ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202201.00088

Standard Error and Confidence Interval Estimation for Cognitive Diagnosis Models: Parallel Bootstrap

Authors: Liu Yanlou, Liu Yanlou

Date: 2022-01-26T18:00:03+00:00

Abstract

The standard errors (Standard Error, SE; or variance-covariance matrix) and confidence intervals (Confidence Interval, CI) of cognitive diagnosis models have important theoretical and practical value in measuring uncertainty in model parameter estimation, differential item functioning detection, model comparison at the item level, Q-matrix validation, and exploring attribute hierarchical relationships. This study proposes two novel methods for calculating SE and CI: the parallel parametric bootstrap method and the parallel nonparametric bootstrap method. Simulation studies found that when the model is correctly specified, these two methods demonstrate good performance in calculating SE and CI for model parameters under conditions of high-quality and medium-quality items; when model parameters are redundant, the SE and CI demonstrate good performance for most permissible model parameters under conditions of high-quality and medium-quality items. The value of the new methods and improvements in computational efficiency are demonstrated through empirical data.

Full Text

Preamble

Standard Errors and Confidence Intervals for Cognitive Diagnostic Models: Parallel Bootstrap Methods

LIU Yanlou
(Academy of Big Data for Education, Qufu Normal University, Jining 273165, China)

Abstract

Standard errors (SE; or variance-covariance matrices) and confidence intervals (CI) for cognitive diagnostic model parameters hold significant theoretical and practical value across multiple domains, including quantifying estimation uncertainty, detecting differential item functioning, conducting item-level model comparisons, validating Q-matrices, and exploring attribute hierarchies. This study proposes two novel methods for computing SEs and CIs: parallel parametric bootstrap and parallel non-parametric bootstrap. Simulation studies reveal that when models are correctly specified, both methods perform well in calculating SEs and CIs for model parameters under high- and medium-quality item conditions. When model parameters are redundant, both methods demonstrate good performance for most permissible model parameters under high- and medium-quality item conditions. The value of these new methods and their computational efficiency improvements are demonstrated through empirical data analysis.

Keywords: cognitive diagnostic model, standard error, confidence interval, bootstrap, parallel computing
Classification Code: B841

1 Introduction

Cognitive Diagnostic Models (CDMs), also known as Diagnostic Classification Models, represent a class of discrete latent variable models (Rupp et al., 2010) that have been widely applied in psychology, education, and biology (e.g., Tjoe & de la Torre, 2014). Latent attributes carry different meanings across domains, such as knowledge, skills, cognitive processes, mental disorders, or even pathogens (Rupp et al., 2010; Wu et al., 2017). When appropriately applied, CDMs enable researchers to infer each individual' s multidimensional latent attribute mastery status from observed response patterns, providing timely feedback, personalized guidance, or targeted remediation.

The standard error (SE) of CDM model parameters quantifies estimation uncertainty (Liu et al., 2021). In psychometric models, two model parameters with identical point estimates may exhibit different confidence intervals due to varying SEs, necessitating simultaneous consideration of point estimates and CIs. For instance, if two items in a CDM both have guessing parameter estimates of 0.2 but SE estimates of 0.08 and 0.05 respectively, their estimation precision differs. Under normal distribution theory, the first guessing parameter' s 95% CI is $[0.2\text{–}1.96\$\times 0.08, 0.2+1.96\times 0.08], whilethesecond'sis[0.2\text{–}1.96\times 0.05, 0.2+1.96\times\$0.05]$. Consequently, numerous psychology journals (e.g., *Acta Psychologica Sinica*, or see American Psychological Association, 2020) require or recommend reporting SEs and 95% CIs. However, few empirical CDM studies report model parameter SEs and CIs, primarily due to the lack of accessible computational methods.

This paper examines two commonly used SE and CI estimation approaches—

analytic methods and bootstrap methods—identifies their current limitations, and proposes a simple, feasible alternative. Model parameter SEs (or more generally, variance-covariance matrices) play a fundamental and central role in CDM inferential statistics (Liu, Xin et al., 2019; Philipp et al., 2018). Beyond CI computation, model parameter SEs are crucial for differential item functioning detection (Liu, Yin, et al., 2019; Ma et al., 2021; Liu et al., 2016), item-level model comparisons (de la Torre & Lee, 2013; Liu, Andersson, et al., 2019; Ma & de la Torre, 2016, 2019), Q-matrix validation (Ma & de la Torre, 2020a), and exploring attribute hierarchies (Liu et al., 2021; Wang & Lu, 2021).

Researchers have proposed various analytic estimation methods (Liu, Xin et al., 2019; Liu et al., 2021; Philipp et al., 2018; Liu et al., 2016), including the Empirical Cross-product Information Matrix (XPD), Observed Information Matrix (Obs), and Sandwich-type Information Matrix (Sw). Under model parameter identifiability conditions (Gu & Xu, 2020; Wang & Lu, 2021), simulation and empirical studies have investigated the performance of these analytic information matrices (Liu et al., 2016; Liu et al., 2016) for computing SEs and CIs of model parameters (including item parameters and structural parameters describing examinee distributions). Regarding item parameter SEs and CIs, researchers have compared XPD, Obs, and Sw performance under ideal conditions (perfect model-data fit) and under misspecification of CDM item response models and/or Q-matrices (Liu, Xin, et al., 2019; Philipp et al., 2018). Findings indicate that when models (including item response models and Q-matrices) are correctly specified or contain minimal misspecification, all three methods show good consistency in estimating item parameter SEs. Under severe model misspecification (e.g., simultaneous substantial errors in item response models and Q-matrices), only Sw remains robust (Liu, Xin, et al., 2019).

Regarding structural parameter SEs and CIs, studies have explored these within the Hierarchical Cognitive Diagnosis Model (HCDM; Templin & Bradshaw, 2014) framework (Liu et al., 2021). When attribute hierarchies are correctly specified (i.e., the structural model is perfectly specified), all three methods achieve good 95% CI coverage rates with sample sizes of 3,000 or more. When attributes have hierarchical relationships but a saturated CDM is used for estimation (i.e., structural model parameters are partially redundant), XPD and Obs methods perform well for permissible structural parameters (those theoretically non-zero according to attribute hierarchies), while XPD performs well for impermissible structural parameters (those theoretically equal to zero) (Liu et al., 2021).

Accurately identifying and validating attribute hierarchies in CDMs enables deeper understanding of examinees' psychological processes, holding important theoretical and practical value (Leighton et al., 2004). However, correctly prespecifying attribute hierarchies in practice is extremely challenging (Hu & Templin, 2020; Liu et al., 2021; Ma & Xu, 2021; Templin & Bradshaw, 2014; Wang & Lu, 2021). When cognitive diagnostic assessments contain attribute hierarchies, using a saturated CDM to fit response data yields structural parameters

approximately equal to zero, providing evidence for attribute hierarchies (Liu et al., 2021; Templin & Bradshaw, 2014). Liu et al. (2021) preliminarily proposed using z-statistics to explore attribute hierarchies when structural parameter SEs are known, expressed as:

$$z = \frac{\hat{\eta}}{SE(\hat{\eta})}$$

where $\hat{\eta}$ represents the structural parameter estimate and $SE(\hat{\eta})$ denotes its standard error.

While XPD, Obs, or Sw methods can effectively compute CDM parameter SEs in most cases, these analytic approaches have two major drawbacks. First, they require positive definiteness of the information matrix. DeCarlo (2011, 2019) found that boundary value problems in CDMs can cause non-positive definiteness when using information matrices to compute variance-covariance matrices. Boundary value issues will be detailed in Section 2. Second, they require positive diagonal elements in the variance-covariance matrix; negative values prevent SE computation. However, computational errors may produce negative diagonal elements when inverting information matrices (Liu & Maydeu-Olivares, 2014). For example, in the empirical data analysis in Section 5, the second structural parameter' s diagonal element in the Obs-based variance-covariance matrix is negative, preventing SE calculation. This means that scenario (1) prevents computation of all model parameter SEs, while scenario (2) prevents computation of specific parameter SEs. These limitations restrict theoretical development and practical applications of analytic information matrices.

Beyond analytic methods, bootstrap methods (Davison & Hinkley, 1997; Efron & Tibshirani, 1993) offer an alternative for computing SEs and CIs, with parametric bootstrap (PB) and non-parametric bootstrap (NPB) being most common. PB and NPB are widely applied (e.g., at least 20 papers in *Acta Psychologica Sinica* from January 2019 to August 2021 used bootstrap), highly generalizable, yet computationally intensive and time-consuming. Unlike analytic information matrices, PB and NPB require fewer assumptions and less formula derivation. These methods involve three steps: (1) obtain resampled datasets from observed data; (2) estimate model parameters from each resampled dataset; and (3) repeat these steps until reaching the predetermined number of resamples, then compute SEs and CIs from the distribution of parameter estimates. PB differs from NPB in that PB first estimates model parameters from observed data, then simulates resampled datasets using these parameters, whereas NPB draws resamples directly from observed data with replacement.

Although researchers suggest bootstrap can compute SEs and CIs in CDMs (Ma & de la Torre, 2020b) and theoretically address analytic information matrix limitations, its estimation accuracy remains understudied. As computationally intensive methods, their heavy computational load and long runtime restrict both theoretical research and practical applications. For instance, in PB and

NPB applications, too few resamples may affect accuracy, while too many reduce efficiency. The optimal resample size remains controversial (e.g., Bai et al., 2016; Efron & Tibshirani, 1993; Guo & Wind, 2021; Hayes, 2009, 2018; Lai, 2021). Furthermore, PB and NPB performance in estimating CDM parameter SEs and CIs across different scenarios requires further investigation. With advances in multi-threading and parallel scheduling technologies, parallel computing has been gradually applied to computationally intensive methods (Denwood, 2016; Khorramdel et al., 2019). For bootstrap specifically, Zhang and Wang (2020) developed the R package bmem using parallel bootstrap for statistical power analysis (Zhang, 2014), and the linear mixed-effects model package lme4 (Bates et al., 2015) also provides parallel bootstrap, which Jiang et al. (2021) used to explore CI estimation for generalizability coefficients.

This paper addresses two main questions: (1) Drawing on previous parallel bootstrap techniques, we develop parallel parametric bootstrap (pPB) and parallel non-parametric bootstrap (pNPB) methods tailored for CDMs to improve computational efficiency. (2) We systematically examine pPB and pNPB performance in estimating CDM parameter SEs and CIs. As demonstrated in this paper, pPB and pNPB are simple, feasible methods that not only effectively address important theoretical questions regarding SE and CI in CDMs but also substantially enhance computational efficiency in practical applications.

The paper proceeds as follows: We first explain problems with analytic information matrix SE computation, then detail the proposed pPB and pNPB methods. Section 4 presents simulation studies exploring performance under correctly specified CDMs and attribute hierarchy conditions. Section 5 demonstrates the role and value of pPB and pNPB through empirical data analysis. Finally, we conclude with discussion and conclusions.

2 Analytic Information Matrix and Its Limitations

This section uses the identity-link G-DINA (Generalized Deterministic Input Noisy Output "AND" gate; de la Torre, 2011) to present three analytic information matrices and explain potential issues with non-positive definiteness and negative diagonal elements in variance-covariance matrices when computing CDM parameter SEs and CIs.

2.1 Saturated CDM

Consider a cognitive diagnostic assessment with N examinees, J items, and K attributes, where both attributes and items are dichotomously scored. The $N \times J$ item response matrix is denoted as $\mathbf{x} = \{x_{nj}\}$, and the $J \times K$ Q-matrix is denoted as $\mathbf{Q} = \{q_{jk}\}$. In the saturated G-DINA model, the probability of examinee n correctly answering item j is:

$$P(X_{nj}=1 \mid \alpha_n, \lambda_j) = \lambda_{j,0} + \sum_{k=1}^{K_j} \lambda_{j,k}\alpha_{nk} + \sum_{k'=k+1}^{K_j} \lambda_{j,kk'}\alpha_{nk}\alpha_{nk'} + \cdots + \lambda_{j,12\cdots K_j} \prod_{k=1}^{K_j} \alpha_{nk}$$

where α_n is the nth examinee' s attribute mastery pattern, $\mathbf{q}_j$ defines the attributes required to correctly answer item j, and λ_j contains all item j parameters. Appropriate constraints on the saturated G-DINA model yield various special models.

For illustration with $K = 2$ and $\mathbf{q}_j' = (1, 1)$, the saturated G-DINA item response function can be expressed as:

$$P_j(\alpha_n) = \lambda_{j,0} + \lambda_{j,1,1}\alpha_{n1} + \lambda_{j,2,1}\alpha_{n2} + \lambda_{j,1,2}\alpha_{n1}\alpha_{n2}$$

where $\lambda_{j,0}$ is the intercept parameter representing the probability of correctly answering item j by guessing without mastering any required attributes, $\lambda_{j,1,1}$ and $\lambda_{j,2,1}$ are main effect parameters for the first (α_1) and second (α_2) attributes, respectively, and $\lambda_{j,1,2}$ is the interaction effect between the two attributes.

When $K = 2$ and no attribute hierarchy exists, all possible attribute mastery patterns are:

$$\alpha = \{(0, 0), (1, 0), (0, 1), (1, 1)\}$$

Using the identity link function, structural parameters η describe the distribution of attribute mastery patterns. Since all attribute mastery pattern probabilities sum to 1, the last structural parameter is constrained as $\eta_4 = 1 - \sum_{c=1}^{3} \eta_c$.

2.2 CDM with Attribute Hierarchies

When attributes have hierarchical relationships, appropriate constraints on saturated model structural and item parameters yield HCDMs (Templin & Bradshaw, 2014). Again using $K = 2$, $\mathbf{q}_j' = (1, 1)$, and assuming a linear hierarchy where α_1 must be mastered before α_2, the possible attribute mastery patterns become:

$$\alpha = \{(0, 0), (1, 0), (1, 1)\}$$

Due to attribute hierarchy constraints, the third mastery pattern $(0, 1)$ from the saturated structural model does not exist. In this example, the HCDM item response function is:

$$P_j(\alpha_n) = \lambda_{j,0} + \lambda_{j,1,1}\alpha_{n1} + \lambda_{j,1,2}\alpha_{n1}\alpha_{n2}$$

If the true model is HCDM but saturated G-DINA is used for estimation, some structural parameters (e.g., η_3) and item parameters (e.g., $\lambda_{j,2,1}$ in saturated G-DINA) have true values of 0, causing redundancy in some CDM parameters. Following previous research (Liu, 2018; Liu et al., 2021), parameters with true values of 0 are termed **impermissible parameters**, while those with non-zero true values are **permissible parameters**.

2.3 Analytic Information Matrix and Its Limitations

Under certain regularity assumptions (Bishop et al., 2007), the difference between maximum likelihood estimates $\hat{\gamma}$ and true values γ follows a multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathcal{I}^{-1}$ (Liu et al., 2016):

$$\sqrt{N}(\hat{\gamma}-\gamma) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}^{-1})$$

where $\mathcal{I}$ is the expected Fisher information matrix calculated using true parameter values and expectations over single examinee response vectors (i.e., all possible response patterns). However, since true parameter values are unknown and possible response patterns grow exponentially with item count, $\mathcal{I}$ has only theoretical value and cannot be applied in practice (Liu, Xin et al., 2019).

To address $\mathcal{I}$' s limitations, researchers developed XPD, Obs, and Sw matrices by substituting parameter estimates $\hat{\gamma}$ for true values γ and observed response matrix $\mathbf{x}$ for expectations over single vectors (Liu, Xin et al., 2019; Philipp et al., 2018; Liu et al., 2016). The XPD matrix uses cross-products of first-order derivatives of the observed data log-likelihood:

$$\text{XPD} = \sum_{n=1}^{N} \left(\frac{\partial \ell_n(\hat{\gamma})}{\partial \gamma}\right) \left(\frac{\partial \ell_n(\hat{\gamma})}{\partial \gamma}\right)'$$

The Obs matrix uses second-order partial derivatives:

$$\text{Obs} = -\sum_{n=1}^{N} \frac{\partial^2 \ell_n(\hat{\gamma})}{\partial \gamma \partial \gamma'}$$

Obs matrix elements can also be expressed as (Liu & Maydeu-Olivares, 2014; Liu, Xin et al., 2019):

$$\text{Obs}_{\gamma_1,\gamma_2} = \sum_{o=1}^{O} \frac{f_o}{p_o(\hat{\gamma})} \frac{\partial p_o(\hat{\gamma})}{\partial \gamma_1} \frac{\partial p_o(\hat{\gamma})}{\partial \gamma_2}$$

where γ_1 and γ_2 represent any item parameter (λ) or structural parameter (η), O is the number of unique response patterns in $\mathbf{x}$, f_o is the observed proportion, and $p_o(\hat{\gamma})$ is the expected proportion for pattern o.

The Sw matrix, named for its shape, is:

$$\text{Sw} = \text{Obs}^{-1} \cdot \text{XPD} \cdot \text{Obs}^{-1}$$

requiring both Obs and XPD matrices.

The limitations of analytic information matrices are threefold. First, boundary value problems severely impact them. In CDMs, at least two scenarios cause boundary values that prevent SE computation or inflate SEs (DeCarlo, 2011, 2019). One scenario involves intercept parameters $\lambda_{j,0}$, which range between [0,1]. When true values equal 0 or 1 (at parameter space boundaries), estimates may approach these boundaries, causing item parameter boundary problems. Another scenario involves impermissible structural parameters. When CDMs contain attribute hierarchies but saturated models are used for estimation, impermissible item and structural parameters inevitably arise. Since structural parameters range [0,1], impermissible parameters' true values fall on boundaries, with estimates potentially approaching 0 (e.g., 10^{-6}). Boundary problems cause analytic information matrices to become unstable or singular (Liu et al., 2021).

Second, if impermissible structural parameter estimates deviate from their true value of 0, these biased estimates violate the assumptions in Equation (5), adversely affecting XPD, Obs, and Sw calculations. Third, since Obs equals XPD minus the rightmost term in Equation (8), computational errors may produce negative diagonal elements in Obs, preventing SE calculation (Liu & Maydeu-Olivares, 2014). These limitations restrict analytic information matrices' theoretical development and practical application.

3 Parallel Bootstrap Methods

3.1 Parallel Non-Parametric Bootstrap

NPB simulates population sampling to compute model parameter SEs. Treating original response matrix $\mathbf{x}$ as a "population," NPB draws new "samples" (resamples, denoted $\mathbf{x}^*$) with replacement, estimates model parameters $\hat{\gamma}^*$ from $\mathbf{x}^*$, repeats this B times, and computes SEs as the standard deviation of the B $\hat{\gamma}^*$ estimates. However, NPB suffers from low computational efficiency (Ma & de la Torre, 2020b).

The proposed pNPB implementation proceeds as follows:

Step 1: Determine resample size B, specify the fitted model, detect CPU core count, and create corresponding parallel worker processes.

Step 2: Parallel sampling phase. In each worker: (a) draw resample $\mathbf{x}^*$ from original data $\mathbf{x}$ with replacement; (b) estimate model parameters $\hat{\gamma}^* = (\hat{\lambda}^*, \hat{\eta}^*)$

using the GDINA package (Ma & de la Torre, 2020b) with the specified CDM. Repeat (a) and (b) in each worker until reaching predetermined resample count B.

Step 3: Compute variance-covariance matrix from the B $\hat{\gamma}^*$ estimates. SEs are obtained by taking square roots of diagonal elements.

3.2 Parallel Parametric Bootstrap

PB uses model parameter estimates $\hat{\gamma}$ as "population parameters" to simulate B resampled datasets $\mathbf{x}^*$, then estimates resampled parameters $\hat{\gamma}^*$.

The proposed pPB implementation proceeds as follows:

Step 1: In addition to pNPB Step 1, estimate item and structural parameters $(\hat{\lambda}, \hat{\eta})$ from original data $\mathbf{x}$ using the specified CDM.

Step 2: Parametric parallel sampling phase. In each worker: (a) simulate attribute mastery patterns for each examinee using structural parameters $\hat{\eta}$; (b) generate response matrix $\mathbf{x}^*$ using examinee attribute patterns and item parameters $\hat{\lambda}$; (c) re-estimate model parameters $\hat{\gamma}^* = (\hat{\lambda}^*, \hat{\eta}^*)$ from $\mathbf{x}^*$ using GDINA (Ma & de la Torre, 2020b). Parallel workers repeat (a)-(c) until reaching predetermined resample count B. Step 3 matches pNPB Step 3.

Compared to analytic information matrices, pNPB and pPB are more generalizable, require no tedious formula derivation, need fewer assumptions (e.g., asymptotic normality of parameter estimates), avoid matrix inversion, are less affected by boundary values, and are particularly suitable for SE and CI computation with impermissible structural parameters. Variance-covariance matrices require only B $\hat{\gamma}^*$ vectors, with diagonal elements never negative. Moreover, compared to traditional NPB and PB, pNPB and pPB offer faster execution and higher efficiency, enabling the first comprehensive, systematic investigation of SE and CI performance using pNPB and pPB in CDMs.

4 Simulation Study

4.1 Research Purpose

This study focuses on pNPB and pPB performance under correctly specified CDMs and boundary value problems. The simulation has two primary objectives: (1) Examine pNPB and pPB performance in estimating SEs and CIs under ideal conditions (correct model specification) and compare with analytic XPD, Obs, and Sw methods. For generalizability, both data-generating and fitted models use identity-link saturated G-DINA. (2) Investigate performance when attribute hierarchies exist (i.e., when structural and item parameters contain impermissible parameters). Note that XPD, Obs, and Sw often encounter non-invertibility issues with attribute hierarchies (Liu et al., 2021), preventing direct comparison with bootstrap methods under identical conditions.

Literature review (e.g., Bai et al., 2016; Efron & Tibshirani, 1993; Guo & Wind, 2021; Hayes, 2009, 2018; Lai, 2021) reveals considerable controversy regarding resample size selection, making this another simulation focus.

4.2 Research Method

We used the GDINA package (Ma & de la Torre, 2020b) for parameter estimation, adapted open-source code from bmem (Zhang & Wang, 2020) and lme4 (Bates et al., 2015) for pNPB and pPB implementation, and used XPD, Obs, and Sw estimation code from Liu et al. (2021) (available from authors). To ensure CDM parameter identifiability, especially under attribute hierarchy conditions (Gu & Xu, 2019, 2020), we adopted the Q-matrix from Ma and Xu (2021) shown in Figure 1. To isolate the effects of experimental factors, we assumed equal structural parameters and equal main/interaction effects across conditions, eliminating parameter magnitude influences. Simulations ran on a cloud server with an Intel i9-10980XE CPU (18 cores, 36 threads), with $R = 500$ replications per condition for stable results.

Figure 1. Q-matrix used in simulation study

Specifically, two data-generating models were used: saturated G-DINA and HCDM with hierarchy ($\alpha_1 \to \alpha_2$, $\alpha_1 \to \alpha_2$). For saturated G-DINA data, five SE estimation methods were compared: XPD, Obs, Sw, pNPB, and pPB. For HCDM data, only pNPB and pPB were used. Resample sizes had four levels: 200, 500, 3000, and 5000. Sample sizes were 1,000 and 3,000. Item quality had three levels: high ($P(0) = 0.1$, $P(1) = 0.9$), medium ($P(0) = 0.2$, $P(1) = 0.8$), and low ($P(0) = 0.3$, $P(1) = 0.7$), where $P(0)$ is guessing probability and $P(1)$ is correct response probability for examinees mastering all required attributes. All conditions used saturated G-DINA for parameter estimation, meaning parameters were correctly specified for saturated G-DINA data and redundant (with some true-zero parameters) for HCDM data.

4.3 Evaluation Indicators

We evaluated SE estimation performance using bias and 95% CI coverage rates. A model parameter' s 95% CI is:

$$[\hat{\gamma} - 1.96 \times SE(\hat{\gamma}), \hat{\gamma} + 1.96 \times SE(\hat{\gamma})]$$

If the 95% CI falls within $[0.95 \pm 1.96 \times \sqrt{0.95 \times (1 - 0.95)/R}]$ (with $R = 500$ replications), interval estimation is considered accurate. Bias is computed as:

$$\text{BIAS} = \frac{1}{R} \sum_{r=1}^{R} SE_r(\hat{\gamma}) - SE_{\text{empirical}}(\hat{\gamma})$$

where $SE_{\text{empirical}}(\hat{\gamma})$ is the standard deviation of 500 parameter estimate vectors $\hat{\gamma}^*$ across replications.

4.4.1 Results Under Correctly Specified CDM Parameters

Figures 2 and 3 show 95% CI coverage rates and SE bias for item parameters using pNPB and pPB under correct model specification. Under high item quality, nearly all item parameter 95% CIs fall within theoretical bounds (gray lines), with bias approaching 0, and both metrics improve with larger sample sizes. Under medium quality with $N = 1000$, most item parameters show good performance despite some 95% CIs falling outside theoretical bounds and slight bias fluctuations; with $N = 3000$, especially when $B \geq 500$, most item parameters demonstrate good 95% CI coverage and bias control. Under low quality, pNPB and pPB performance diverges: with $N = 1000$, pNPB tends to overestimate SE (most 95% CIs above theoretical bounds), while pPB tends to underestimate SE (most below bounds). Performance improves with larger sample sizes, with pPB outperforming pNPB. When $B \geq 500$, results are highly consistent across conditions, with no noticeable differences between $B = 3000$ and $B = 5000$.

Figure 2. 95% CI coverage rates for item parameters using pNPB and pPB under correctly specified CDM parameters

Figure 3. SE bias for item parameters using pNPB and pPB under correctly specified CDM parameters

Figures 4 and 5 show 95% CI coverage rates and SE bias for item parameters using analytic XPD, Obs, and Sw methods. Under high and medium quality, item parameter SEs perform well. With $N = 1000$, Sw slightly outperforms XPD and Obs; all improve with $N = 3000$. Comparing XPD, Obs, Sw, pNPB, and pPB, Sw and Obs generally perform slightly better. Low quality severely impacts XPD, Obs, and Sw performance: with $N = 1000$, XPD and Obs 95% CIs mostly fall below theoretical bounds (underestimating SE), while Sw' s mostly fall above (overestimating SE). With $N = 3000$, most 95% CIs fall within theoretical bounds. Notably, under low quality and $N = 1000$, XPD and Sw produce extreme SE estimates (e.g., >1000) for 9 and 86 parameters respectively, indicating unstable performance. Overall, Obs performs slightly better than other methods under low quality.

Figure 4. 95% CI coverage rates for item parameters using XPD, Obs, and Sw under correctly specified CDM parameters

Figure 5. SE bias for item parameters using XPD, Obs, and Sw under correctly specified CDM parameters

Figures 6 and 7 show 95% CI coverage rates and SE bias for structural parameters using bootstrap methods. Under high quality, both pNPB and pPB perform excellently, with all 95% CIs within or on theoretical bounds and bias nearly zero. Under medium quality with $N = 1000$, structural parameter 95% CIs show increased variability but most SEs perform well with minimal bias;

with $N = 3000$, both methods perform well. Low quality severely impacts performance: with $N = 1000$, pNPB 95% CIs mostly exceed theoretical bounds with positive bias, while pPB 95% CIs all fall below with negative bias, and increasing B shows no clear improvement. With $N = 3000$, performance improves, with pPB slightly outperforming other B levels, though B increases have minimal impact on pNPB.

Figure 6. 95% CI coverage rates for structural parameters using pNPB and pPB under correctly specified CDM parameters

Figure 7. SE bias for structural parameters using pNPB and pPB under correctly specified CDM parameters

Figures 8 and 9 show 95% CI coverage rates and SE bias for structural parameters using analytic methods. Under high and medium quality, XPD, Obs, and Sw perform well, with nearly all 95% CIs within or on theoretical bounds and bias near zero. Low quality severely impacts performance: with $N = 1000$, XPD and Obs 95% CIs fall below theoretical bounds (negative bias), while Sw' s mostly exceed (positive bias). With $N = 3000$, performance improves, especially for Sw. Notably, under low quality and $N = 1000$, Sw produces extreme SE estimates for 1 and 3 parameters respectively (outside plot ranges), again due to extreme SE values. Overall, Sw performs comparably or better than other methods, except under low quality and $N = 1000$ where all methods perform poorly.

Figure 8. 95% CI coverage rates for structural parameters using XPD, Obs, and Sw under correctly specified CDM parameters

Figure 9. SE bias for structural parameters using XPD, Obs, and Sw under correctly specified CDM parameters

4.4.2 Results Under Redundant CDM Parameters

As noted, fitting saturated G-DINA to HCDM data causes boundary value problems and unstable SE estimates from analytic information matrices. Bootstrap avoids matrix inversion, but pNPB and pPB performance requires investigation.

Under parameter redundancy, results are presented separately for permissible and impermissible parameters. To display full results, 95% CI coverage rate ranges are set to [0.3, 1]. Figures 10 and 11 show 95% CI coverage rates and SE bias for permissible item parameters. Despite good performance for most parameters under high and medium quality, some parameters show 95% CIs substantially below theoretical bounds with large negative bias. These extreme deviations remain consistent across experimental conditions and worsen with $N = 3000$. This occurs because fitting saturated models to HCDM data incorrectly treats "impermissible" attribute patterns as "permissible," biasing some item parameter estimates and affecting their 95% CI coverage and bias. Comparing Equations (3) and (4), when the true model is HCDM with linear hierarchy but saturated CDM is used, impermissible pattern $(0, 1)$ is incorrectly included,

making structural parameter η_3 and item parameter $\lambda_{j,2,1}$ have true values of 0. Aside from these extreme deviations, increasing B from 200 to 3000 slightly improves 95% CI coverage for parameters near theoretical bounds, with high consistency between $B = 3000$ and $B = 5000$. Under low quality, permissible item parameter 95% CI coverage shows substantial variability.

Figure 10. 95% CI coverage rates for permissible item parameters under redundant CDM parameters

Figure 11. SE bias for permissible item parameters under redundant CDM parameters

Figures 12 and 13 show 95% CI coverage rates and SE bias for impermissible item parameters. Overall, most impermissible item parameter 95% CIs fall below theoretical bounds with negative bias, showing high consistency within item quality levels. Sample size, item quality, and resample size have minimal impact. pNPB slightly outperforms pPB for impermissible item parameter SE estimation.

Figure 12. 95% CI coverage rates for impermissible item parameters under redundant CDM parameters

Figure 13. SE bias for impermissible item parameters under redundant CDM parameters

Figures 14 and 15 show 95% CI coverage rates and SE bias for permissible structural parameters. Under high and medium quality, pNPB and pPB perform well, with 95% CIs within or on theoretical bounds, improving with larger sample sizes and resample counts, and bias nearly zero. Item quality substantially impacts structural parameter 95% CI coverage and bias, with increased variability and deviation from zero as quality decreases. Under low quality with $N = 1000$, pPB 95% CIs all fall below theoretical bounds (underestimating SE), while pNPB 95% CIs mostly exceed (overestimating SE). Larger sample sizes improve performance, but increased resample counts have minimal impact.

Figure 14. 95% CI coverage rates for permissible structural parameters under redundant CDM parameters

Figure 15. SE bias for permissible structural parameters under redundant CDM parameters

Figures 16 and 17 show 95% CI coverage rates and SE bias for impermissible structural parameters. As previously discussed, redundant structural parameters affect item parameter estimates and their SEs. Eliminating impermissible structural parameters is valuable. Liu et al. (2021) explored using analytic SEs with z-statistics (Equation 1) for significance testing. Accurate structural parameter SEs are crucial for this purpose, but analytic methods suffer from boundary and singular matrix problems. Bootstrap methods lack these limitations, making impermissible structural parameter SE performance important. Under high quality, pNPB and pPB 95% CIs slightly exceed theoretical bounds

because bootstrap variation $SE(\hat{\eta}^*)$ exceeds empirical variation $SE(\hat{\eta})$. However, absolute differences are minimal. Performance differences are small across conditions, with resample size increases showing no improvement.

Figure 16. 95% CI coverage rates for impermissible structural parameters under redundant CDM parameters

Figure 17. SE bias for impermissible structural parameters under redundant CDM parameters

Under medium quality, pNPB 95% CIs mostly exceed theoretical bounds, while pPB' s mostly fall within bounds, indicating pNPB' s non-parametric resampling yields $SE(\hat{\eta}^*)$ closer to $SE(\hat{\eta})$. Larger sample sizes improve performance (except pPB' s third structural parameter). Resample size increases show no improvement.

Under low quality, sample size substantially impacts impermissible structural parameter SE performance. With $N = 1000$, pNPB 95% CIs exceed theoretical bounds while pPB' s fall below, because non-parametric resampling produces larger $SE(\hat{\eta}^*)$ than $SE(\hat{\eta})$. Performance improves with larger samples, but increasing B from 200 to 5000 has negligible impact.

Figure 18. Q-matrix for ECPE dataset

5 Empirical Data Analysis

The ECPE (Examination for Certificate of Proficiency in English; Templin & Bradshaw, 2014) is a classic dataset in CDM research. Our ECPE data, obtained from the CDM package (Robitzsch et al., 2020), contains responses from 2,922 examinees on 28 dichotomously scored English grammar items.

Content and psychometric experts identified three attributes: α_1 (morphosyntactic rules), α_2 (cohesive rules), and α_3 (lexical rules). Figure 18 shows the ECPE Q-matrix (Templin & Hoffman, 2013). These attributes may have a linear hierarchy: $\alpha_1 \rightarrow \alpha_2 \rightarrow \alpha_3$ (Liu et al., 2021; Templin & Bradshaw, 2014; Wang & Lu, 2021). Since structural parameter SEs are valuable for exploring attribute hierarchies, we compare structural parameter SE estimates from our methods with previous results (Liu et al., 2021) to demonstrate theoretical and practical value.

Table 1. SEs of structural parameter estimates for ECPE data

Method	η_1	η_2	η_3	η_4	η_5	η_6	η_7
XPD	0.012	-	0.008	0.009	0.008	0.004	0.006
Obs	0.012	-	0.008	0.009	0.008	0.004	0.006
Sw	0.012	0.008	0.008	0.009	0.008	0.004	0.006
pNPB-200	0.012	0.009	0.008	0.009	0.008	0.005	0.006

Method	η_1	η_2	η_3	η_4	η_5	η_6	η_7
pNPB-500	0.012	0.009	0.008	0.009	0.008	0.005	0.006
pNPB-3000	0.012	0.009	0.008	0.009	0.008	0.005	0.006
pPB-200	0.012	0.008	0.008	0.009	0.008	0.004	0.006
pPB-500	0.012	0.008	0.008	0.009	0.008	0.004	0.006
pPB-3000	0.012	0.008	0.008	0.009	0.008	0.004	0.006

Note: Numbers after pNPB and pPB indicate resample size. "-" indicates non-computable SE.

5.1 Data Analysis Method

We estimated parameters using identity-link saturated G-DINA and SEs using pPB and pNPB, comparing computation times with PB and NPB. Parameter estimation used GDINA; XPD, Obs, and Sw SE code was adapted from dcminfo (Liu & Xin, 2017). All programs ran on a cloud server. Key points: (1) The saturated structural model has $2^3 = 8$ attribute mastery patterns; since structural parameters sum to 1, the eighth parameter is constrained as $\eta_8 = 1 - \sum_{c=1}^{7} \eta_c$. (2) Theoretically, more resamples yield more accurate SEs. We included $B = 10,000$ for pPB and pNPB but did not examine PB and NPB runtimes due to excessive time requirements.

Figure 19. All possible attribute mastery patterns and corresponding structural parameter estimates for ECPE data

5.2 Research Results

Figure 19 shows the eight attribute mastery patterns and their structural parameter estimates. Table 1 presents SEs for these estimates. Comparing methods, pPB SEs are numerically very close to XPD SEs, while pNPB SEs are closer to Sw SEs. pNPB SEs are generally larger than pPB SEs, consistent with simulation results for permissible and impermissible structural parameters under redundancy.

If linear hierarchy $\alpha_1 \rightarrow \alpha_2 \rightarrow \alpha_3$ exists in ECPE data, structural parameters 2, 3, and 6 (gray in Figure 19) should approximate 0 (Templin & Bradshaw, 2014). However, whether $\hat{\eta}_6 = 0.014$ is "approximately zero" requires statistical testing. Liu et al. (2021) used XPD, Obs, and Sw SEs in z-statistics (Equation 1) to test structural parameter significance. After significance level correction, all methods (except Obs, which could not compute SE for parameter 2) consistently confirmed the linear hierarchy. Since structural parameter estimates

$\hat{\eta}$ are identical across methods, only $SE(\hat{\eta})$ varies. Our pNPB and pPB SEs for parameters 2, 3, and 6 fall between the minimum and maximum analytic SEs, so computed z-statistics also fall between analytic method extremes. Thus, both methods confirm the linear hierarchy. Importantly, when CDMs contain attribute hierarchies, XPD, Obs, and Sw frequently encounter non-invertibility issues, and Obs may produce negative diagonal elements preventing SE computation (e.g., parameter 2). Bootstrap computes SEs directly from resampled parameter estimates, avoiding matrix inversion. Consistent with simulations, increasing resample size has minimal impact on SE estimates, particularly for $B \geq 3000$.

To illustrate computational efficiency gains, we compared runtimes for 200, 500, and 3000 resamples: pNPB took 10.93s, 25.43s, and 135.36s; pPB took 15.42s, 36.01s, and 200.96s; NPB took 158.43s, 392.97s, and 2282.33s; PB took 220.77s, 537.15s, and 3201.17s. pNPB and pPB substantially improve computational efficiency.

6 Discussion and Outlook

Estimating model parameter SEs and CIs in CDM research is valuable yet challenging (de la Torre, 2011; Liu et al., 2021; Ma & de la Torre, 2019; von Davier, 2014). Analytic information matrices XPD, Obs, and Sw perform well in many applications (Liu, Xin et al., 2019; Philipp et al., 2018; Liu et al., 2016) but require positive definiteness and suffer from boundary problems (DeCarlo, 2011, 2019). Traditional bootstrap methods (NPB, PB) have fewer assumptions and strong generality but are computationally inefficient and time-consuming (Ma & de la Torre, 2020b). This study proposes pNPB and pPB for CDM parameter SE and CI computation, systematically examining effects of model specification, sample size, resample count, item quality, and estimation method. We demonstrate pNPB and pPB' s effectiveness and efficiency in analyzing ECPE data with potential attribute hierarchies.

Notably, besides analytic information matrices and bootstrap, other methods like MCMC can compute CDM parameter SEs and CIs. MCMC can estimate parameters and compute SEs from posterior standard deviations. However, MCMC estimation can be extremely time-consuming (e.g., >1 hour), and studying SEs and CIs requires many replications (e.g., 500+) for reliable simulation results (Liu, Xin et al., 2019; Philipp et al., 2018; Liu et al., 2016). Additionally, Bayesian methods may be sensitive to prior distributions (Jing et al., 2021). Therefore, this study does not examine MCMC performance for CDM parameter SE and CI estimation.

6.1 Discussion

(1) Bootstrap performance in SE and CI estimation

Fundamentally, both NPB and PB simulate population sampling: treating the sample or sample-estimated parameters as the "population"for resampling. Boot-

strap cannot generate information beyond its "sample." Thus, when observed CDM data contain more accurate information about unknown parameters, bootstrap performs better. Simulation results show that model specification, sample size, and item quality substantially impact pNPB and pPB performance. Under correct specification, observed data fit the model perfectly, while redundancy conditions show the opposite: fitting saturated models to hierarchical data biases parameter estimates due to impermissible parameters, highlighting the importance of attribute hierarchy testing (Hu & Templin, 2020; Liu et al., 2021; Ma & Xu, 2021). Larger samples contain more parameter information, yielding more accurate estimates. Higher item quality better distinguishes attribute mastery patterns, providing more information and improving pNPB and pPB performance. An interesting simulation finding is that under low quality, the last four items (each measuring 3 attributes with 8 parameters to estimate) show substantially worse 95% CI coverage and bias than earlier items, indicating less available information.

(2) Resample size effects on bootstrap

Bootstrap is computationally intensive: more resamples require more time (Efron & Tibshirani, 1993), though theoretically increasing accuracy (Hayes, 2009, 2018). Optimal resample size remains controversial (Bai et al., 2016; Guo & Wind, 2021; Lai, 2021). Leveraging parallel bootstrap efficiency, we examined $B = 200, 500, 3000$, and 5000. Overall, resample size has minimal impact on pNPB and pPB performance. When $B \geq 500$, results stabilize, with $B = 3000$ and $B = 5000$ producing nearly identical results. Under correct specification, increasing B from 200 to 3000 slightly improves some parameters' 95% CI coverage and bias. Under non-ideal conditions (low quality, impermissible parameters), resample size increases have minimal impact. Empirical analysis shows pNPB results with $B = 200, 500$, and 3000 differ only slightly from $B = 10,000$, and pPB with $B = 3000$ is nearly identical to $B = 10,000$. Theoretically, CDM information matrices measure information about model parameters in observed data (Liu, Xin et al., 2019), while SEs quantify estimation uncertainty (Liu et al., 2021). Thus, the amount of "information" in observed data is the primary factor affecting SE performance. Our simulation and empirical results support this theory, suggesting that information quantity, not resample count, most influences bootstrap performance. However, whether this conclusion generalizes requires further study.

6.2 Research Outlook

Several important questions warrant future research. (1) This study used 30 items and 4 attributes; future work should examine effects of different item and attribute counts on pNPB and pPB. (2) We only examined hierarchy ($\alpha_1 \rightarrow \alpha_2$, $\alpha_1 \rightarrow \alpha_2$); SE performance under different attribute hierarchies, especially for structural parameters, needs exploration. Real applications may involve both attribute hierarchies and correlations (Hu & Templin, 2020; Liu et al., 2021), which this study did not consider. pNPB and pPB performance in exploring

and validating attribute hierarchies deserves further investigation. (3) Beyond the 95% CI method used here, other bootstrap-based CI methods warrant attention (e.g., Jiang, 2021; Lai, 2021). (4) Analytic information matrices often fail to invert with attribute hierarchies, preventing direct method comparisons. Liu et al. (2021) preliminarily proposed a two-stage estimation approach to sequentially eliminate impermissible structural parameters, a valuable direction. Under correct specification, analytic methods (e.g., Obs, Sw) sometimes slightly outperform pNPB or pPB. Future studies could compare analytic methods using two-stage estimation with pNPB and pPB. (5) pNPB and pPB have potential beyond SE and CI computation, including differential item functioning detection, item-level model comparison, and Q-matrix validation. (6) While this study examined pNPB and pPB within CDMs, these highly generalizable methods could be developed and applied to other statistical and measurement models to resolve conflicting conclusions from prior research (e.g., Efron & Tibshirani, 1993; Hayes, 2009, 2018; Lai, 2021).

Results show: (1) Under correctly specified CDMs with high or medium item quality, pNPB and pPB perform well for item and structural parameter 95% CI coverage and bias, improving with larger samples and better item quality. Low quality severely impacts performance, with pNPB overestimating and pPB underestimating SEs. (2) Under redundant parameters with high or medium quality, most permissible item parameters and nearly all permissible structural parameters show good 95% CI coverage and bias, though some item parameters exhibit extreme deviations. Impermissible parameter 95% CI coverage is poor under most conditions. (3) Empirical analysis shows pNPB and pPB SEs confirm previous findings of linear attribute hierarchies in ECPE data, with massive computational efficiency gains over NPB and PB. (4) Based on simulation and empirical results, $B = 200$ resamples may suffice for quick SE preview, while $B = 3000$ or more is recommended for accurate estimation.

Key words: cognitive diagnostic model, standard error, confidence interval, bootstrap, parallel computing method

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

Source: ChinaXiv —Machine translation. Verify with original.