithm 1.

Algorithm 1 MLE Person Parameter Estimation under IT-DINA

```

1: Require: Examinee response  $\mathbf{y}_i$ , fixed item parameters, Q-matrix, attribute number  $K$ ,
    $C = 2^K$  latent attribute patterns; backtracking hyperparameters  $\rho = 0.8$ ,  $c_1 = 10^{-4}$ ;
   max iter  $M = 500$ , conv threshold  $\epsilon = 0.001$ , random restart count  $J_{max} = 5$ 
2: Enumerate all latent attribute patterns  $\{\alpha_1, \alpha_2, \dots, \alpha_C\}$ 
3: for  $c = 1, \dots, C$  do
4:   for  $r = 1, \dots, J_{max}$  do
5:     if  $r = 1$  then
6:       Initialize  $\xi_{i,r}^{(0)} = \mathbf{0}$ 
7:     else
8:       Sample  $\xi_{i,r}^{(0)} \sim N(\mathbf{0}, 0.25\mathbf{I})$ 
9:     end if
10:    Set iteration counter  $m \leftarrow 0$ 
11:    while  $m < M$  do
12:      Compute gradient  $\nabla_{\xi} \ell(\xi_{i,r}^{(m)} \mid \mathbf{y}_i, \alpha_c)$ 
13:      if  $\|\nabla_{\xi} \ell\| < \epsilon$  then
14:        Break
15:      end if
16:      Set initial step  $\gamma = 1$ , backtrack to satisfy Armijo condition
17:      Update:  $\xi_{i,r}^{(m+1)} = \xi_{i,r}^{(m)} + \gamma \nabla_{\xi} \ell$ 
18:       $m \leftarrow m + 1$ 
19:    end while
20:    Store converged  $\hat{\xi}_{i,r}^{(c)}$  and conditional log-likelihood  $\ell_r$ 
21:  end for
22:  Select restart with maximum likelihood:  $\hat{\xi}_i^{(c)} = \arg \max_{r=1 \dots J_{max}} \ell_r$ 
23:  Save objective value  $\ell_c = \ell(\hat{\xi}_i^{(c)} \mid \mathbf{y}_i, \alpha_c)$ 
24: end for
25: Optimal attribute class:  $\hat{c} = \arg \max_{c=1 \dots C} \ell_c$ ,  $\hat{\alpha}_i^{\text{MLE}} = \alpha_{\hat{c}}$ 
26: Corresponding testlet estimate:  $\hat{\xi}_i^{\text{MLE}} = \hat{\xi}_i^{(\hat{c})}$ 
27: Output: Binary attribute profile  $\hat{\alpha}_i^{\text{MLE}}$ , continuous testlet effect  $\hat{\xi}_i^{\text{MLE}}$ 

```

2.3.2. MAP Implementation

The MAP procedure mirrors the MLE structure but adds the log-prior term to the objective function.

- Step 1: Continuous optimization with prior. For each latent class c , we maximize the log-posterior with respect to ξ_i :

$$\hat{\xi}_i^{(c)} = \arg \max_{\xi_i} \{ \ell(\xi_i \mid \mathbf{y}_i, \alpha_c) + \log \pi(\xi_i) \}. \quad (12)$$

The gradient update includes contributions from both the log-likelihood and the log-prior:

$$\begin{aligned} \xi_i^{(m+1)} = \xi_i^{(m)} + \gamma^{(m)} [& \nabla_{\xi} \ell(\xi_i^{(m)} \mid \mathbf{y}_i, \alpha_c) \\ & + \nabla_{\xi} \log \pi(\xi_i^{(m)})]. \end{aligned} \quad (13)$$

- Step 2: Discrete selection with prior. The optimal attribute profile is chosen by comparing the log-posterior values across classes:

$$\begin{aligned} \hat{\alpha}_i^{\text{MAP}} = \arg \max_{\alpha_c} \{ & \ell(\hat{\xi}_i^{(c)} \mid \mathbf{y}_i, \alpha_c) \\ & + \log \pi(\hat{\xi}_i^{(c)}) + \log \pi(\alpha_c) \}. \end{aligned} \quad (14)$$

Under a uniform prior on α , $\log \pi(\alpha_c)$ is constant and can be omitted. Details are provided in Algorithm 2.

Algorithm 2 MAP Person Parameter Estimation under IT-DINA

```

1: Require:  $\mathbf{y}_i$ , item params,  $Q$ -matrix,  $K, C = 2^K$ ; prior: uniform  $\pi(\alpha)$ , standard normal  $\pi(\xi)$ ; hyperparams  $\rho = 0.8, c_1 = 10^{-4}, M = 500, \epsilon = 0.001, J_{max} = 5$ 
2: Enumerate all latent attribute patterns  $\{\alpha_1, \alpha_2, \dots, \alpha_C\}$ 
3: for  $c = 1, \dots, C$  do
4:   for  $r = 1, \dots, J_{max}$  do
5:     if  $r = 1$  then
6:       Initialize  $\xi_{i,r}^{(0)} = \mathbf{0}$ 
7:     else
8:       Sample  $\xi_{i,r}^{(0)} \sim N(\mathbf{0}, 0.25I)$ 
9:     end if
10:     $m \leftarrow 0$ 
11:    while  $m < M$  do
12:      Gradient of log-posterior:  $\nabla_{\xi} [\ell + \log \pi(\xi)]$ 
13:      if  $\|\nabla_{\xi} [\ell + \log \pi]\| < \epsilon$  then
14:        Break
15:      end if
16:      Backtracking line search for step  $\gamma$ 
17:       $\xi_{i,r}^{(m+1)} = \xi_{i,r}^{(m)} + \gamma \nabla_{\xi} [\ell + \log \pi(\xi)]$ 
18:       $m \leftarrow m + 1$ 
19:    end while
20:    Store converged  $\hat{\xi}_{i,r}^{(c)}$  and log-posterior  $\mathcal{P}_r = \ell + \log \pi(\hat{\xi}_{i,r}^{(c)})$ 
21:  end for
22:  Pick best restart:  $\hat{\xi}_i^{(c)} = \arg \max_r \mathcal{P}_r$ 
23:  Save  $\mathcal{P}_c = \ell(\hat{\xi}_i^{(c)}) + \log \pi(\hat{\xi}_i^{(c)})$ 
24: end for
25: Select attribute pattern with maximum posterior:  $\hat{c} = \arg \max_c \mathcal{P}_c, \hat{\alpha}_i^{\text{MAP}} = \alpha_{\hat{c}}$ 
26: Testlet estimate:  $\hat{\xi}_i^{\text{MAP}} = \hat{\xi}_i^{(\hat{c})}$ 
27: Output: Binary attribute profile  $\hat{\alpha}_i^{\text{MAP}}$ , regularized testlet effect  $\hat{\xi}_i^{\text{MAP}}$ 

```

2.3.3. EAP Implementation

The EAP estimator avoids optimization over ξ_i and instead relies on numerical integration. The implementation yields separate estimates for the discrete and continuous components.

- Step 1: Marginal likelihood computation. For each latent class c , we compute the marginal likelihood of the response vector by integrating over the testlet-effect space:

$$P(\mathbf{y}_i | \alpha_c) = \int P(\mathbf{y}_i | \xi_i, \alpha_c) \pi(\xi_i) d\xi_i. \quad (15)$$

In practice, this integral is approximated using Gauss–Hermite quadrature. The posterior probability of attribute profile c is then

$$P(\alpha_c | \mathbf{y}_i) = \frac{P(\mathbf{y}_i | \alpha_c) \pi(\alpha_c)}{\sum_{d=1}^C P(\mathbf{y}_i | \alpha_d) \pi(\alpha_d)}. \quad (16)$$

- Step 2: EAP estimate of the discrete attribute profile. For each attribute k , the posterior probability of mastery is obtained by summing the posterior probabilities of all latent classes that possess attribute k :

$$\begin{aligned}\hat{\alpha}_{ik}^{\text{EAP}} &= P(\alpha_{ik} = 1 \mid \mathbf{y}_i) \\ &= \sum_{c=1}^C \mathbb{I}(\alpha_{ck} = 1) P(\alpha_c \mid \mathbf{y}_i),\end{aligned}\quad (17)$$

where $\mathbb{I}(\cdot)$ is the indicator function. For each attribute, the posterior mastery probability was obtained by summing posterior probabilities over all latent classes containing that attribute.

- Step 3: EAP estimate of the continuous testlet effects. The posterior mean of the testlet-effect vector is given by

$$\begin{aligned}\hat{\boldsymbol{\xi}}_i^{\text{EAP}} &= \int \boldsymbol{\xi}_i p(\boldsymbol{\xi}_i \mid \mathbf{y}_i) d\boldsymbol{\xi}_i \\ &= \frac{\sum_{c=1}^C P(\alpha_c \mid \mathbf{y}_i) \int \boldsymbol{\xi}_i p(\mathbf{y}_i \mid \boldsymbol{\xi}_i, \alpha_c) \pi(\boldsymbol{\xi}_i) d\boldsymbol{\xi}_i}{\sum_{c=1}^C P(\mathbf{y}_i \mid \alpha_c) \pi(\alpha_c)}.\end{aligned}\quad (18)$$

This can be computed as a weighted average of the conditional expectations $\mathbb{E}[\boldsymbol{\xi}_i \mid \mathbf{y}_i, \alpha_c]$, where the weights are the posterior class probabilities $P(\alpha_c \mid \mathbf{y}_i)$. Details are provided in Algorithm 3.

Algorithm 3 EAP Person Parameter Estimation under IT-DINA

- 1: **Require:** $\mathbf{y}_i$, item params, Q -matrix, $K, C = 2^K$; priors $\pi(\alpha), \pi(\boldsymbol{\xi})$; 20 Gauss–Hermite nodes per dimension (total $R^2 = 400$ tensor product quadrature points)
 - 2: Precompute all attribute patterns $\{\alpha_1, \dots, \alpha_C\}$ and GHQ nodes/weights $\{\boldsymbol{\xi}_q, w_q\}_{q=1}^{R^2}$
 - 3: **for** $c = 1, \dots, C$ **do**
 - 4: Initialize marginal likelihood $M_c \leftarrow 0$, conditional moment $E_c \leftarrow \mathbf{0}$
 - 5: **for** $q = 1, \dots, R^2$ **do**
 - 6: Compute item response likelihood $L_{cq} = P(\mathbf{y}_i \mid \boldsymbol{\xi}_q, \alpha_c)$
 - 7: $M_c \leftarrow M_c + w_q \cdot L_{cq}$
 - 8: $E_c \leftarrow E_c + w_q \cdot \boldsymbol{\xi}_q \cdot L_{cq}$
 - 9: **end for**
 - 10: Marginal likelihood: $P(\mathbf{y}_i \mid \alpha_c) = M_c$
 - 11: Conditional expectation $\mathbb{E}[\boldsymbol{\xi} \mid \mathbf{y}_i, \alpha_c] = E_c / M_c$
 - 12: **end for**
 - 13: **for** $c = 1, \dots, C$ **do**
 - 14: Posterior class weight: $P(\alpha_c \mid \mathbf{y}_i) = \frac{P(\mathbf{y}_i \mid \alpha_c) \pi(\alpha_c)}{\sum_{d=1}^C P(\mathbf{y}_i \mid \alpha_d) \pi(\alpha_d)}$
 - 15: **end for**
 - 16: **for** each attribute $k = 1, \dots, K$ **do**
 - 17: Mastery posterior: $\hat{\alpha}_{ik}^{\text{EAP}} = \sum_{c=1}^C \mathbb{I}(\alpha_{ck} = 1) \cdot P(\alpha_c \mid \mathbf{y}_i)$
 - 18: Binary classification: $\hat{\alpha}_{ik}^{\text{binary}} = \mathbb{I}(\hat{\alpha}_{ik} \geq 0.5)$
 - 19: **end for**
 - 20: Posterior mean testlet effect: $\hat{\boldsymbol{\xi}}_i^{\text{EAP}} = \sum_{c=1}^C P(\alpha_c \mid \mathbf{y}_i) \cdot \mathbb{E}[\boldsymbol{\xi} \mid \mathbf{y}_i, \alpha_c]$
 - 21: **Output:** Continuous mastery posterior probabilities, binary attribute profile $\hat{\alpha}_i^{\text{binary}}$, posterior mean testlet $\hat{\boldsymbol{\xi}}_i^{\text{EAP}}$
-

In summary, this section has presented the DINA, T-DINA, and IT-DINA models and has provided a unified description of three person-parameter estimation methods within the IT-DINA framework. The estimation procedures explicitly separate the enumeration over discrete latent classes for attribute profile determination from the optimization or integration over continuous testlet effects. These methods form the analytical foundation for the simulation study and empirical application that follow.

3. Simulation Study

3.1. Design

We conducted a simulation study to compare three estimation approaches—MLE, MAP, and EAP—for the IT-DINA model. To provide a structured comparison of person parameter estimation methods, five simulation cases were considered. Cases 1–4 were designed to examine estimator performance under correctly specified models, overparameterized fitting, and model simplification. Case 5 was conducted as a robustness condition involving higher attribute dimensionality, a non-uniform attribute distribution, and mixed testlet loadings. Across all five cases, we compared MLE, MAP, and EAP under varying testlet lengths and numbers of testlets. The number of testlets was set to either three or four, with the number of items J_s in each testlet being either 10, 20, 30, 40, or 50; the sample size was 10,000. The Q-matrix was constructed from an existing cognitive-diagnostic assessment item pool as described in Xu et al. (2024a). Cases 1–4 used $K = 3$ and a uniform attribute distribution as a baseline consistent with the three-attribute empirical application, whereas Case 5 relaxed this baseline using $K = 5$ and a non-uniform attribute distribution.

Following Xu et al. (2024a), we used conventional DINA slipping and guessing values as reference quantities for deriving the item intercept and slope parameters. Specifically, s_j and g_j were first sampled from $U(0.05, 0.20)$, and the corresponding DINA-scale item parameters were obtained as

$$\delta_{0j} = \Phi^{-1}(g_j), \quad \delta_{1j} = \Phi^{-1}(1 - s_j) - \Phi^{-1}(g_j).$$

These values were used to set the baseline item parameters before introducing the testlet-effect loadings.

After the testlet-effect loadings were incorporated, the model-implied marginal probabilities of correct response for masters and non-masters depend on δ_{0j} , δ_{1j} , β_{js} , γ_{js} , and the variance of the person-specific testlet effects. Specifically, with $\xi_1, \xi_2 \sim N(0, \sigma_\xi^2)$, the model-implied quantities are

$$c_{js}^* = \Phi\left(\frac{\delta_{0j} + \delta_{1j}}{\sqrt{1 + \sigma_\xi^2(\beta_{js}^2 + \gamma_{js}^2)}}\right), \quad (19)$$

$$g_{js}^* = \Phi\left(\frac{\delta_{0j}}{\sqrt{1 + \sigma_\xi^2(\beta_{js}^2 + \gamma_{js}^2)}}\right), \quad (20)$$

where $\sigma_\xi^2 = 0.25$ in the present simulation. Thus, the implied slipping and guessing quantities, $1 - c_{js}^*$ and g_{js}^* , are not constrained to remain exactly within the original $U(0.05, 0.20)$ interval. To make the resulting item-quality conditions explicit, we report the true model-implied slipping and guessing quantities after reparameterization in Table 1.

Table 1. Range of model-implied guessing and slipping probabilities across five cases.

Case	β_{js}	γ_{js}	g_{js}^* Range	$1 - c_{js}^*$ Range
Case 1	0.5	0.5	[0.0605, 0.2137]	[0.0605, 0.2137]
Case 2	0.0	0.0	[0.0500, 0.2000]	[0.0500, 0.2000]
Case 3	1.0	0.0	[0.0706, 0.2257]	[0.0706, 0.2257]
Case 4	0.5	1.0	[0.0755, 0.2312]	[0.0755, 0.2312]
Case 5	± 0.5	± 1.0	[0.0755, 0.2312]	[0.0755, 0.2312]

Note: Case 5 uses mixed positive and negative loadings. Because the scaling factor depends on $\beta_{js}^2 + \gamma_{js}^2$, the signs of the loadings do not change the implied probability ranges; therefore, the ranges are identical to those in Case 4.

The loading values β_{js} and γ_{js} were set to 0, 0.5, or 1 to represent absent, moderate, and relatively strong testlet-effect components. Both ξ_1 and ξ_2 were generated from $\mathcal{N}(0, 0.5^2)$. This variance was chosen to introduce moderate person-specific testlet heterogeneity while avoiding a setting in which testlet effects dominate the attribute effects (Xu et al., 2024a). Because the same data-generating specification was used when comparing MLE, MAP, and EAP, this choice provides a common basis for evaluating the relative performance of the three estimators.

The five cases were specified as follows:

- Case 1: $\gamma_{js} = 0.5$, $\beta_{js} = 0.5$, $K = 3$, and attribute profiles were generated from a discrete uniform distribution, i.e., $\pi(\alpha) = 1/8$ for all $\alpha \in \{0, 1\}^3$. This case represents the full IT-DINA setting and was used to evaluate the relative performance of MLE, MAP, and EAP under the baseline correctly specified IT-DINA condition.
- Case 2: $\gamma_{js} = 0$, $\beta_{js} = 0$, $K = 3$, and $\pi(\alpha) = 1/8$ for all $\alpha \in \{0, 1\}^3$. This case represents an overparameterized fitting condition in which data contain no testlet effects but are analyzed using the full IT-DINA model.
- Case 3: $\gamma_{js} = 0$, $\beta_{js} = 1$, $K = 3$, and $\pi(\alpha) = 1/8$ for all $\alpha \in \{0, 1\}^3$. This case represents a partially overparameterized fitting condition in which data contain testlet main effects but no interaction effects, whereas the fitted IT-DINA model includes both components.
- Case 4: $\gamma_{js} = 1$, $\beta_{js} = 0.5$, $K = 3$, and $\pi(\alpha) = 1/8$ for all $\alpha \in \{0, 1\}^3$. This case was designed to examine the robustness of the three methods when the fitted model omits the interaction-related component of the testlet effect. In this scenario, data were generated from the full IT-DINA model, but parameter estimation was carried out using the reduced T-DINA model (which contains only the β -related testlet effect ξ_1).
- Case 5: Each testlet contains two γ_{js} and two β_{js} with negative values, with $|\gamma_{js}| = 1$ and $|\beta_{js}| = 0.5$. The number of attributes is set to $K = 5$, and the population probability π follows a non-uniform distribution as specified by Culpepper (2015): the probability of belonging to a latent class with three or more mastered skills is twice as large as the probability of belonging to a class with zero, one, or two mastered skills. Specifically, $\pi_c = 1/48$ for all attribute profiles α_c with two or fewer skills, and $\pi_c = 1/24$ for all α_c with three or more skills.

For MAP and EAP, the same non-informative prior specifications were used across all cases for comparability: a uniform prior over the 2^K attribute profiles for α , and independent standard normal priors for the continuous testlet effects.

Table 2 provides a concise summary of the five simulation cases, including the true generating model, the fitted model, and the presence or absence of testlet main effects and interaction effects.

Table 2. Summary of Simulation Cases.

Case	True Model	Fitted Model	Testlet Effect	Interaction Effect	Purpose
1	IT-DINA	IT-DINA	Yes	Yes	Correctly specified full IT-DINA condition
2	DINA	IT-DINA	No	No	Overparameterized fitting when testlet effects are absent
3	T-DINA	IT-DINA	Yes	No	Overparameterized fitting when interaction effects are absent
4	IT-DINA	T-DINA	Yes	Omitted in fitted model	Robustness to model simplification
5	IT-DINA	IT-DINA	Yes (mixed-sign loadings)	Yes (mixed-sign loadings)	Robustness under negative loadings, more attributes, and non-uniform population distribution

Across all five cases, the number of testlets and the number of items per testlet were systematically varied to examine the effects of testlet structure on estimation performance. The three estimation methods were evaluated in terms of both continuous testlet-effect recovery and discrete attribute-pattern classification.

3.2. Evaluation Criteria

The evaluation criteria were organized to mirror the two components of the person parameter vector $\lambda_i = (\alpha_i, \xi_i)$: continuous testlet effects and discrete attribute profiles.

3.2.1. Recovery of Continuous Testlet Effects

To evaluate the recovery of continuous testlet effects, we computed the average bias (Bias) and root mean squared error (RMSE) across all examinees. Let ξ_i denote the true testlet-effect vector for examinee i , and $\hat{\xi}_i$ its estimate. The indices were calculated as

$$\text{Bias}(\hat{\xi}) = \frac{1}{N} \sum_{i=1}^N (\hat{\xi}_i - \xi_i), \quad (21)$$

$$\text{RMSE}(\hat{\xi}) = \sqrt{\frac{1}{N} \sum_{i=1}^N \|\hat{\xi}_i - \xi_i\|^2}. \quad (22)$$

3.2.2. Recovery of Discrete Attribute Profiles

For the discrete attribute profiles, classification performance was assessed using the pattern-wise correct classification rate (PCCR), defined as the proportion of examinees for whom the entire estimated attribute profile matches the true profile:

$$\text{PCCR} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{\alpha}_i = \alpha_i), \quad (23)$$

where $\mathbb{I}(\cdot)$ is the indicator function. PCCR requires every attribute to be correctly recovered simultaneously and therefore provides a stringent criterion for profile-level classification accuracy.

For MLE and MAP, $\hat{\alpha}_i$ is the discrete maximizer defined in Equations (12) and (15), and its entries are already binary. For EAP, the attribute-level output is the posterior mastery probability $P(\alpha_{ik} = 1 \mid \mathbf{y}_i)$ defined in Equation (18); a binary estimate was obtained by applying a 0.50 cutoff,

$$\hat{\alpha}_{ik}^{\text{EAP}} = \mathbb{I}(P(\alpha_{ik} = 1 \mid \mathbf{y}_i) \geq 0.50), \quad (24)$$

and the resulting $\hat{\alpha}_i^{\text{EAP}}$ was then used to compute PCCR. This thresholding step ensures that the three estimators are evaluated on the same binary scale and that PCCR values are directly comparable across methods.

All simulation conditions were independently replicated $R = 500$ times. For each replication, new attribute profiles, person-specific testlet effects, and item response data were generated, while the true item-parameter settings were held fixed within each simulation condition. For PCCR, simulation variability was summarized using the Monte Carlo standard error (MCSE) of the average PCCR across replications:

$$\widehat{\text{MCSE}}(\overline{\text{PCCR}}) = \frac{\text{SD}(\text{PCCR}^{(1)}, \dots, \text{PCCR}^{(R)})}{\sqrt{R}}, \quad (25)$$

where $PCCR^{(r)}$ denotes the PCCR obtained from the r -th replication, and $SD(\cdot)$ denotes the sample standard deviation across the R replications. When PCCR is reported as a percentage, the corresponding MCSE is also reported in percentage points.

3.3. Results

3.3.1. Results Under the Correctly Specified IT-DINA Setting (Case 1)

Under the full IT-DINA setting (Case 1), Figures 1 and 2 present the estimation bias and RMSE for the testlet-effect parameters ζ_1 and ζ_2 across three and four testlets, respectively. The figures address the question of how well the three estimation methods recover the continuous testlet effects under the most complex latent structure considered in this study.

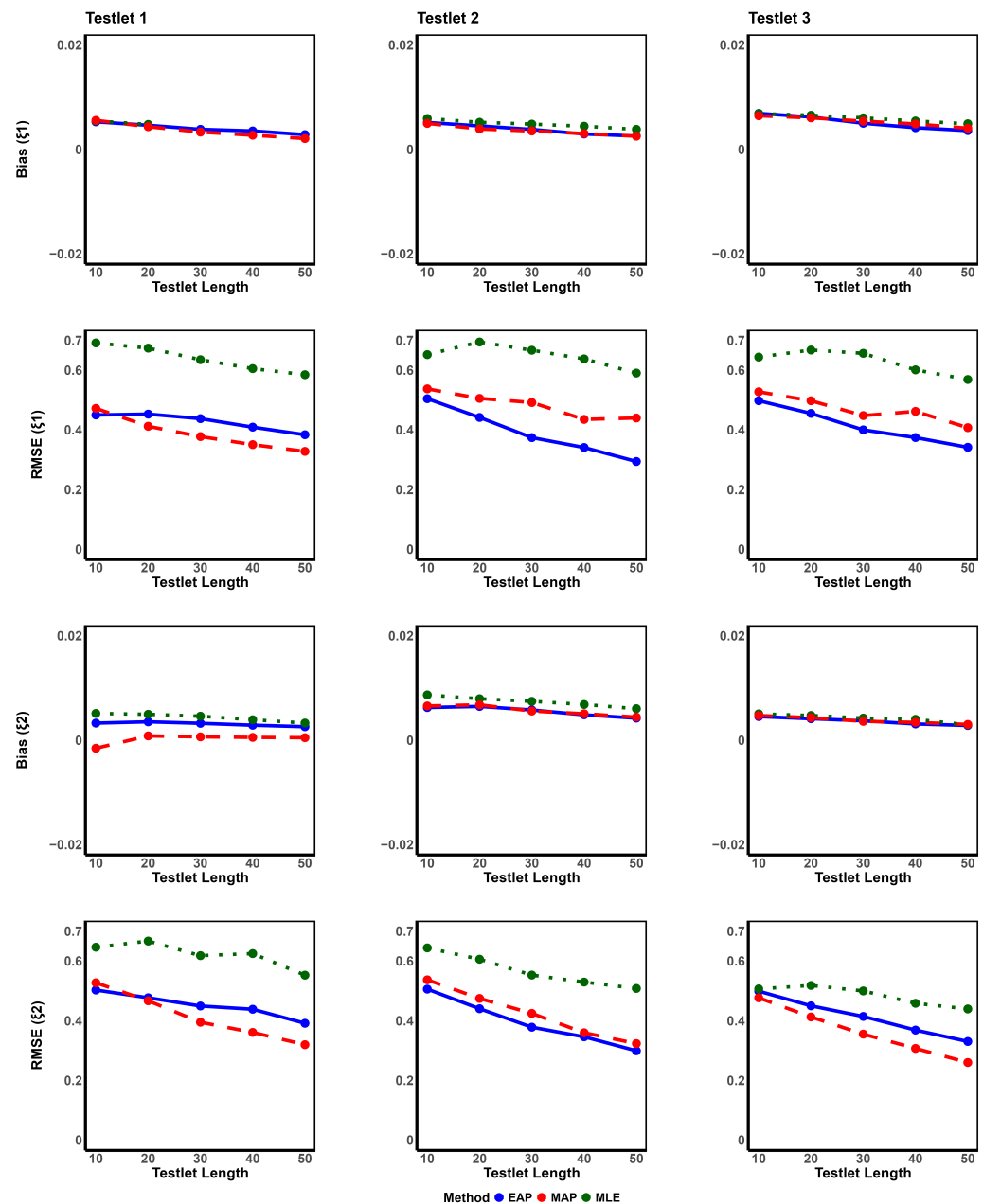


Figure 1. Bias and RMSE of ζ_1 and ζ_2 by Estimation Method for Three Testlets (Case 1).

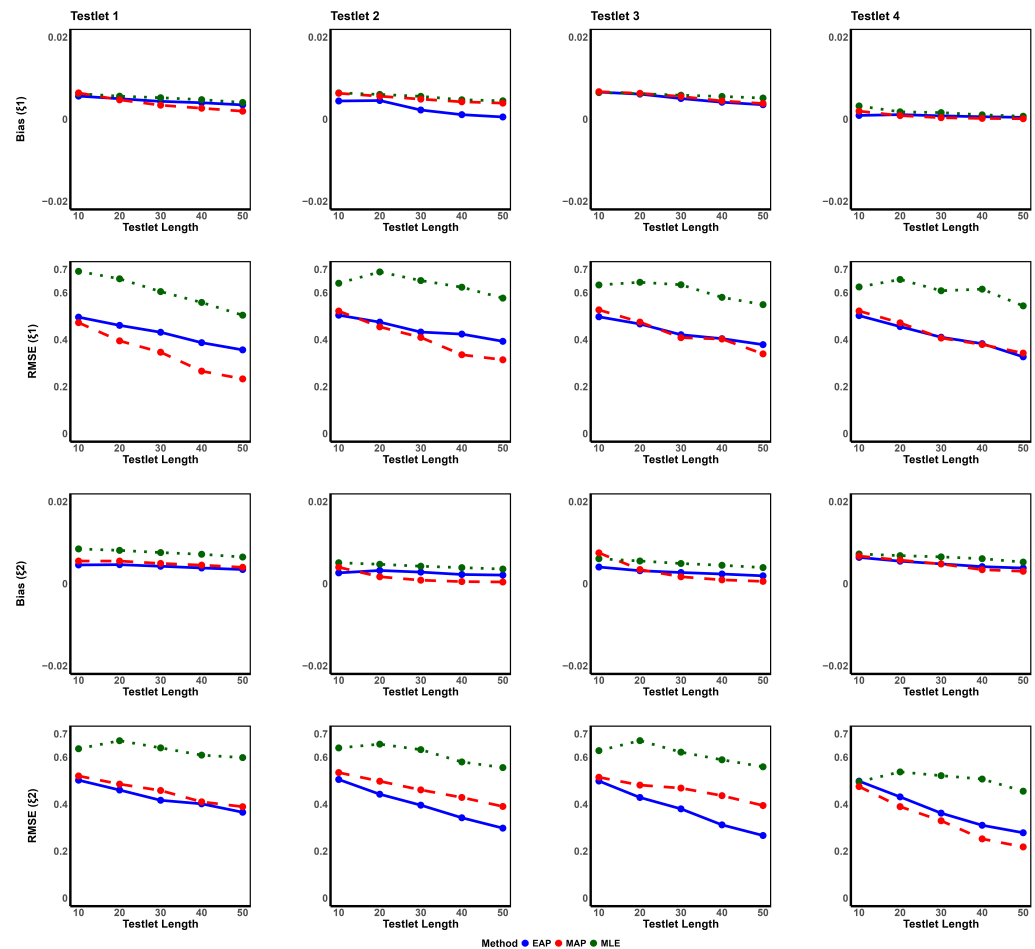


Figure 2. Bias and RMSE of ζ_1 and ζ_2 by Estimation Method for Four Testlets (Case 1).

In terms of bias, the biases of EAP and MAP remain consistently close to zero (absolute value less than 0.005) and show no clear systematic variation with changes in the number of items per testlet. In contrast, the bias of MLE exhibits noticeable fluctuations. When there are only 10 items per testlet, the absolute bias of MLE for some testlet-effect parameters reaches 0.02 to 0.04. As testlet length increases, the bias decreases but remains around 0.01 even with 50 items, which is approximately 5 to 10 times that of the Bayesian methods. Regarding RMSE, the RMSE of EAP and MAP decreases from about 0.35 to 0.20 as testlet length increases, with the two methods performing very similarly. For MLE, the RMSE is as high as 0.55 to 0.65 for short testlets (10 items), about 1.6 to 1.9 times that of the Bayesian methods. Even when testlet length increases to 50 items, the RMSE of MLE remains stable at 0.30 to 0.35, higher than that of EAP and MAP. This gap indicates that in the presence of testlet effects, Bayesian methods appear to leverage prior information to constrain estimation, yielding somewhat higher accuracy.

Tables 3 and 4 report the pattern-wise classification accuracy for all cases. Under Case 1, the Bayesian methods (EAP and MAP) achieved classification accuracy rates consistently above 97%, rising to approximately 98.5% as testlet length reached 50 items. MLE, by contrast, produced accuracy rates between 93% and 95%, approximately 2 to 4 percentage points lower than the Bayesian methods. These results suggest that, under the correctly specified IT-DINA condition, EAP and MAP showed higher recovery accuracy than MLE for both continuous testlet effects and attribute-pattern classification.

Table 3. Pattern-wise Correct Classification Rate (PCCR, %) of Three Estimation Methods under Three Testlets.

Case	Method	Testlet Length				
		10	20	30	40	50
Case1	EAP	97.65	97.69	97.98	98.06	98.50
	MAP	97.68	97.77	97.91	97.99	98.01
	MLE	93.34	93.57	94.10	94.44	94.59
Case2	EAP	99.10	99.59	99.75	99.87	99.96
	MAP	99.43	99.55	99.35	99.71	99.91
	MLE	96.30	95.89	95.77	97.89	96.93
Case3	EAP	97.10	97.59	98.05	98.80	98.96
	MAP	97.43	98.55	98.35	99.21	98.21
	MLE	95.30	95.89	94.77	95.89	95.91
Case4	EAP	96.07	97.14	97.57	97.79	97.88
	MAP	94.22	95.54	96.10	96.55	96.82
	MLE	90.12	91.50	92.47	93.16	93.67
Case5	EAP	87.15	88.69	88.58	90.06	91.50
	MAP	88.58	88.67	89.81	89.98	90.51
	MLE	90.44	90.57	91.10	91.74	92.59

Table 4. Pattern-wise Correct Classification Rate (PCCR, %) of Three Estimation Methods under Four Testlets.

Case	Method	Testlet Length				
		10	20	30	40	50
Case1	EAP	97.71	97.86	97.95	98.16	98.47
	MAP	98.54	98.57	98.46	98.83	98.01
	MLE	94.52	94.48	94.17	94.74	95.55
Case2	EAP	99.26	99.67	99.80	99.88	99.97
	MAP	99.36	99.46	99.58	99.69	99.85
	MLE	96.68	95.99	96.31	97.29	97.71
Case3	EAP	97.30	97.49	98.55	98.90	98.91
	MAP	97.53	98.25	98.45	99.20	99.36
	MLE	95.24	95.47	94.89	95.72	96.71
Case4	EAP	96.53	97.05	97.34	97.60	97.87
	MAP	94.22	95.51	96.09	96.55	96.83
	MLE	90.15	91.56	92.38	93.08	93.59
Case5	EAP	88.43	88.52	88.57	90.56	91.53
	MAP	89.48	89.63	89.84	89.92	90.48
	MLE	90.24	90.47	91.18	91.48	92.67

3.3.2. Results When DINA-Generated Data Were Fitted by IT-DINA (Case 2)

In Case 2, responses were generated from the DINA model without testlet effects but were fitted using the full IT-DINA model. Thus, this case represents an overparameterized fitting condition in which the fitted model contains additional testlet-effect components that are absent from the data-generating model. Figures 3 and 4 examine whether the three estimation methods can keep these redundant testlet-effect components stable when no true testlet dependence exists.

The estimation results show that under both the three-testlet and four-testlet conditions, the EAP and MAP methods yielded negligible bias (close to zero), with RMSE values clustering around 0.49–0.50, closely matching the generating standard deviation of the

parameters. This suggests that these two methods reproduce the distributional scale of the random effects well, with relatively stable estimation precision. In contrast, the MLE method shows some bias (with absolute values approximately 0.003 to 0.006), and its RMSE generally ranges from 0.55 to 0.69, higher than that of the Bayesian methods. Although the RMSE of MLE shows a decreasing trend as the number of items per testlet increases, it remains far from the level of accuracy achieved by EAP and MAP. This suggests that, when the full IT-DINA model is fitted to data without true testlet effects, Bayesian regularization helps prevent the overparameterized model from producing unstable person-level testlet-effect estimates.

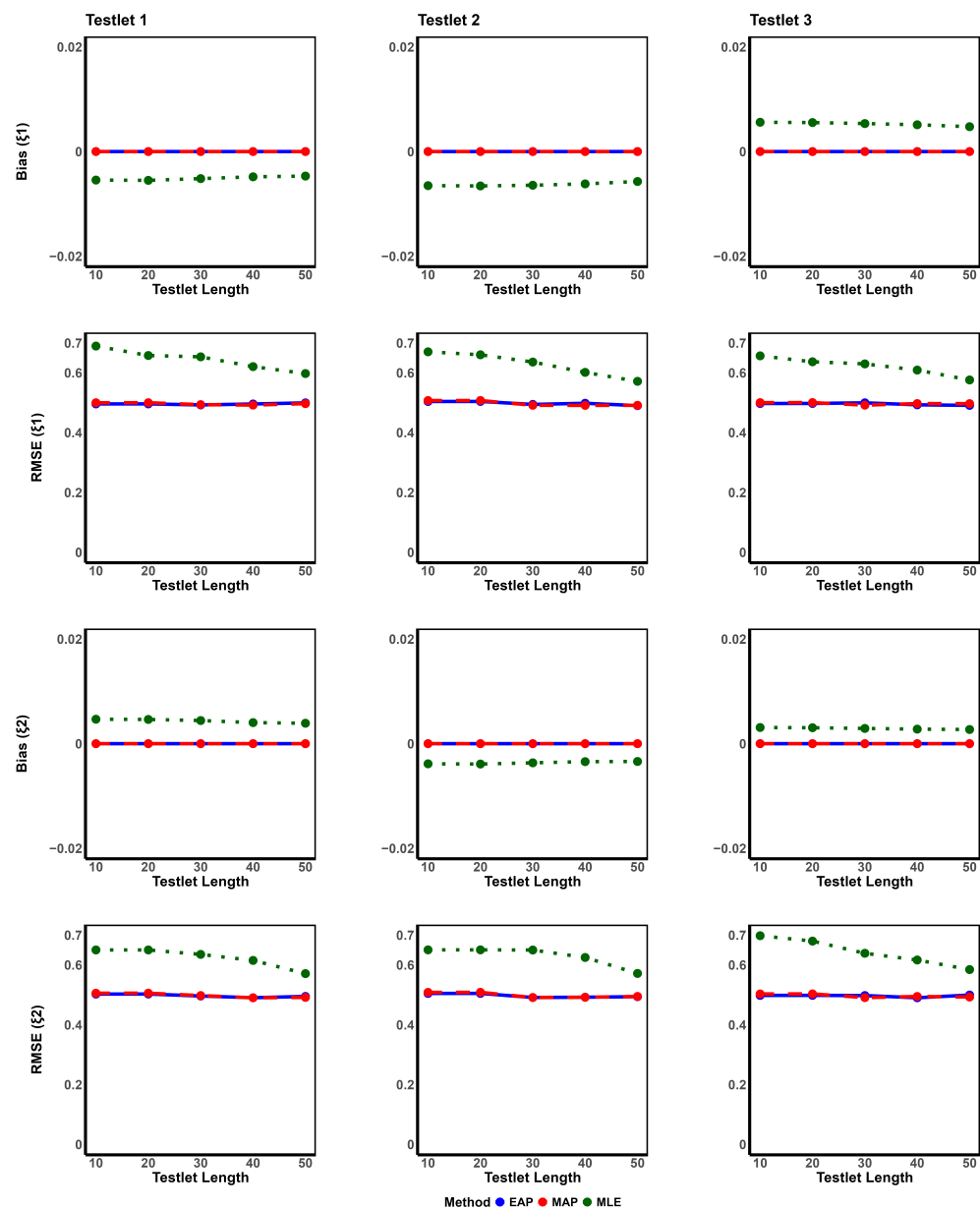


Figure 3. Bias and RMSE of ζ_1 and ζ_2 by Estimation Method for Three Testlets (Case 2).

As shown in Tables 3 and 4, attribute-pattern classification accuracy under Case 2 was the highest among the five cases. The Bayesian methods approached nearly perfect classification as testlet length increased, whereas MLE remained slightly less accurate and less stable, especially under shorter testlets. These findings indicate that even when testlet effects are absent from the data-generating process, EAP and MAP can maintain stable

estimates under the overparameterized IT-DINA fitting framework. In contrast, MLE is more sensitive to the redundant continuous components introduced by the fitted model.

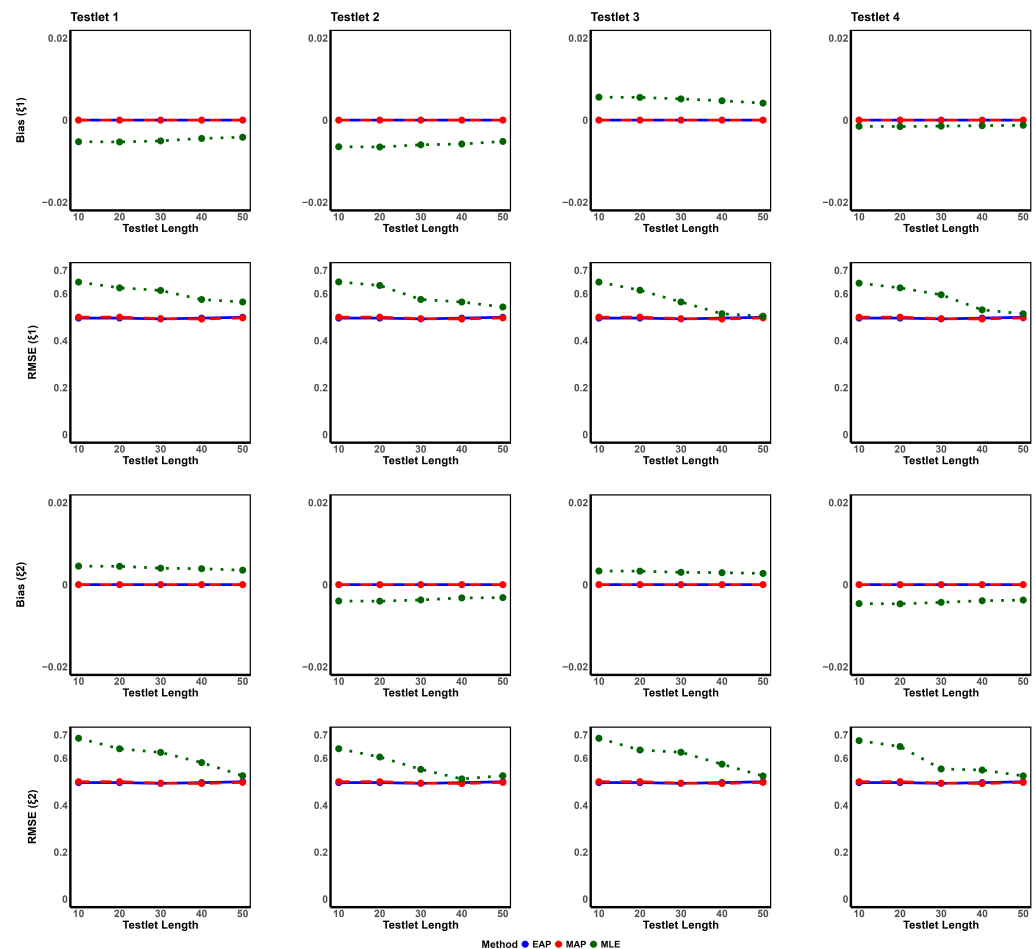


Figure 4. Bias and RMSE of ζ_1 and ζ_2 by Estimation Method for Four Testlets (Case 2).

3.3.3. Results When T-DINA-Generated Data Were Fitted by IT-DINA (Case 3)

In Case 3, responses were generated from the T-DINA model with testlet main effects only, but were fitted using the full IT-DINA model. This case, therefore, represents a partially overparameterized fitting condition: the fitted IT-DINA model includes both the main testlet-effect component and the interaction-related component, whereas the data-generating model contains only the main testlet effect. Figures 5 and 6 display the estimation results for ζ_1 and ζ_2 and examine whether the estimation methods can recover the true main testlet-effect component while controlling the unnecessary interaction-related component.

For parameters with main effects, the bias of EAP is extremely small (absolute values all below 0.001), and its RMSE decreases from approximately 0.43 to 0.34 as testlet length increases. MAP exhibits slightly larger bias than EAP (absolute values approximately 0.001 to 0.006), with RMSE ranging from 0.40 to 0.50. MLE shows the largest bias (absolute values approximately 0.004 to 0.02), with RMSE between 0.50 and 0.69. For the parameters without effects, both EAP and MAP showed negligible bias (effectively near zero), with RMSE remaining close to 0.49–0.50. In contrast, MLE estimates exhibit considerable bias (absolute values approximately 0.002 to 0.014), and their RMSE is higher than that of the Bayesian methods. This pattern indicates that Bayesian methods, especially EAP, are better able to control redundant components when the fitted model is more complex than the true data-generating model.

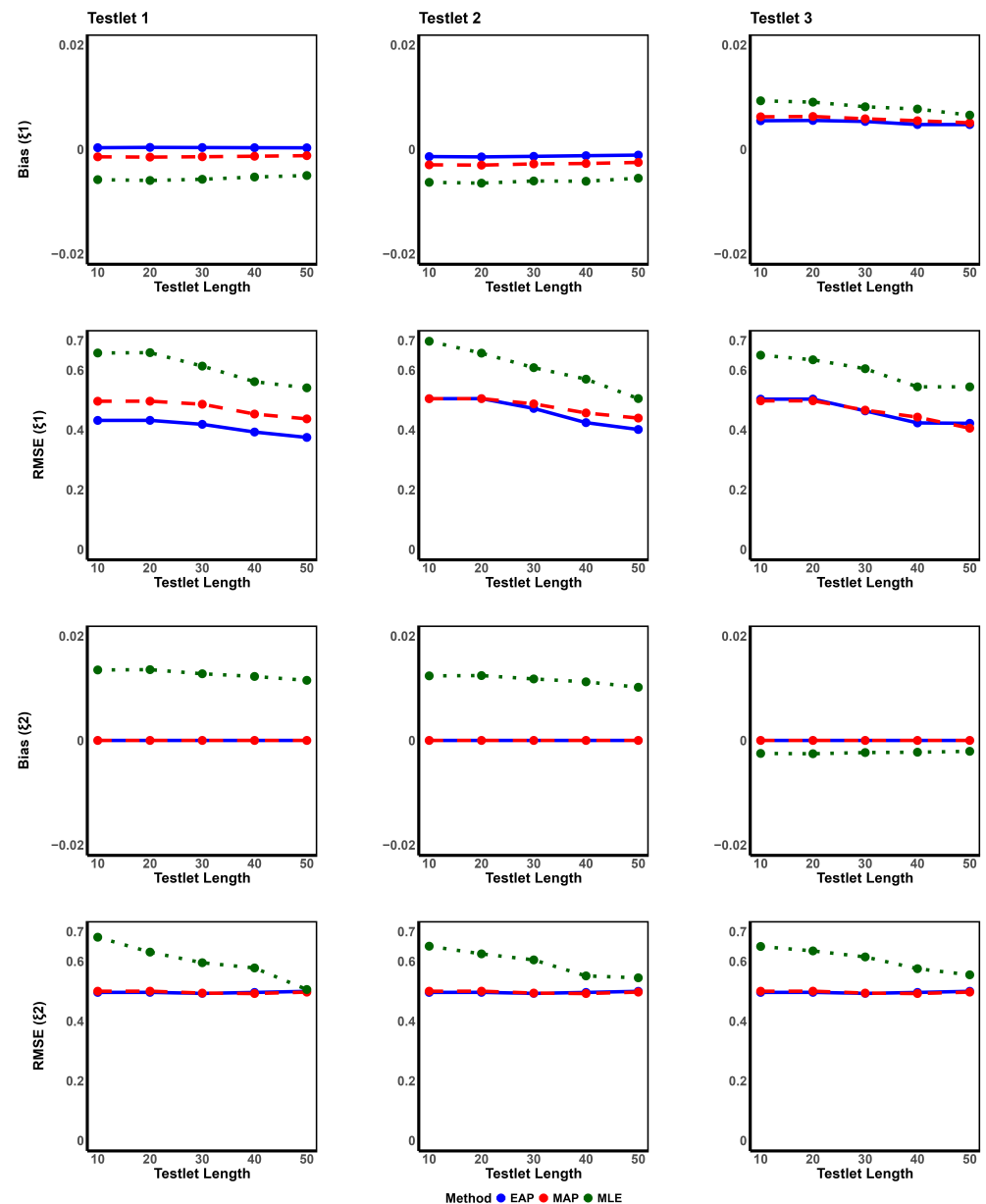


Figure 5. Bias and RMSE of ζ_1 and ζ_2 by Estimation Method for Three Testlets (Case 3).

Classification accuracy under Case 3, reported in Tables 3 and 4, fell between that of Case 1 and Case 2. EAP and MAP achieve accuracy rates from around 97% to over 99%, while MLE ranges from 95% to 96%. These results suggest that Case 3 represents an intermediate structural mismatch condition: the fitted model is more complex than the true model, but the presence of true testlet main effects makes estimation more challenging than in the DINA-generated condition of Case 2. Overall, Bayesian methods maintained a more stable recovery of both the main testlet-effect component and the discrete attribute profiles.

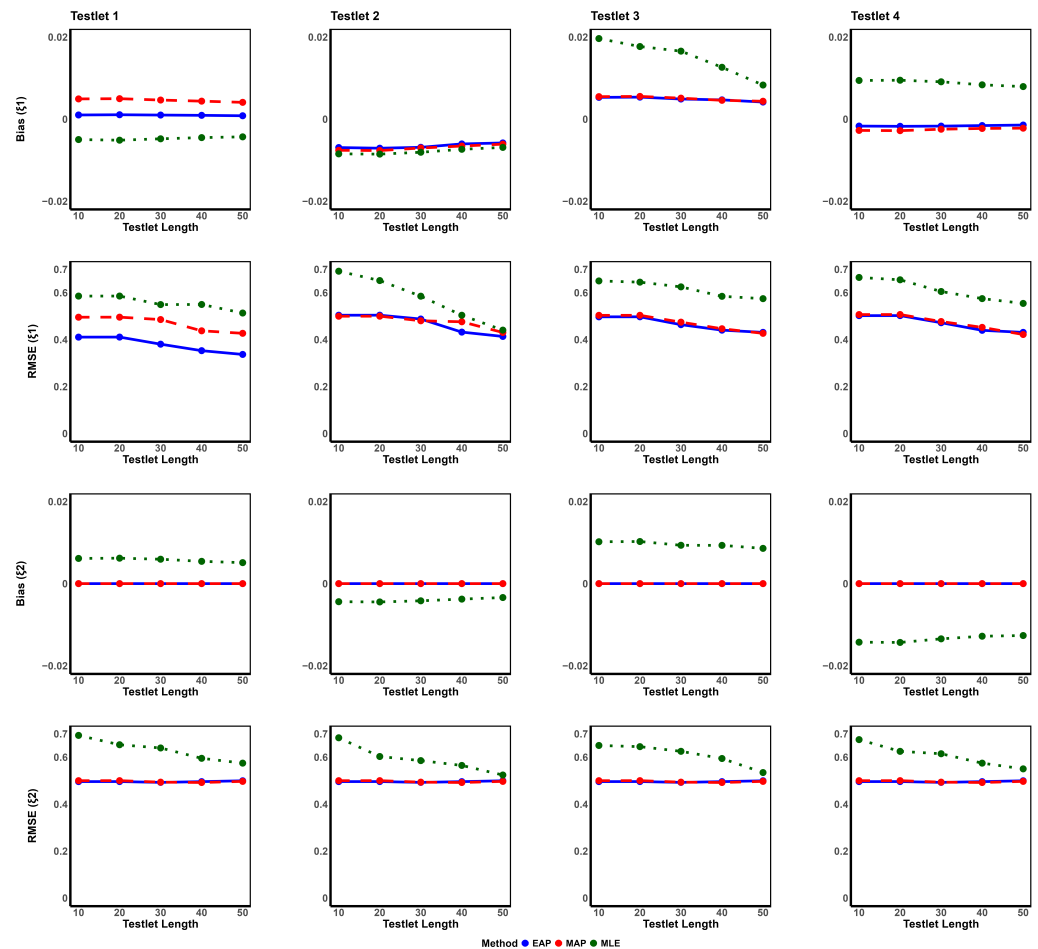


Figure 6. Bias and RMSE of ζ_1 and ζ_2 by Estimation Method for Four Testlets (Case 3).

3.3.4. Robustness Under Model Simplification (Case 4)

To evaluate robustness to model simplification, we next fitted the reduced T-DINA model to data generated from the IT-DINA model (Case 4). In this scenario, the true generating model includes both main testlet effects ($\beta_{js} = 0.5$) and interaction effects ($\gamma_{js} = 1$), but the fitted T-DINA model estimates only the β -related testlet effect ζ_1 , effectively omitting the interaction-related component.

Figures 7 and 8 present the estimation results for ζ_1 under the reduced T-DINA model.

As shown in these figures, EAP generally produced the smallest bias and RMSE under both the three-testlet and four-testlet settings. MAP showed performance close to EAP in most testlets, although a slightly larger bias was observed in some conditions. MLE showed larger fluctuations in bias and RMSE, particularly when testlets were short. These results suggest that EAP was less affected by the omission of the interaction-related random effect under the present simulation setting, whereas MLE was more sensitive to this form of model simplification.

Compared with the correctly specified IT-DINA results in Case 1, the bias and RMSE of EAP show little change, those of MAP increase slightly, while the bias of MLE increases by a factor of 2 to 5 relative to the correctly specified condition, with RMSE exceeding 0.80 for short testlets. These findings suggest that EAP was relatively less affected by the omission of the interaction-related random effect, whereas MLE was more sensitive to this form of model simplification.

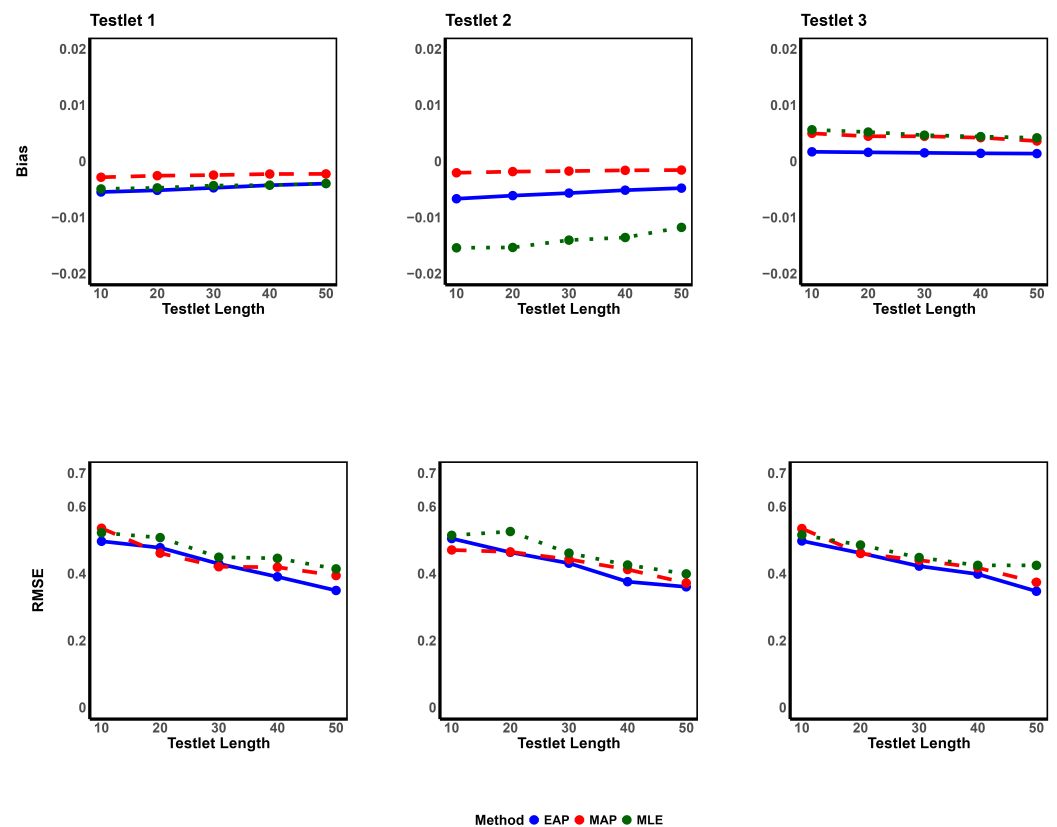


Figure 7. Bias and RMSE of ζ_1 by Estimation Method for Three Testlets (Case 4).

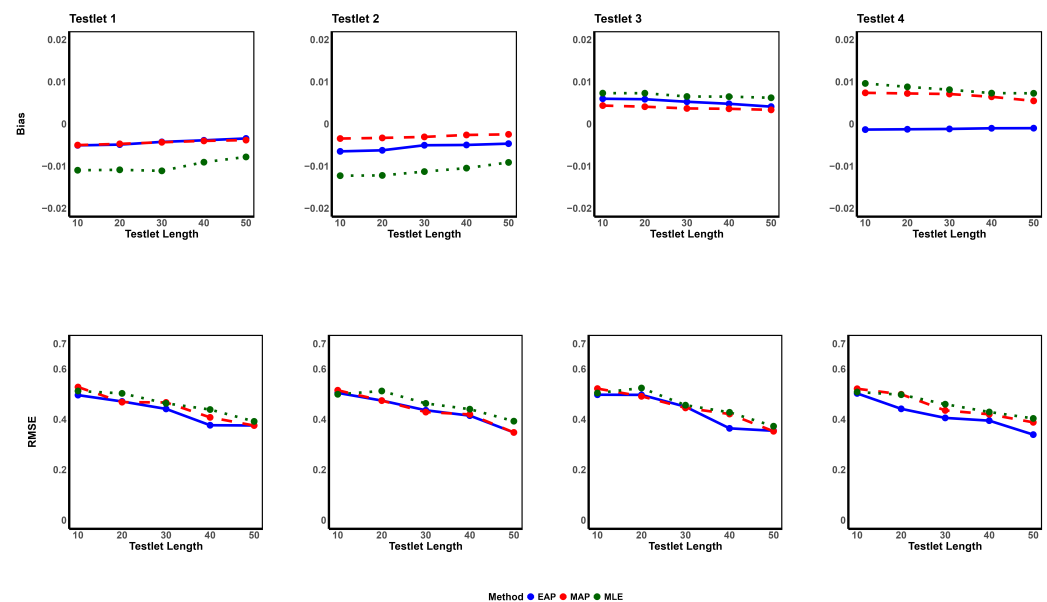


Figure 8. Bias and RMSE of ζ_1 by Estimation Method for Four Testlets (Case 4).

Classification accuracy under the reduced T DINA model (shown in Tables 3 and 4) followed a similar pattern: EAP maintained accuracy comparable to correctly specified conditions, MAP showed minor degradation, and MLE exhibited the largest drop in accuracy. For short testlets, the difference between EAP and MLE reached approximately 6–7 percentage points, suggesting that estimator choice can meaningfully affect attribute-pattern classification when the fitted model omits interaction-related components.

3.3.5. Results Under Non-Uniform Attribute Distribution and Mixed-Sign Loadings (Case 5)

This subsection presents the results for Case 5. As shown in Figures 9 and 10, EAP and MAP generally produced small biases for both ζ_1 and ζ_2 , whereas MLE showed larger RMSE values, especially when the testlet length was short. As the number of items per testlet increased, the RMSE of all three methods decreased, but MLE generally remained less accurate than EAP and MAP in recovering continuous testlet effects. These results suggest that the normal prior used by the Bayesian estimators helped stabilize continuous testlet-effect estimation under this more complex condition.

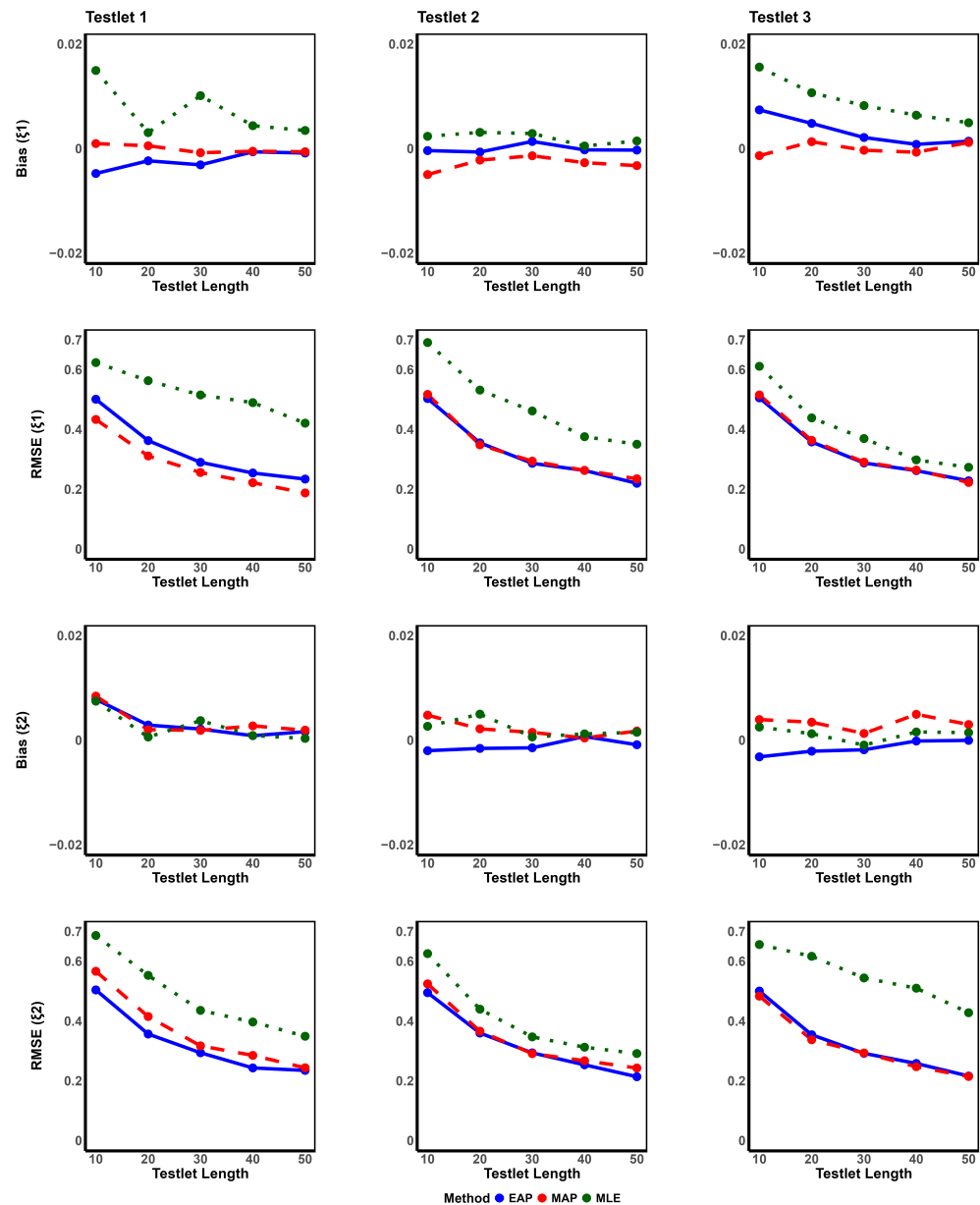


Figure 9. Bias and RMSE of ζ_1 and ζ_2 by Estimation Method for Three Testlets (Case 5).

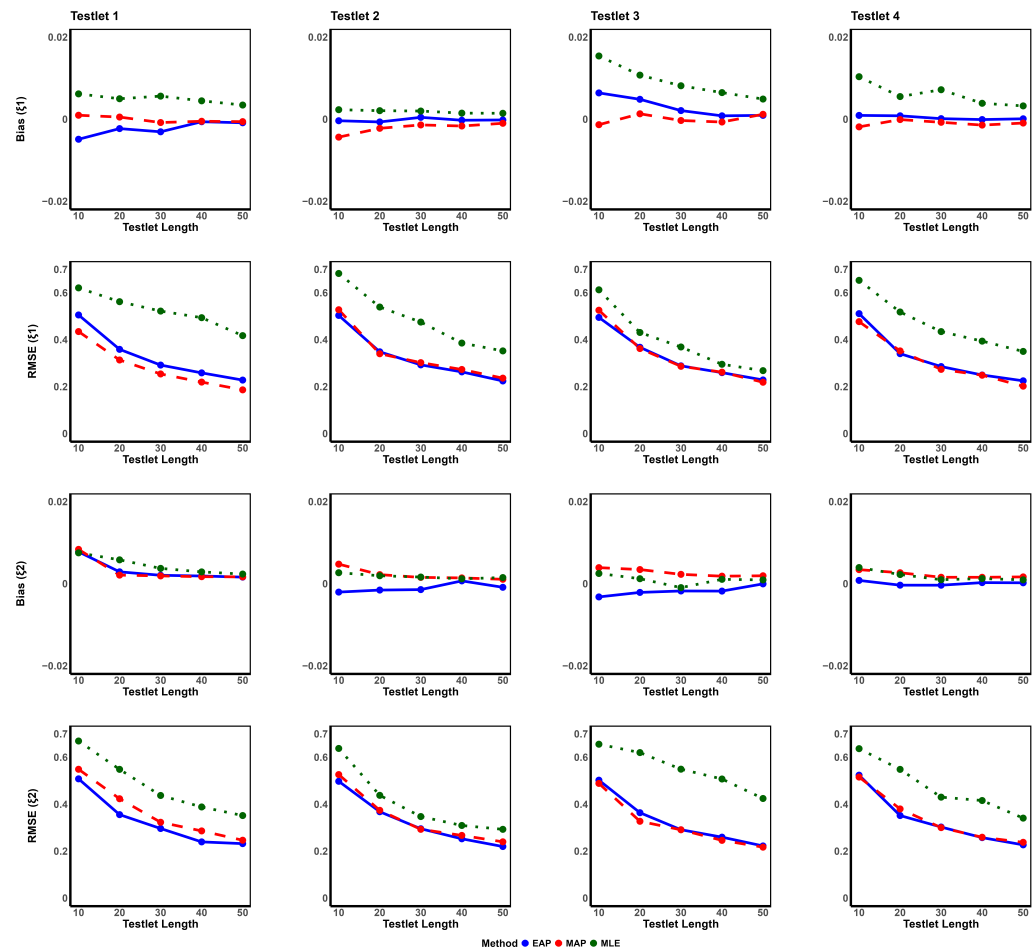


Figure 10. Bias and RMSE of ξ_1 and ξ_2 by Estimation Method for Four Testlets (Case 5).

The classification results are reported in Tables 3 and 4. The PCCR values of EAP and MAP ranged from approximately 87% to 91%, whereas MLE achieved slightly higher values, approximately 90% to 93%. The advantage of EAP and MAP in continuous testlet-effect recovery did not extend to attribute-pattern classification in Case 5. One possible explanation is the mismatch between the non-uniform true attribute distribution and the uniform attribute-profile prior used by MAP and EAP. Under this prior misspecification, Bayesian classification may be pulled away from the true population structure, whereas MLE is not directly affected by the assumed attribute-profile prior.

3.3.6. Simulation Variability of PCCR

To assess the simulation variability of the classification results, we computed Monte Carlo standard errors (MCSEs) for PCCR, as reported in Tables 5 and 6. Across all simulation conditions, the MCSE values ranged from 0.08% to 1.50%, indicating that the estimated PCCR values were stable.

Specifically, the MCSEs were generally below 1% in Cases 1–3, where classification accuracy was relatively high. In Case 4, the MCSEs increased slightly, especially for MLE, reflecting greater variability under model simplification. The largest MCSEs appeared in Case 5, ranging approximately from 1.17% to 1.50%, which is consistent with the lower classification accuracy and higher attribute dimensionality in this condition. Overall, the MCSE results indicate that the simulation estimates of PCCR were sufficiently stable for comparing the three estimation methods across the studied conditions.

Table 5. Monte Carlo Standard Errors (in %) for PCCR under Three Testlets.

Case	Method	Testlet Length				
		10	20	30	40	50
Case1	EAP	0.68	0.67	0.63	0.62	0.54
	MAP	0.67	0.66	0.64	0.63	0.63
	MLE	1.12	1.10	1.05	1.02	1.01
Case2	EAP	0.42	0.29	0.22	0.16	0.09
	MAP	0.34	0.30	0.36	0.24	0.13
	MLE	0.84	0.89	0.90	0.64	0.77
Case3	EAP	0.75	0.69	0.62	0.49	0.45
	MAP	0.71	0.54	0.57	0.40	0.59
	MLE	0.95	0.89	1.00	0.89	0.89
Case4	EAP	0.87	0.75	0.69	0.66	0.64
	MAP	1.04	0.92	0.87	0.82	0.78
	MLE	1.33	1.25	1.18	1.13	1.09
Case5	EAP	1.50	1.42	1.42	1.34	1.25
	MAP	1.42	1.42	1.35	1.34	1.31
	MLE	1.32	1.31	1.27	1.23	1.17

Table 6. Monte Carlo Standard Errors (in %) for PCCR under Four Testlets.

Case	Method	Testlet Length				
		10	20	30	40	50
Case1	EAP	0.67	0.65	0.63	0.60	0.55
	MAP	0.54	0.53	0.55	0.48	0.63
	MLE	1.02	1.02	1.05	1.00	0.92
Case2	EAP	0.38	0.26	0.20	0.15	0.08
	MAP	0.36	0.33	0.29	0.25	0.17
	MLE	0.80	0.88	0.84	0.73	0.67
Case3	EAP	0.72	0.70	0.54	0.47	0.46
	MAP	0.69	0.59	0.55	0.40	0.36
	MLE	0.95	0.93	0.99	0.90	0.80
Case4	EAP	0.82	0.76	0.72	0.68	0.65
	MAP	1.04	0.93	0.87	0.82	0.78
	MLE	1.33	1.24	1.19	1.14	1.10
Case5	EAP	1.43	1.43	1.42	1.31	1.25
	MAP	1.37	1.36	1.35	1.35	1.31
	MLE	1.33	1.31	1.27	1.25	1.17

Overall, the simulation results suggest that EAP and MAP generally improved estimation stability under the studied testlet-based CDM conditions, particularly for short testlets and model-structure mismatch. At the same time, Case 5 shows that the classification advantage of Bayesian methods may weaken when the assumed attribute-profile prior is inconsistent with the true population distribution.

4. Application

The empirical analysis serves two purposes: (a) to examine whether the relative differences among MLE, MAP, and EAP observed in the simulations also emerge in real assessment data; and (b) to illustrate the practical consequences of ignoring testlet effects in cognitive diagnosis.

4.1. Data Set

TIMSS (Trends in International Mathematics and Science Study), as one of the most influential international educational assessment programs, provides extensive cross-national comparative data on student mathematics and science literacy (Fishbein et al., 2025).

This empirical illustration utilizes student response data from the TIMSS 2023 eighth-grade mathematics assessment, obtained from the official TIMSS 2023 international database released in SPSS-format data files. The original TIMSS 2023 eighth-grade mathematics data included 308,492 students and 68 candidate mathematics items across assessment booklets. Missing values were processed by uniformly converting unanswered or invalid responses into missing indicators. Subsequently, to ensure the reliability of model estimation, students with fewer than five valid responses were excluded, resulting in a final sample of 118,547 students for further analysis.

In the empirical analysis, TIMSS item blocks were used as the operational units for defining testlets, following the TIMSS rotated booklet design (von Davier et al., 2024). Here, operational testlets refer to TIMSS item blocks used as block-level grouping units for modeling residual within-block dependence. Although TIMSS blocks are not necessarily testlets in the narrow sense of a single shared-stimulus item set, they serve as reasonable block-level units for examining whether responses to items administered within the same block exhibit residual dependence beyond the measured cognitive attributes. TIMSS mathematics blocks are organized by difficulty level. In this empirical illustration, we focused on the difficult-level eighth-grade mathematics blocks because they were expected to provide adequate response variation for diagnostic classification. Among these blocks, B_MD3 and B_MD5 were selected because they were administered to the same group of respondents in the selected booklet and retained sufficient item-level response variation after preprocessing. Accordingly, B_MD3 and B_MD5 were treated as two operational testlets, allowing person parameters to be estimated jointly across the two blocks.

To maintain a relatively homogeneous response format for this empirical illustration, the final analysis was based on the constructed-response items retained in B_MD3 and B_MD5. Because the DINA, T-DINA, and IT-DINA models considered in this study are formulated for dichotomous item responses, the original polytomous scores of these items were converted into binary responses. Specifically, non-zero scores were recoded as correct, and zero scores were recoded as incorrect. Items with no response variation after recoding were excluded because they provide little statistical information for distinguishing among attribute mastery patterns. After preprocessing and screening, the final analyzed data set contained 27 binary items, including 15 items from B_MD3 and 12 items from B_MD5.

The Q-matrix was constructed according to the official TIMSS 2023 eighth-grade mathematics cognitive-domain classification. TIMSS classifies each mathematics item according to its primary cognitive demand: Knowing, Applying, or Reasoning (von Davier et al., 2024). In the present empirical illustration, these three cognitive domains were treated as three binary diagnostic attributes. Because each retained item was assigned to exactly one cognitive domain in the TIMSS framework, the resulting Q-matrix has a simple structure in which each row contains one entry equal to 1 and no all-zero rows. Among the 27 retained items, 11 were assigned to Applying, 4 to Knowing, and 12 to Reasoning, as shown in Table 7.

Table 8 reports the pre-calibrated item-parameter estimates used in the empirical analysis. These parameters were treated as fixed when comparing MLE, MAP, and EAP for person parameter estimation. A few negative δ_1 estimates were retained as part of the pre-calibrated model and should be interpreted cautiously.

Table 7. Q-matrix Specification for TIMSS Mathematics Items.

Operational Testlet	Item	Applying	Knowing	Reasoning
B_MD3	ME82214A	0	1	0
	ME82214B	1	0	0
	MQ82N01A	1	0	0
	MQ82N01B	0	0	1
	MQ82N02	0	0	1
	MQ82N03	1	0	0
	MQ82N04	0	0	1
	MQ82N07	0	1	0
	MQ8N208A	0	0	1
	MQ8N208B	0	0	1
	ME82711	1	0	0
	ME82111	0	1	0
	ME82320	1	0	0
	ME82707	1	0	0
	ME82802	1	0	0
B_MD5	MQ82C02	0	0	1
	MQ82C03	0	1	0
	MQ82C04	1	0	0
	MQ82C05	0	0	1
	ME82116	1	0	0
	ME82322	0	0	1
	ME82199A	0	0	1
	ME82199B	0	0	1
	ME82723	0	0	1
	ME82106	0	0	1
	ME82807	1	0	0
	ME82309	1	0	0

All analyses were conducted in the R 4.2.0 environment, employing parallel computing techniques to handle the large-scale sample data. The model convergence criterion was set such that the absolute difference in parameter changes between adjacent iterations was less than 0.001, with a maximum of 500 iterations.

Table 8. Item-Parameter Estimation Results for the IT-DINA Model.

Operational Testlet	Item	δ_0	SE (δ_0)	δ_1	SE (δ_1)	Pass Rate
B_MD3	ME82214A	−3.11	0.0057	3.43	0.0051	0.14
	ME82214B	−3.25	0.0142	3.20	0.0071	0.11
	MQ82N01A	−1.73	0.0124	3.05	0.0184	0.24
	MQ82N01B	−1.61	0.0103	−2.45	0.0063	0.05
	MQ82N02	−1.32	0.0068	−2.74	0.0081	0.08
	MQ82N03	−4.06	0.0009	3.57	0.0084	0.07
	MQ82N04	−1.64	0.0108	−2.42	0.0074	0.05
	MQ82N07	−4.06	0.0093	3.88	0.0072	0.09
	MQ82N08A	−1.63	0.0106	−2.43	0.0081	0.05
	MQ82N08B	−1.92	0.0165	−2.15	0.0075	0.03
	ME82711	−3.42	0.0103	3.36	0.0071	0.11
	ME82111	−2.91	0.0092	3.08	0.0048	0.13
	ME82320	−4.06	0.0017	3.75	0.0076	0.08
	ME82707	−3.42	0.0137	3.41	0.0071	0.11
	ME82802	−4.06	0.0081	3.22	0.0076	0.04

Table 8. Cont.

Operational Testlet	Item	δ_0	SE (δ_0)	δ_1	SE (δ_1)	Pass Rate
B_MD5	MQ82C02	−3.01	0.0074	2.89	0.0090	0.07
	MQ82C03	−0.86	0.0077	−3.20	0.0077	0.15
	MQ82C04	−1.02	0.0046	−3.05	0.0068	0.12
	MQ82C05	−2.55	0.0065	2.82	0.0073	0.09
	ME82116	−1.16	0.0055	−2.91	0.0069	0.10
	ME82322	−3.25	0.0082	3.33	0.0080	0.08
	ME82199A	−2.65	0.0071	2.52	0.0071	0.07
	ME82199B	−3.11	0.0064	3.09	0.0083	0.07
	ME82723	−2.77	0.0091	2.69	0.0091	0.07
	ME82106	−2.77	0.0085	2.83	0.0065	0.08
	ME82807	−1.37	0.0073	−2.69	0.0076	0.07
	ME82309	−1.13	0.0053	−2.93	0.0078	0.10

4.2. Results

4.2.1. Comparison of Person Parameter Estimation Methods Under IT-DINA

We first compared the three estimation methods—MLE, MAP, and EAP—within the IT-DINA framework. The comparison focused on four aspects: skill pattern distributions, testlet-effect estimate distributions, rank consistency, and classification transition matrices.

Table 9 presents the frequency and percentage distribution of skill patterns estimated by the three methods. The distributions of EAP and MAP across the eight skill patterns are highly consistent, with differences in percentages for each pattern being within 0.5%. In contrast, the distribution from MLE shows noticeable deviation, with a frequency of 32,479 (27.40%) for pattern (0,0,1), which is much higher than that of EAP (15.48%) and MAP (15.66%), while frequencies for other patterns are correspondingly lower. This empirical pattern is consistent with the simulation observation that MLE may yield more dispersed or less regularized person-level estimates under complex testlet structures.

Table 9. Frequency and Percentage of Skill Patterns Estimated by Three Methods.

Skill Pattern	EAP		MAP		MLE	
	Freq	%	Freq	%	Freq	%
(0,0,0)	12,045	10.16	11,892	10.03	8273	6.98
(1,0,0)	10,342	8.72	10,101	8.52	7464	6.30
(0,1,0)	8976	7.57	8,743	7.37	6739	5.68
(0,0,1)	18,342	15.48	18,567	15.66	32,479	27.40
(1,1,0)	9543	8.05	9321	7.86	7158	6.04
(1,0,1)	14,237	12.01	14,512	12.24	12,089	10.20
(0,1,1)	12,876	10.86	13,045	11.00	11,724	9.89
(1,1,1)	32,186	27.15	32,366	27.31	32,621	27.52
Total	118,547	100	118,547	100	118,547	100

Figure 11 compares the distributions of the testlet-effect parameter estimates across the three methods. The EAP estimates are most concentrated around zero with the narrowest interquartile ranges and fewest outliers, reflecting the strong shrinkage induced by the standard normal prior. MAP estimates are slightly more dispersed but still far more stable than MLE estimates. MLE produced wider boxes, longer whiskers, and more extreme values, indicating greater dispersion in the estimated testlet effects. This pattern is consistent with the simulation results, in which Bayesian methods produced more concentrated estimates of continuous testlet effects.

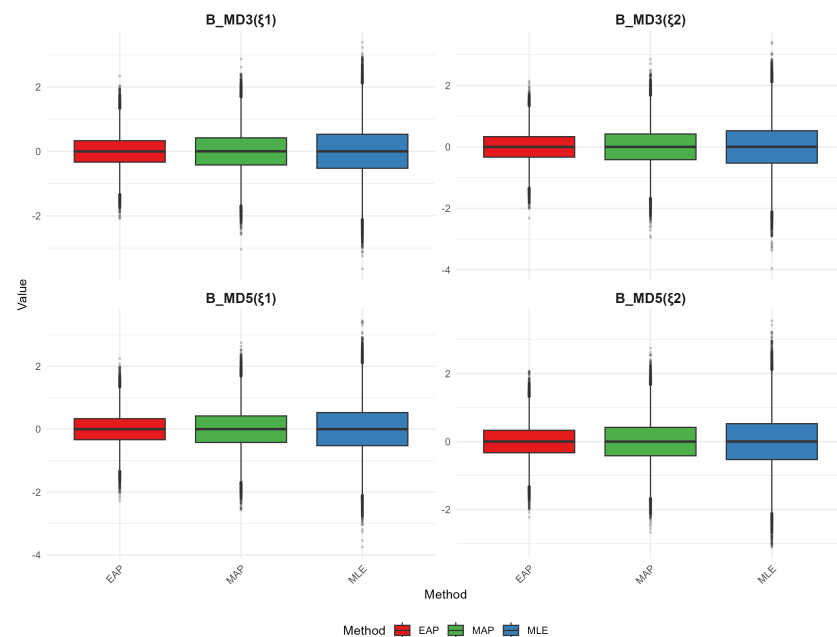


Figure 11. Boxplots of Testlet-Effect Estimates by Estimation Method.

Table 10 reports Kendall rank correlation coefficients for the testlet-effect parameter estimates. The correlations between EAP and MAP exceed 0.96, indicating that the two Bayesian methods yielded highly similar individual rankings. Correlations between EAP and MLE, and between MAP and MLE, range from 0.55 to 0.65, indicating only moderate agreement. This pattern aligns with the simulation-based observation that MLE diverges systematically from Bayesian estimates in the presence of testlet effects.

Table 10. Rank Correlations of Testlet-Effect Estimates.

Parameter	EAP-MAP	EAP-MLE	MAP-MLE
ξ_1 (T1)	0.976	0.634	0.616
ξ_1 (T2)	0.981	0.578	0.642
ξ_2 (T1)	0.969	0.561	0.655
ξ_2 (T2)	0.982	0.553	0.649

Note: T1 and T2 denote the two block-level operational testlets corresponding to TIMSS blocks B_MD3 and B_MD5, respectively.

Table 11 shows that EAP and MAP are highly consistent in classifying skill patterns, with diagonal proportions exceeding 93% for all patterns. Table 12 highlights systematic differences between EAP-based and MLE-based classifications. A substantial proportion of students classified into patterns (0,0,0), (1,0,0), (0,1,0), (1,1,0), and (1,0,1) by EAP were reclassified as pattern (0,0,1) by MLE. For instance, among the 9543 students in EAP pattern (1,1,0), 1571 (16.5%) were assigned to pattern (0,0,1) by MLE, and even larger shifts were observed from patterns (0,0,0) and (1,0,0) to (0,0,1) (24.4% and 25.2%, respectively). This pronounced shift toward a single pattern is consistent with the simulation observation that MLE may yield less regularized person-level classifications under complex testlet structures.

Table 11. Transition Matrix: EAP to MAP.

EAP	MAP							
	(0,0,0)	(1,0,0)	(0,1,0)	(0,0,1)	(1,1,0)	(1,0,1)	(0,1,1)	(1,1,1)
(0,0,0)	11,327	87	102	92	101	94	87	155
(1,0,0)	152	9731	78	85	69	58	65	104
(0,1,0)	110	61	8350	47	41	97	91	179
(0,0,1)	96	72	64	18,024	30	23	9	24
(1,1,0)	77	45	42	87	9011	86	82	113
(1,0,1)	55	46	44	76	20	13,977	7	12
(0,1,1)	47	34	41	98	21	91	12,497	47
(1,1,1)	28	25	22	58	28	86	207	31,732

Table 12. Transition Matrix: EAP to MLE.

EAP	MLE							
	(0,0,0)	(1,0,0)	(0,1,0)	(0,0,1)	(1,1,0)	(1,0,1)	(0,1,1)	(1,1,1)
(0,0,0)	6931	282	519	2937	465	302	297	312
(1,0,0)	342	6281	278	2608	235	204	184	210
(0,1,0)	291	263	5354	2379	192	173	153	171
(0,0,1)	185	156	138	17,521	98	87	76	81
(1,1,0)	196	186	174	1571	5427	583	596	810
(1,0,1)	135	122	114	2832	268	7839	503	2424
(0,1,1)	117	106	98	1758	371	1464	6429	2533
(1,1,1)	76	68	64	873	102	1437	3486	26,080

4.2.2. Practical Consequences of Ignoring Testlet Effects

To illustrate the practical importance of modeling local dependence, we compared the IT-DINA model with the traditional DINA model, which assumes local item independence. If relevant block-level operational testlet effects are ignored, person parameter estimates and diagnostic classifications may differ across model specifications. Therefore, the DINA–IT-DINA comparison illustrates the practical consequences of model misspecification for empirical classification. (The T-DINA model, which includes main testlet effects but not interactions, was not separately compared here because the empirical focus is on the contrast between ignoring testlet effects entirely and accounting for them in a flexible manner; the simulation study already covered the intermediate T-DINA case).

Table 13 presents model fit indices. The IT-DINA model yields substantially lower AIC and BIC values than the DINA model (reductions of approximately 12,000), and the likelihood ratio test is highly significant ($\Delta LL = 6,234.45$, $\Delta df = 4$, $p < 0.001$). These results provide evidence that the IT-DINA model offered better relative fit than the DINA model in this empirical example.

Table 13. Model Fit Comparison Between DINA and IT-DINA Models.

Model	Log-Likelihood (LL)	Parameters	AIC	BIC
DINA	−1,523,456.23	54	3,047,020.5	3,047,520.3
IT-DINA	−1,517,221.78	58	3,034,559.6	3,035,019.8

Note: The IT-DINA model estimates two additional testlet-effect parameters per block-level operational testlet. This application involves two operational testlets, corresponding to TIMSS blocks B_MD3 and B_MD5; hence, 4 additional parameters are estimated.

Table 14 compares the skill pattern classifications from the two models (using EAP for IT-DINA). The differences are substantial. The DINA model assigns 32,479 students

(27.40%) to pattern (0,0,1), whereas IT-DINA assigns only 18,342 students (15.48%) to this pattern—a difference of over 14,000 students. Conversely, DINA underestimates low-mastery patterns (0,0,0), (1,0,0), and (0,1,0) by a cumulative 8887 students, and underestimates multi-attribute patterns (1,1,0), (1,0,1), and (0,1,1) by approximately 5685 students. Only the full-mastery pattern (1,1,1) shows comparable frequencies across models.

Table 14. Comparison of Skill Pattern Classification Results Between DINA and IT-DINA Models.

Skill Pattern	IT-DINA (EAP)		DINA		Frequency Difference
	Frequency	Percentage	Frequency	Percentage	
(0,0,0)	12,045	10.16%	8273	6.98%	−3772
(1,0,0)	10,342	8.72%	7464	6.30%	−2878
(0,1,0)	8976	7.57%	6739	5.68%	−2237
(0,0,1)	18,342	15.48%	32,479	27.40%	14,137
(1,1,0)	9543	8.05%	7158	6.04%	−2385
(1,0,1)	14,237	12.01%	12,089	10.20%	−2148
(0,1,1)	12,876	10.86%	11,724	9.89%	−1152
(1,1,1)	32,186	27.15%	32,621	27.52%	435

These shifts suggest that, when testlet effects are ignored, local dependence may be absorbed into attribute classification, leading to different person-level diagnostic conclusions. In this empirical illustration, the DINA model assigned a much larger proportion of students to the (0,0,1) pattern than the IT-DINA model. Because many retained items were associated with the Reasoning attribute, this difference indicates that model choice can substantially affect the interpretation of attribute mastery patterns. This finding is consistent with the simulation results, where model misspecification affected person-level estimates and attribute-pattern classification.

4.3. Summary of Empirical Findings

This empirical illustration using TIMSS 2023 mathematics data yielded three main findings. First, under the IT-DINA model, EAP and MAP produced more concentrated testlet-effect estimates and more similar classification distributions than MLE, as reflected in the distributional comparisons, transition matrices, and rank correlations. Second, the empirical results were consistent with the simulation observation that Bayesian methods, especially EAP, can produce more regularized person-level estimates in complex testlet settings.

5. Discussion

This study compared MLE, MAP, and EAP for person parameter estimation in testlet-based cognitive-diagnostic models, using the IT-DINA model as the main framework. The simulation results showed that EAP and MAP generally produced more stable recovery of continuous testlet effects than MLE, especially when testlets were short or when the fitted model involved redundant or omitted testlet-effect components. For attribute-pattern classification, Bayesian methods achieved higher accuracy in the baseline and moderately misspecified conditions, whereas Case 5 showed that this advantage could weaken when the true attribute distribution was non-uniform but the Bayesian estimators used a uniform prior. The empirical TIMSS analysis provided an illustration of the proposed comparison framework, showing that the IT-DINA and DINA models produced different person-level diagnostic classifications and that the three estimation methods yielded different distributions of testlet-effect estimates. Overall, these findings suggest that Bayesian estimation can be useful for person parameter estimation in testlet-based CDMs, but its advantage depends on the model structure, prior specification, and estimation target.

The superior stability of EAP can be attributed to several interrelated factors rooted in the mixed latent structure of the IT-DINA model. First, EAP relies on posterior expectation rather than posterior mode. By averaging over the entire posterior distribution, EAP naturally applies shrinkage to individual estimates, which is particularly valuable when testlet-level information is sparse. In contrast, MLE and MAP both optimize point-valued criteria—likelihood and posterior density, respectively—making them more susceptible to local instabilities when the data provide limited evidence about the continuous testlet effects. Second, the likelihood surface in the IT-DINA model is inherently complex due to the coexistence of discrete latent classes (attribute profiles) and continuous latent variables (testlet effects). For a given attribute pattern, the conditional likelihood as a function of the testlet effects can be flat, multimodal, or poorly identified, especially with short testlets. MLE, which lacks any regularization mechanism, is particularly vulnerable to these pathologies, leading to erratic estimates and convergence failures. Bayesian methods, by incorporating a prior distribution over the testlet effects, effectively regularize the estimation problem and mitigate the influence of sampling variability. Third, the difference between MAP and EAP lies in their use of the posterior distribution. MAP selects the posterior mode, which can remain sensitive to local fluctuations when the posterior is diffuse or asymmetric. EAP, by taking the posterior mean, smooths over such irregularities and produces estimates that are more stable across replications. This distinction explains why EAP consistently exhibited smaller bias and RMSE than MAP under conditions of model-structure mismatch and short testlets.

An additional finding with implications for test design is that increasing the number of items per testlet produced larger improvements in estimation accuracy than increasing the number of testlets. This pattern suggests that, under the present simulation settings, person parameter estimation benefited more from richer within-testlet information than from a larger number of shorter testlets. However, this finding should be interpreted within the design boundaries of the simulation. The high classification rates observed in some conditions were obtained under a relatively simple diagnostic structure and favorable item-quality settings, and therefore should not be directly generalized to assessments with higher attribute dimensionality, weaker item information, or noisier response data.

The comparison between the DINA and IT-DINA models suggests that ignoring local dependence may affect person-level diagnostic classifications. In the TIMSS illustration, the two models produced markedly different distributions of attribute patterns, especially for the $(0,0,1)$ pattern. This result indicates that model choice can have substantive implications for diagnostic interpretation, even when the empirical analysis is used primarily as a model-based illustration. Therefore, the practical relevance of modeling testlet effects lies not only in improving relative model fit but also in examining how local dependence may influence person-level diagnostic conclusions.

Several limitations of the present work and directions for future research should be noted. First, the model boundary is confined to the DINA, T-DINA, and IT-DINA specifications. While these models are widely used and provide a clear basis for examining testlet effects, the findings may not automatically generalize to more flexible frameworks such as the G-DINA family of testlet-based cognitive-diagnostic models. Second, the polytomous TIMSS constructed-response items were dichotomized to fit the DINA-type model framework. This transformation discards partial-credit information and may reduce item-level measurement precision. The empirical attribute classification results and model fit statistics are therefore specific to this binary scoring scheme, and should not be directly generalized to diagnostic analyses that retain full polytomous response structures. Third, the empirical application was intended as a model-based illustration rather than a population-representative analysis. The current implementation did not explicitly incorpo-

rate complex survey weighting into model calibration or fit evaluation. For applications in which sampling weights are available and population-level inference is required, future work could incorporate survey weights and design-adjusted fit indices (Lumley & Scott, 2015). Finally, the current comparison of MLE, MAP, and EAP is contingent on the specific numerical implementations employed, including the optimization algorithm, quadrature approximations, and prior specifications. Although care was taken to use standard and well-documented procedures, alternative implementation choices could yield somewhat different quantitative results. Future research could extend the current findings by examining a broader class of testlet-based CDMs, incorporating polytomous response data directly, and investigating the sensitivity of person parameter estimates to alternative prior formulations and computational strategies.

Author Contributions: Conceptualization, X.X.; methodology, X.X. and M.Y.; software, M.Y.; validation, M.Y., J.Y. and X.X.; formal analysis, M.Y.; investigation, M.Y.; resources, X.X.; data curation, M.Y.; writing—original draft preparation, M.Y.; writing—review and editing, J.Y. and X.X.; visualization, M.Y.; supervision, X.X.; project administration, X.X.; funding acquisition, X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Fundamental Research Funds for the Central Universities.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The TIMSS 2023 international database can be obtained from the TIMSS & PIRLS International Study Center at <https://timss2023.org/data/> (accessed on 7 July 2026). The simulation results reported in this study were generated according to the model specifications and simulation conditions described in the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

- Ackerman, T. A. (1987). A comparison of multidimensional IRT and testlet response theory. *Applied Psychological Measurement*, 11(2), 89–106.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, 64(2), 153–168. [\[CrossRef\]](#)
- Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, 75(4), 581–612. [\[CrossRef\]](#)
- Culpepper, S. A. (2015). Bayesian estimation of the DINA model with higher-order structure. *Psychometrika*, 80(4), 1093–1114.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1), 115–130. [\[CrossRef\]](#)
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76(2), 179–199. [\[CrossRef\]](#)
- de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, 81(2), 253–273. [\[CrossRef\]](#) [\[PubMed\]](#)
- Fang, G., Guo, J., Xu, X., Ying, Z., & Zhang, S. (2021). Identifiability of bifactor models. *Statistica Sinica*, 31, 2309–2330.
- Fishbein, B., Taneva, M., & Kowolik, K. (2025). *TIMSS 2023 user guide for the international database*. Boston College, TIMSS & PIRLS International Study Center. Available online: <https://www.iea.nl/publications/user-guides/timss-2023-user-guide-international-database> (accessed on 7 July 2026).
- Gane, B. D., Clark, A. K., & Kavitsky, E. (2025). Parent interpretation and use of diagnostic mastery-based score reports. *Applied Measurement in Education*, 38(3–4), 363–381. [\[CrossRef\]](#)
- Guo, J., Xu, X., & Xin, T. (2023). A note on latent traits estimates under IRT models with missingness. *Journal of Educational Measurement*, 60(4), 575–625. [\[CrossRef\]](#)
- Guo, L., Zhou, W., & Li, X. (2024). Cognitive diagnosis testlet model for multiple-choice items. *Journal of Educational and Behavioral Statistics*, 49(1), 32–60. [\[CrossRef\]](#)

- Hansen, M. P. (2013). *Hierarchical item response models for cognitive diagnosis* [Doctoral dissertation, University of California].
- Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, 71(2), 407–419. [CrossRef]
- Itamiya, T. (2025). A review of methods for analyzing test data with local dependence structures: An item response theory perspective. *Japanese Journal for Research on Testing*, 21(1), 125–152.
- Jiao, H., & Liao, M. (2021). Digital module 25: Testlet response theory. *Educational Measurement: Issues and Practice*, 40(2), 110. [CrossRef]
- Jin, K.-Y., & Siu, W. L. (2025). Modeling directional testlet effects on multiple open-ended questions. *Journal of Educational Measurement*, 62(1), 81–100.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258–272. [CrossRef]
- Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.
- Lim, Y. S. (2025). Parametric bootstrap Mantel-Haenszel statistic for aggregated testlet effects. *Journal of Educational Measurement*, 62(4), 503–530. [CrossRef]
- Liu, C.-W. (2026). Inferring individual attributes using testlet-based visual analogue scaling and beta copula diagnostic classification models. *Journal of Educational and Behavioral Statistics*, 51(3), 455–496.
- Lumley, T., & Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3(1), 1–18. [CrossRef]
- Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, 93(14), 1–26. [CrossRef] [PubMed]
- Ma, W., Iaconangelo, C., & de la Torre, J. (2023). A testlet diagnostic classification model with attribute hierarchies. *Applied Psychological Measurement*, 47(3), 173–188. [CrossRef] [PubMed]
- Ma, W., & Jiang, Z. (2021). Estimating cognitive diagnosis models in small samples: Bayes modal estimation and monotonic constraints. *Applied Psychological Measurement*, 45(2), 95–111. [CrossRef]
- Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, 2(2), 99–120.
- Ravand, H., & Baghaei, P. (2020). Cognitive diagnostic modeling: Current issues and future directions. *Frontiers in Psychology*, 11, 1302. [CrossRef]
- Shi, Q., Ma, W., Robitzsch, A., Sorrel, M. A., & Man, K. (2021). Cognitively diagnostic analysis using the G-DINA model in R. *Psych*, 3(4), 812–835. [CrossRef]
- Sorrel, M. A., Escudero, S., Nájera, P., Kreitchmann, R. S., & Vázquez-Lira, R. (2023). Exploring approaches for estimating parameters in cognitive diagnosis models with small sample sizes. *Psych*, 5(2), 336–349. [CrossRef]
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4), 345–354.
- Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using MCMC: An introduction with R and WinBUGS. *Journal of Educational and Behavioral Statistics*, 38(1), 50–87. [CrossRef] [PubMed]
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61(2), 287–307.
- von Davier, M., Fishbein, B., & Kennedy, A. M. (Eds.). (2024). *TIMSS 2023 technical report: Methods and procedures*. Boston College, TIMSS & PIRLS International Study Center. Available online: <https://timss2023.org/methods/> (accessed on 7 July 2026).
- Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2007). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum Associates. [CrossRef]
- Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24(3), 185–201. [CrossRef]
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427–450. [CrossRef]
- Xu, X., Fang, G., Guo, J., Ying, Z., & Zhang, S. (2024a). Diagnostic classification models for testlets: Methods and theory. *Psychometrika*, 89(3), 851–876. [PubMed]
- Xu, X., Guo, J., & Xin, T. (2025). Inference of correlations among testlet effects: A latent variable selection method. *Applied Psychological Measurement*, 49(3), 126–155. [CrossRef]
- Xu, X., Ren, S., Zhang, D., & Xin, T. (2024b). Knowledge integration in science learning: Tracking students' knowledge development and skill acquisition with cognitive diagnosis models. *Educational Measurement: Issues and Practice*, 43(1), 66–82. [CrossRef]
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213. [CrossRef]

- Zhan, P., Li, X., Wang, W.-C., Bian, Y., & Wang, L. (2015). The multidimensional testlet-effect cognitive diagnostic models. *Acta Psychologica Sinica*, 47(5), 689–701.
- Zhou, M., Liu, Y., & Teng, Y. (2024). The choice of examinee parameter estimation methods in cognitive diagnostic models. *Journal of Psychological Science*, 47(1), 229–236.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.