The Reliability of Model Parameter Point Estimation: A Case Study of CDM

Authors: Liu Yanlou, Chen Qishan, Wang Yiming, Jiang Xiaotong, Liu Yanlou

Date: 2023-05-11T00:00:00+00:00

Abstract

In psychological research, inappropriate model parameter estimation frameworks or convergence criteria severely compromise the reliability of model parameter point estimates, thereby undermining the reliability of research conclusions. This study proposes a novel CDM model parameter estimation framework based on MLE-EM, along with a new convergence judgment method. Through simulation studies and empirical data analysis, the performance of the new parameter estimation framework and convergence judgment method was examined and compared with existing model parameter estimation frameworks and convergence judgment methods. The results demonstrate that the new model parameter estimation framework and convergence criteria outperform existing frameworks and criteria, effectively enhancing the reliability of model parameter point estimates.

Full Text

Preamble

On the Reliability of Point Estimation of Model Parameters: Taking CDMs as an Example

LIU Yanlou[1],[2], CHEN Qishan[3],[4], WANG Yiming[2], JIANG Xiaotong[2]
([1] Academy of Big Data for Education; [2] School of Psychology, Qufu Normal University, Jining 273165, China)
([3] Philosophy and Social Science Laboratory of Reading and Development in Children and Adolescents (South China Normal University), Ministry of Education; [4] School of Psychology, South China Normal University, Guangzhou 510631, China)

Abstract
In psychological research, inappropriate model parameter estimation frameworks or convergence criteria can severely compromise the reliability of point

estimates of model parameters, thereby affecting the reliability of research conclusions. This study proposes a novel framework for estimating CDM model parameters based on MLE-EM, along with a new convergence assessment method. Through simulation studies and empirical data analysis, we explore the performance of the new parameter estimation framework and convergence assessment method, comparing them with existing model parameter estimation frameworks and convergence assessment methods. Results demonstrate that the new model parameter estimation framework and convergence criteria outperform existing ones, effectively improving the reliability of point estimates of model parameters.

Keywords: parameter estimation, point estimation, convergence criterion, cognitive diagnostic model

Across natural and social sciences, the reliability of research conclusions—the degree to which findings can be trusted—has received considerable attention, particularly regarding the replicability of research results (see Hu et al., 2016; Begley & Ellis, 2012; Ioannidis, 2005, 2008; Tajika et al., 2015). A survey conducted by *Nature* found that over 70% of researchers could not replicate others' experiments, and more than 50% could not replicate their own experiments (Baker, 2016).

Within psychology, researchers have investigated the prevalence and potential causes of replicability issues and proposed solutions from both statistical methodology and research practice perspectives (e.g., see the submission guidelines and self-checklist reports in *Acta Psychologica Sinica* or American Psychological Association, 2020).

In psychological research, investigators use model parameters to describe the relationship between examinees' observable behaviors (or observed data) and their latent traits. When fitting psychometric models to data, researchers tend to treat the measurement model and parameter estimation software as a "black box," rarely considering whether the model parameter estimates are reliable. For instance, maximum likelihood estimation is one of the most widely used model parameter estimation methods, with one prerequisite for a unique global optimum being that the likelihood function is convex. However, this assumption may not hold in practice, resulting in two or more local optima for model parameters.

When analyzing the same data with the same model, different starting values may lead to convergence at different local optima. According to maximum likelihood theory, different likelihood function values indicate different model parameter estimates; larger differences in likelihood function values correspond to larger differences between local optima. For example, if γ is any parameter in the model and the difference between the first point estimate and the second point estimate $\widehat{\gamma}^{(1)} - \widehat{\gamma}^{(2)}$ is not approximately zero, this indicates that the estimates and 95% CIs for model parameter γ differ between these two estimations.

The reliability of point estimates of model parameters forms the foundation for

reliable research conclusions. Therefore, how to improve the reliability of model parameter estimates and consequently enhance the replicability of research findings constitutes the primary question this paper addresses.

Cognitive diagnosis (or diagnostic classification) uses psychometric models to infer the relationship between examinees' observable external behaviors and their latent multidimensional, fine-grained psychological attributes (such as psychological structures, skills, processing procedures, or strategies, collectively referred to as attributes) (Rupp et al., 2010). Cognitive diagnostic models (CDMs) have garnered increasing attention across psychology, education, sociology, biology, and other fields (Sorrel et al., 2016; Wu et al., 2017). Consequently, this paper uses CDMs as an example to explore the reliability of point estimation of model parameters.

Currently, maximum likelihood estimation using the expectation-maximization algorithm (MLE-EM) represents one of the most widely used methods for estimating CDM model parameters (de la Torre, 2009, 2011; von Davier, 2008). For example, MLE-EM can be used to estimate CDM model parameters in R packages such as CDM (George et al., 2016) and GDINA (Ma & de la Torre, 2020), as well as in software like flexMIRT, Latent GOLD, mdltm, and Mplus (Sen & Terzi, 2020; Templin & Hoffman, 2013). Under ideal conditions, MLE-EM can yield point estimates with desirable asymptotic properties and consistency. However, researchers have noted that estimating CDM model parameters with MLE-EM may encounter issues such as non-convergence of model parameters, extreme item parameter values, (poor) local optima, and boundary values (DeCarlo, 2011, 2019; Ma & Guo, 2019; Ma & Jiang, 2021; Philipp et al., 2018; Templin & Bradshaw, 2014; Zeng et al., 2022). The general MLE-EM process involves specifying initial values for model parameters, iteratively performing E-steps (expectation steps) and M-steps (maximization steps), and stopping iteration when specific convergence criteria are satisfied, outputting point estimates of model parameters. Therefore, we can address the reliability of point estimation of model parameters by improving the parameter estimation framework (including initial value specification and EM procedures) and convergence criteria.

This paper will elaborate in Section 2 on the problems existing in CDM model parameter estimation frameworks and convergence criteria, and their impact on parameter estimation reliability. Section 3 details the newly proposed model parameter estimation framework and convergence criteria. Section 4 compares the performance of new and existing methods in model parameter estimation reliability through simulation studies. Section 5 presents empirical data analysis to examine the performance of the newly proposed model parameter estimation framework and convergence criteria when estimating CDM model parameters, comparing them with the performance of the GDINA package. Finally, discussion and conclusions are presented.

2 CDM and Problems in Model Parameter Estimation

This section first introduces saturated CDM and hierarchical cognitive diagnostic models (HCDM), then uses these as a basis to elaborate on problems in model parameter estimation such as non-convergence, extreme item parameter values, (poor) local optima, and boundary values.

2.1 Saturated CDM and HCDM

For convenience of expression, consider a cognitive diagnostic test with N examinees, K attributes, and J items, where both attributes and items are scored 0-1. Let matrix $\mathbf{Y} = \{y_{nj}\}_{N\times J}$ represent examinees' observed response matrix, where $y_{nj} = 1$ indicates examinee n answered item j correctly, and $y_{nj} = 0$ indicates an incorrect response. Matrix $\mathbf{Q} = \{q_{jk}\}_{J\times K}$ represents the relationship between attributes and test items, where $q_{jk} = 1$ indicates item j measures attribute k, and $q_{jk} = 0$ indicates it does not. Matrix $\alpha = \{\alpha_{lk}\}_{L\times K}$ represents all possible attribute mastery patterns, where $\alpha_{lk} = 1$ indicates examinees with the l-th attribute mastery pattern have mastered attribute k, $\alpha_{lk} = 0$ indicates they have not, and L represents the number of all possible attribute mastery patterns. Following previous research (Tian et al., 2014; Dempster et al., 1977), we refer to the examinees' item response matrix $\mathbf{Y}$ as "incomplete" data, and the matrix combining item response matrix and examinees' attribute mastery patterns as "complete" data. That is, the complete data matrix $\mathbf{X}$ can be expressed as:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \\ \vdots \\ \mathbf{x}_N \end{pmatrix} = \begin{pmatrix} \mathbf{y}_1 & \alpha_1 \\ \vdots & \vdots \\ \mathbf{y}_n & \alpha_n \\ \vdots & \vdots \\ \mathbf{y}_N & \alpha_N \end{pmatrix}$$

The structural model of CDM defines the distribution proportions of all possible attribute mastery patterns α in the examinee population. Let $\pi = (\pi_1, \ldots, \pi_l, \ldots, \pi_L)^T$ represent the structural parameter vector with $\pi_L = 1 - \sum_{l=1}^{L-1} \pi_l$, where π_l denotes the proportion of the l-th attribute mastery pattern $\alpha_l = (\alpha_{l1}, \ldots, \alpha_{lk}, \ldots, \alpha_{lK})^T$ in the examinee population, and the symbol "T" denotes transpose. In saturated CDM, $L = 2^K$. The item response model of CDM represents the probability of correct response for examinee i with the l-th attribute mastery pattern α_l on test item j. The conditional probability of correct response in saturated CDM can be expressed as:

$$P_{lj} = P(y_{nj} = 1 \mid \alpha_l, \mathbf{q}_j) = \lambda_{j,0} + \sum_{k=1}^{K} \lambda_{j,1,(k)} \alpha_{lk} q_{jk} + \cdots + \lambda_{j,K,(1,\ldots,K)} \prod_{k=1}^{K} \alpha_{lk} q_{jk}$$

In Equation (2), $\mathbf{q}_j = (q_{j1}, \ldots, q_{jk}, \ldots, q_{jK})^T$ represents the vector corresponding to item j in the Q-matrix; in item j, $\lambda_{j,0}$ denotes the intercept term, $\lambda_{j,1,(k)}$ denotes the main effect term corresponding to attribute k, and $\lambda_{j,K,(1,\ldots,K)}$ denotes the highest-order interaction effect term. The item parameter vector λ and structural parameter vector π constitute the model parameters $\gamma = (\lambda^T, \pi^T)^T$.

The difference between saturated CDM and HCDM lies in the definition of the structural model and item response model, with HCDM nested within saturated CDM. To illustrate this relationship in detail, consider an example. Suppose in a test $K = 2$, $\mathbf{q}_j = (1,1)^T$, examinee n's attribute mastery pattern is $\alpha_n = (1,1)^T$; and mastering the first attribute (α_1) is a prerequisite for mastering the second attribute (α_2). Then, in saturated CDM, all possible attribute mastery patterns can be expressed as:

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$$

Thus, the structural parameters of saturated CDM can be expressed as $\pi = (\pi_1, \pi_2, \pi_3, 1 - \sum_{l=1}^{3} \pi_l)^T$; according to Equation (2), the item response function of saturated CDM in this example can be expressed as:

$$P_{lj} = \lambda_{j,0} + \lambda_{j,1,(1)} + \lambda_{j,1,(2)} + \lambda_{j,2,(1,2)}$$

Based on the attribute hierarchical relationship, the attribute mastery patterns allowed in HCDM are:

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$$

Thus, the structural parameters of HCDM can be expressed as $\pi = (\pi_1, \pi_2, 1 - \pi_1 - \pi_2)^T$; the item response function in HCDM can be expressed as:

$$P_{lj} = \lambda_{j,0} + \lambda_{j,1,(1)} + \lambda_{j,2,(1,2)}$$

Comparing expressions (3) and (5), and expressions (4) and (6), we can see that HCDM can be obtained by constraining some structural parameters and item parameters in saturated CDM to 0. That is, if the "true" model is HCDM but saturated CDM is used to estimate model parameters, the true values of some model parameters equal 0. Some structural model parameters having true values equal to 0 means these parameters are at the lower boundary of the parameter space. If such boundary value problems are not resolved, they may cause MLE-EM parameter estimation to have multiple solutions.

2.2 Potential Problems in CDM Model Parameter Estimation

When using CDMs to fit response data, issues such as non-convergence of model parameters, extreme item parameter values, or multiple local optima may arise if there are too many model parameters, small sample sizes, or boundary values exist in model parameters, especially in structural parameters (Ma & Jiang, 2021; Templin & Bradshaw, 2014).

Both the probability of correct item response and structural parameters in CDMs range between [0,1]. When estimating model parameters, one may encounter problems where item parameters or structural parameters are at the upper or lower boundaries of the parameter space, which may prevent model parameter estimation or cause excessively large standard errors or even unsolvable parameters. Ma and Jiang (2021) proposed Bayesian modal estimation with monotonic constraints for estimating item parameters of the G-DINA model. However, their research indicated that item parameters obtained through Bayesian modal estimation or algorithms combining Bayesian modal with monotonic constraints may be biased; additionally, they noted that prior distribution selection requires extreme caution in practical applications because inappropriate prior information may lead to misleading or even erroneous results. To constrain model parameter estimates within appropriate boundaries, Yamaguchi (2022) further proposed constraining structural parameters as well. However, when hierarchical relationships exist among attributes but saturated structural models are used for parameter estimation, some structural parameters have true values equal to 0, making it inappropriate to use priors that constrain them away from 0.

Using MLE-EM to estimate CDM model parameters requires setting initial values for model parameters $\gamma^{(0)}$. Based on $\gamma^{(0)}$, the E-step calculates the expectation of the complete-data likelihood function, and the M-step finds model parameters that maximize the expectation function. Each iteration (denoted as rep) produces a model parameter estimate vector $\gamma^{(\text{rep})}$. Iteration stops when the convergence assessment method value is less than the convergence tolerance or when the maximum number of iterations is reached (George et al., 2016; Ma & de la Torre, 2020). If iteration stops because the convergence assessment method value is less than the convergence tolerance, model parameters are considered converged, and parameters from the final iteration serve as model parameter estimates $\hat{\gamma}$; otherwise, convergence is not achieved.

The following discussion addresses how structural parameter boundary values, maximum iteration count, and initial values may affect the M-step and iteration count, elaborating on problems with existing methods. Specifically, in the E-step of MLE-EM: given observed data $\mathbf{Y}$ and model parameters $\gamma^{(\text{rep})}$, the expectation of the log-likelihood function of complete data $\mathbf{X}$ is calculated:

$$\mathcal{Q}(\gamma \mid \mathbf{Y}; \gamma^{(\text{rep})}) = \mathbb{E}_{\mathbf{X}|\mathbf{Y},\gamma^{(\text{rep})}} \left[\log \left(\prod_{n=1}^{N} \pi_n \prod_{j=1}^{J} P_{nj}^{y_{nj}} (1 - P_{nj})^{1-y_{nj}} \right) \right]$$

In addition to obtaining the above expression, the E-step also calculates the expected counts of all attribute mastery patterns $n_l^{(\text{rep})}$ and the expected number of examinees correctly answering item j under each attribute mastery pattern $r_{lj}^{(\text{rep})}$ in the rep-th iteration. The M-step finds model parameters $\gamma^{(\text{rep}+1)}$ that maximize the function $\mathcal{Q}(\gamma \mid \mathbf{Y}; \gamma^{(\text{rep})})$. Then $\gamma^{(\text{rep}+1)}$ replaces $\gamma^{(\text{rep})}$ in the E-step, and iterations continue until convergence conditions are met or the preset maximum iteration count is reached.

Taking saturated G-DINA model parameter estimation as an example, in the M-step, through formula derivation (see de la Torre, 2009, 2011), the expression for the updated probability of correct response for item j under the l-th attribute mastery pattern can be obtained:

$$P_{lj}^{(\text{rep}+1)} = \frac{r_{lj}^{(\text{rep})}}{n_l^{(\text{rep})}}$$

Based on $P_{lj}^{(\text{rep}+1)}$, CDM item parameter estimates $\lambda_j^{(\text{rep}+1)}$ can be easily obtained; the updated structural parameter estimate can be expressed as:

$$\pi_l^{(\text{rep}+1)} = \frac{n_l^{(\text{rep})}}{N}$$

At least two situations in CDM research can cause structural parameter boundary problems (DeCarlo, 2011, 2019; Templin & Bradshaw, 2014; Yamaguchi, 2022). The first situation occurs when hierarchical relationships exist among attributes but saturated models are used for estimation. Comparing saturated CDM and HCDM reveals that if the "true" model is HCDM but saturated CDM is used to fit the data, some structural parameters in the model are "non-permissible" parameters, meaning their true values are 0. The second situation occurs when small sample sizes may result in few or zero examinees corresponding to certain attribute mastery patterns. In both situations, the true value of structural parameter π_l equals 0. Since $n_l = N \times \pi_l$, this may cause the expected count of attribute mastery patterns $n_l^{(\text{rep})}$ in the M-step to equal 0. That is, the numerator and denominator in Equation (8) may equal 0, causing abnormal termination of iteration. The structural parameter boundary value problem is closely related to model convergence assessment and the reliability of CDM parameter estimation.

For problems that boundary values may cause, there are at least three current solutions. The first uses prior distributions to constrain correct response probabilities (Liu et al., 2016; Ma & Jiang, 2021). This approach requires extreme caution as it can lead to biased parameter estimates, especially in contexts where hierarchical relationships exist among attributes. The second is the default method adopted in the GDINA software package (Ma & de la Torre, 2022). The specific approach is: if the denominator in Equation (8) is less than 0.001,

add correction coefficients 0.0005 and 0.001 to the numerator and denominator respectively, i.e., set $P_{lj}^{(\text{rep}+1)} = 0.0005/0.001 = 0.5$. However, the reasonableness of this setting is questionable. The third is the method adopted in the CDM software package, which adds a very small value 10^{-10} to the denominator $n_l^{(\text{rep})}$ in Equation (8) in each iteration (Robitzsch et al., 2022). However, the reasonableness of this setting in some special cases (e.g., when both numerator and denominator values are close to 10^{-10}) is also questionable.

[Figure 1: see original paper] Simple example of local or global optimum for a single parameter

MLE-EM requires setting initial values for model parameters before iteration begins. In CDM model parameter estimation, the setting of the initial value vector $\gamma^{(0)}$ may affect MLE-EM performance. When estimating model parameters, MLE-EM starts from parameter initial values $\gamma^{(0)}$ and gradually converges to (local) optimal model parameter estimates through iteration. Under ideal circumstances, the function expression (7) has only a global optimum, and initial values $\gamma^{(0)}$ will not affect the final model parameter estimates $\hat{\gamma}$. However, when the likelihood function of CDM has multiple local optima, different initial values $\gamma^{(0)}$ will yield different final estimates $\hat{\gamma}$. That is, when the model meets specific convergence criteria, model parameter estimates $\hat{\gamma}$ may only be a poor local optimum (Ma & Guo, 2019; Zeng et al., 2022). To improve the reliability of CDM model parameter estimates, researchers have proposed using multiple starting values (e.g., 300) to estimate model parameters (Ma & Guo, 2019); or generating multiple starting values (e.g., 200), calculating their likelihood function values, and then selecting the set of model parameters with the largest likelihood function value as the starting values for MLE-EM iteration. Figure 1 presents a simple example of local and global optima for a single parameter. In this example, there are two local optima and one global optimum. Suppose $\gamma^{(0)}$ is any parameter's initial value in the CDM model. If $\gamma^{(0)}$ is at point A, it will eventually converge to the local optimum on the left; if $\gamma^{(0)}$ is at point C, it will converge to the local optimum on the right; if $\gamma^{(0)}$ is at point B, it will converge to the global optimum. Among these three solutions, the solution with initial value at point A is the worst. It should be noted that the CDM parameter estimation process is far more complex than the process presented in Figure 1, and a single initial value cannot guarantee obtaining better model parameter estimates.

2.3 Convergence Criteria for CDM Model Parameter Estimation

Convergence criteria are used to determine whether model parameter estimates are sufficiently close to the optimal solution. Generally, convergence criteria consist of three components: the convergence assessment method, convergence tolerance, and maximum iteration count (Paek & Cai, 2013). Convergence tolerance is a small value preset by researchers before model parameter estimation to determine whether the model has converged (e.g., 10^{-4} or 10^{-6}, or even smaller). In model parameter estimation, if the actual iteration count

has not reached the preset maximum iteration count and the difference in the convergence assessment method between iterations is less than the convergence tolerance, model parameter estimates are considered converged; if the actual iteration count reaches the maximum iteration count but the difference in the convergence assessment method between iterations is not less than the convergence tolerance, model parameter estimates are considered non-converged, and the model's maximum likelihood estimates cannot be obtained.

Currently, at least six methods can be used to determine whether CDM model parameter estimation has converged (George et al., 2016; Ma & de la Torre, 2020; Ma & de la Torre, 2022; Robitzsch et al., 2022; Rupp & van Rijn, 2018).

The first is the absolute difference in item parameters. This method assumes that if the maximum value of the absolute difference between the item parameter vector after iteration $\lambda^{(\text{rep}+1)}$ and before iteration $\lambda^{(\text{rep})}$, $\max\{\text{abs}(\lambda^{(\text{rep}+1)} - \lambda^{(\text{rep})})\}$, is less than the preset convergence tolerance, model parameters are considered converged and iteration stops. The advantage of this convergence assessment method is that its convergence tolerance directly reflects the precision of item parameters.

The second is the absolute difference in model parameters. This method is similar to the absolute difference in item parameters; the difference is that the absolute difference in model parameters also incorporates the difference in structural parameters into convergence assessment, i.e., when the maximum value of $\max\{\text{abs}(\gamma^{(\text{rep}+1)} - \gamma^{(\text{rep})})\}$ is still less than the convergence tolerance, model parameters are considered converged.

The third is the absolute difference in item correct response probabilities. This method compares whether the maximum value of the absolute difference in correct response probabilities for all items under all attribute mastery patterns between iterations, $\max\{\text{abs}(\mathbf{P}^{(\text{rep}+1)} - \mathbf{P}^{(\text{rep})})\}$, where $\mathbf{P} \in \{P_{lj}\}_{L\times J}$, is less than a preset convergence tolerance. The fourth is the absolute difference in the vector composed of item correct response probabilities and structural parameters. This method builds on the third method by also incorporating structural parameters, so it will not be elaborated further. It can be observed that the above four convergence assessment methods are based on all or part of the model parameters. In CDMs, item correct response probabilities are generally composed of combinations of item parameters, meaning that compared to item parameters, the method based on item correct response probability differences more easily satisfies model convergence criteria.

The fifth is the log-likelihood function difference. The log-likelihood function difference calculates the absolute value of the difference between the negative two times log-likelihood functions of observed data at the rep-th and rep+1-th iterations, $\text{abs}\{-2[\ell(\gamma^{(\text{rep}+1)} \mid \mathbf{Y}) - \ell(\gamma^{(\text{rep})} \mid \mathbf{Y})]\}$. This method assumes that when the difference in log-likelihood function values between iterations is less than the convergence tolerance, the likelihood function has reached its maximum. However, some researchers have pointed out that the limitation of

this method is that log-likelihood function values are affected by the number of items and examinees, thus recommending the use of relative likelihood difference instead.

The sixth is the relative likelihood difference. The relative likelihood difference method incorporates log-likelihood function values into the convergence assessment calculation. It attempts to eliminate the impact of log-likelihood function magnitude on convergence criteria. This method compares whether the absolute value of the ratio of the difference between two likelihood functions to the current likelihood function is less than the preset convergence tolerance. The GDINA software package uses $\text{abs}\{2[\ell(\gamma^{(\text{rep}+1)} \mid \mathbf{Y}) - \ell(\gamma^{(\text{rep})} \mid \mathbf{Y})]/\ell(\gamma^{(\text{rep}+1)} \mid \mathbf{Y})\}$ (Ma et al., 2022). The limitation of this method is that $\ell(\hat{\gamma} \mid \mathbf{Y})$ is unknown before model parameter estimation, making it difficult to preset an appropriate tolerance based on this unknown value.

In CDM model parameter estimation, researchers show clear differences in their use of convergence assessment methods, convergence tolerance, and maximum iteration count. The most frequently used convergence assessment method is the absolute difference in item parameters, with corresponding convergence tolerance of 10^{-4} or 10^{-5} (see de la Torre 2009, 2011; Ma & de la Torre, 2016; Paulsen & Valdivia, 2022; Sen & Terzi, 2020). Some researchers using absolute difference in item parameters set even smaller convergence tolerance, such as 10^{-6} (George et al., 2016), 10^{-7} (Rupp & van Rijn, 2018), or 10^{-8} (Chiu et al., 2022); some researchers use log-likelihood function difference for convergence assessment, setting convergence tolerance to 10^{-3} or 10^{-4} (Khorramdel et al., 2019; Ma & Guo, 2019). However, Rupp and van Rijn (2018) argue that log-likelihood function difference depends on the number of items and examinees, and relative likelihood difference may be better for model parameter convergence assessment. Yet they did not investigate the performance of relative likelihood difference or the appropriate convergence tolerance for this method.

Additionally, researchers tend to use software default settings when estimating model parameters and rarely modify these defaults, but default settings for CDM model parameter estimation software also differ substantially. For example, the default convergence criteria in GDINA and CDM packages differ significantly (Ma et al., 2022; Robitzsch et al., 2022). The GDINA package defaults to: absolute difference in the vector of item correct response probabilities and structural parameters, convergence tolerance of 10^{-6}, and maximum iteration count of 2000. The CDM package uses a combination of convergence methods with different default settings across functions. In the CDM package, the gdina function defaults to a combination of absolute difference in item parameters with convergence tolerance 10^{-6} and absolute difference in log-likelihood function with convergence tolerance 10^{-3}, with a maximum iteration count of 1000.

It is evident that researchers use substantially different convergence criteria. Therefore, whether different convergence criteria affect the reliability of point estimates of model parameters under the same measurement model; if so, which

of the currently available model parameter estimation convergence assessment methods performs best; or whether a broadly applicable method can be developed to improve the reliability of point estimates of CDM model parameters—these are important questions that need to be addressed.

3 New Model Parameter Estimation Framework and Convergence Criteria

As previously described, boundary values, local optima, extreme item parameter values, non-convergence of model parameters, and convergence criterion settings in CDM model parameter estimation may affect the reliability of point estimates of model parameters, potentially impacting the replicability of research findings. Therefore, this paper proposes a new model parameter estimation framework to address the potential problems in model parameter estimation mentioned in Section 2.2, and a new convergence criterion to address the potential problems in convergence criteria mentioned in Section 2.3.

First, we address the solution to boundary value problems. Through Section 2.2, we can see that the three current solutions to boundary values all have some potential shortcomings. Drawing on settings in GDINA and CDM packages, this paper uses: if the denominator in Equation (8) is less than 10^{-10}, add 10^{-12} to the denominator. The expected number of examinees correctly answering item j under the l-th attribute mastery pattern (numerator) is not greater than the expected number of examinees under this attribute mastery pattern (denominator), so using this method ensures that the maximum value of $P_{lj}^{(\text{rep}+1)}$ in Equation (8) will not exceed 0.01. That is, this method ensures the denominator is not equal to 0 while minimizing the impact of correction coefficients on correct response probabilities. Interested readers may try other values, but we believe that as long as the denominator is not equal to 0 and $P_{lj}^{(\text{rep}+1)}$ is small (e.g., less than 0.01), different correction coefficients will not produce noticeable effects on model parameter estimation results.

Second, we address the comprehensive solution to local optima, extreme item parameter values, and non-convergence of model parameters.

In model parameter convergence assessment, the sole purpose of setting a maximum iteration count is to prevent the model parameter estimation program from falling into infinite (or nearly infinite) loops. However, when model parameters should converge, setting the maximum convergence count too small may cause MLE-EM to terminate loops prematurely, resulting in erroneous non-convergence. The first step in solving non-convergence is to set a sufficiently large convergence count; therefore, this study sets the maximum convergence count to 50000.

A prerequisite for CDM model parameters having only a global optimum is that Equation (7) is a convex function. However, this prerequisite may not always hold, leading to poor reliability of model parameters. Therefore, referencing Ma

and Guo's (2019) research, this paper proposes using multiple starting values to calculate CDM model parameters. That is, when encountering non-convergence or extreme item parameter values, regenerate initial values and recalculate. If model parameters converge under new initial value conditions, the log-likelihood function value is greater than the previous value, and item parameters have no extreme values, use the new estimates as the final model parameter estimates. In subsequent sections, this new model parameter estimation framework is referred to as mCDM, and various convergence criteria are explored based on this framework. Since mCDM may require model parameter estimation under multiple different initial values for the same observed data matrix $\mathbf{Y}$ under specific conditions, computational load may be substantial. Therefore, referencing previous research (Liu, 2022), computationally intensive parts of the mCDM program are implemented in C++ with parallel computing. It should be noted that the mCDM program has been uploaded to the Science Data Bank, and interested readers can download and use it.

Finally, we elaborate on the newly proposed convergence assessment method.

The principle of maximum likelihood estimation is to find model parameter values that maximize the log-likelihood function of observed data and use them as estimates of the model parameters' "true values." The purpose of convergence assessment methods is to determine whether the log-likelihood function value of observed data has approximately reached its maximum. However, a single assessment method may have deficiencies under specific conditions. Using absolute difference in log-likelihood function and absolute difference in model parameters as examples for illustration. The log-likelihood function difference method assumes that when the difference in log-likelihood functions between the rep-th and rep+1-th iterations is less than the preset convergence tolerance, the likelihood function value has reached its maximum. Figure 2 [Figure 2: see original paper] presents a simple example of potential defects in the log-likelihood function difference convergence assessment method. Assume point B is any parameter's initial value $\gamma^{(0)}$ in CDM. When model parameter $\gamma^{(\text{rep})}$ approaches the global optimum, if the likelihood function curve is relatively flat (see Farrell & Lewandowsky, 2018), the problem of large changes in absolute difference of model parameters but very small changes in log-likelihood function difference may occur. That is, the absolute difference in model parameters performs better than log-likelihood function difference. The potential problem with absolute difference in model parameters is that likelihood function magnitude is affected not only by model parameter values but also by the number of items and examinees (see Rupp & van Rijn, 2018).

[Figure 2: see original paper] Simple example of potential defects in log-likelihood function difference convergence assessment method

Theoretically, when conducting CDM model parameter estimation, the stricter the model parameter estimation convergence assessment method and convergence tolerance setting (meaning more iterations under the same convergence tolerance), the more likely one is to obtain model parameter estimates that

maximize $\ell(\hat{\gamma} \mid \mathbf{y})$. However, in practice, due to factors such as sample size, number of items, number of attributes, item response models, attribute hierarchical relationships, and potential misspecification of Q-matrix elements, it is difficult to predetermine which method and corresponding convergence tolerance is the strictest. Therefore, referencing previous research (George et al., 2016; Ma & de la Torre, 2020; Ma et al., 2022; Robitzsch et al., 2022; Rupp & van Rijn, 2018; von Davier, 2008; Xu & von Davier, 2008), to overcome potential defects of single assessment methods, this paper proposes comprehensively using absolute difference in model parameters, absolute difference in the vector of item correct response probabilities and structural parameters, log-likelihood function difference, and relative likelihood difference for model parameter convergence assessment based on a given convergence tolerance, and calls it the comprehensive method.

It should be noted that compared to item parameters (or item correct response probabilities), the number of structural parameters is relatively small. Examinees' observed response data can provide more information for each structural parameter, allowing their estimates to stabilize relatively quickly. That is, theoretically, whether structural parameters are included in the convergence assessment method should not make a noticeable difference. However, for prudence, the following research adopts methods that include structural parameters. Additionally, unlike the GDINA package, the relative likelihood difference calculation formula used in this study's mCDM program is $\text{abs}\{[\ell(\gamma^{(\text{rep}+1)} \mid \mathbf{Y}) - \ell(\gamma^{(\text{rep})} \mid \mathbf{Y})]/\ell(\gamma^{(\text{rep}+1)} \mid \mathbf{Y})\}$.

In summary, this study proposes a new framework and new convergence criteria for CDM model parameter estimation based on MLE-EM to improve the reliability of point estimates of model parameters. The new model parameter estimation framework includes improvements to the E-step and M-step of the MLE-EM method. The main improvement to the E-step is using different initial values to recalculate expectation counts in the E-step and conduct subsequent iterations when necessary (e.g., when model parameters do not converge or item parameters have extreme values). The main improvement to the M-step is ensuring the denominator in Equation (8) is not equal to 0 and that $P_{lj}^{(\text{rep}+1)}$ is small.

4 Simulation Study

4.1 Research Purpose

This study focuses on whether the newly proposed model parameter estimation framework and convergence criteria can effectively improve the reliability of point estimates of model parameter values. That is, whether the comprehensive assessment method under the newly proposed mCDM framework outperforms existing methods, and whether it can obtain parameter estimates that maximize the likelihood function while ensuring parameters remain within reasonable ranges.

Specifically, this includes: (1) the performance of various convergence criteria when both the data-generating model and fitted model are saturated G-DINA, i.e., under completely correct model specification; (2) the performance of various convergence criteria when the data-generating model is HCDM but saturated G-DINA is used for fitting, i.e., when boundary values exist in the model.

4.2 Research Methods

The performance of model parameter convergence criteria depends on specific model parameter estimation methods. In addition to the newly developed mCDM program in this study, the open-source software packages CDM (version 8.2-6; Robitzsch et al., 2022) and GDINA (version 2.9.3; Ma et al., 2022) can also be used for model parameter estimation. However, the CDM package's default setting uses attribute mastery pattern simplification to estimate structural parameters when $K \geq 4$ (Xu & von Davier, 2008). The authors' preliminary research found that some structural parameter estimates obtained under this method are biased. Therefore, this study uses two model parameter estimation frameworks: GDINA and mCDM. There are five convergence assessment methods: absolute difference in model parameters, absolute difference in the vector of item correct response probabilities and structural parameters, absolute difference in log-likelihood function, relative likelihood difference, and comprehensive method. These five convergence assessment methods are abbreviated as dp, ip, ll, rl, and comp, respectively. Referencing previous research, this study considers three convergence tolerances: 10^{-4}, 10^{-6}, and 10^{-8}. To distinguish different convergence criteria, we add the letter "G" before the abbreviation for GDINA framework methods and "m" for mCDM framework methods, and append the number of decimal places of the three convergence tolerances after the method abbreviation. For example, the convergence criterion with absolute difference in model parameters as the assessment method and convergence tolerance of 10^{-4} under the GDINA framework is abbreviated as Gdp4; the convergence criterion with comprehensive method as the assessment method and convergence tolerance of 10^{-6} under the mCDM framework is abbreviated as mcomp6. Thus, this study examines the performance of 30 convergence criteria composed of 2 estimation frameworks, 5 convergence assessment methods, and 3 convergence tolerances under various influencing factors.

The simulation study considers 2 data-generating models: saturated G-DINA model and HCDM with linear hierarchical relationships among attributes (α_1, α_2, α_3). Given that it is difficult to pre-specify appropriate hierarchical relationships before CDM model parameters in practice, saturated G-DINA is selected as the fitting model. Sample size and number of items significantly affect model parameter estimation accuracy and may also affect convergence criteria performance. This study considers 3 sample sizes: $N = 500$, 1000, and 4000; 2 levels of item numbers: $J = 16$ and 32; and fixes the number of attributes at 4. To ensure CDM model parameters are identifiable (Gu & Xu 2019, 2020), this study uses the Q-matrix presented in Figure 3 [Figure 3: see original paper] when the

number of items is 16; the Q-matrix for 32 items is constructed by repeating the Q-matrix in Figure 3 twice.

[Figure 3: see original paper] Q-matrix for J = 16 in the simulation study

To better approximate CDM application scenarios, referencing Liu (2018) and Liu et al. (2022) research designs, the following steps are used to generate true values of item parameters and structural parameters: (1) Intercept terms (i.e., guessing parameters) $P(\mathbf{0})$ in item parameters are randomly drawn from a uniform distribution $[0.05, 0.4]$; correct response probability parameters $P(\mathbf{1})$ are randomly drawn from a uniform distribution $[0.6, 0.95]$; and main effect terms and interaction effect terms are set equal, i.e., both main effects and interaction effects equal $[P(\mathbf{1}) - P(\mathbf{0})]$ divided by their count. (2) True values of structural parameters are generated based on a multivariate normal distribution: first, set the mean vector of the multivariate normal distribution to 0, and draw values for off-diagonal elements of the variance-covariance matrix from a uniform distribution $[0.3, 0.7]$. Then, randomly generate one million examinees from the multivariate normal distribution and dichotomize each examinee's value vector using 0 as the cutoff point, i.e., set values greater than 0 in the vector to 1 and other cases to 0, thereby converting to attribute mastery patterns. Finally, when the data-generating model is saturated G-DINA, directly calculate the distribution proportions of these one million examinees' attribute mastery patterns across L attribute mastery patterns and use them as true values of structural parameters; when the data-generating model is HCDM, only calculate the proportions of permissible attribute mastery patterns among these one million examinees' patterns and use them as true values of structural parameters in HCDM. Each experimental condition combination is replicated 500 times to obtain stable simulation results, and the maximum iteration count for both mCDM and GDINA is set to 50000.

4.3 Evaluation Metrics

The purpose of convergence criteria is to determine whether model parameters in the iteration process have maximized the likelihood function. Therefore, this study's evaluation metrics are primarily constructed around the log-likelihood function, including: number of best likelihood functions (LL_{best}), mean of likelihood functions (LL_{mean}), maximum of likelihood functions (LL_{max}), minimum of likelihood functions (LL_{min}), and standard deviation of likelihood functions (LL_{sd}). The number of best likelihood functions refers to the number of times each of the 30 convergence criteria achieves the best likelihood function value in 500 replications: $LL_{\text{best}} = \sum_{R=1}^{500} I\{LL_{\text{criterion}R} = LL_{\text{max}R}\}$, where $LL_{\text{criterion}R}$ represents the log-likelihood function value corresponding to each convergence criterion in the R-th replication, $LL_{\text{max}R} = \max\{LL_{\text{criterion}R}\}$ represents the maximum log-likelihood function value among all convergence criteria in the R-th replication, and I is an indicator function used to determine whether the two function values are equal. If $LL_{\text{criterion}R}$ equals $LL_{\text{max}R}$, function I equals 1, otherwise it equals 0. Regarding LL_{best}, it should be noted that multiple con-

vergence criteria may simultaneously achieve the best likelihood function value in a single replication; larger LL_{best} values indicate better convergence criterion performance. LL_{mean}, LL_{max}, LL_{min}, and LL_{sd} represent the mean, maximum, minimum, and standard deviation of log-likelihood function values corresponding to the 30 convergence criteria across 500 replications, respectively, e.g., $LL_{\text{mean}} = \text{mean}\{LL_{\text{criterion}R}\}$.

Other evaluation metrics include: average time for a single run of the model parameter estimation program across 500 replications for the 30 convergence criteria (t_{mean}, in seconds), average iteration count (Itr_{mean}), maximum actual iteration count (Itr_{max}), total number of extreme values across all item parameters (defining item parameters greater than 1 or less than -1 as extreme values, denoted as λ_{ext}), and total number of non-convergences of the model parameter estimation program.

4.4 Simulation Results

Before presenting specific results, two general findings are noted. Simulation results under all experimental condition combinations in this study show: (1) When the maximum iteration count is 50000, model parameters converge in all replications, with no non-convergence cases. That is, the non-convergence count metric is 0 for all conditions. (2) Under the same model parameter estimation framework (GDINA or mCDM) and convergence tolerance (10^{-4}, 10^{-6}, or 10^{-8}), the rl method performs worst among all methods, with LL_{best} equal to 0. Therefore, simulation results for the rl method are no longer presented.

4.4.1 Performance of Convergence Criteria with Saturated CDM Data Generation Table 1 presents the performance of 24 convergence criteria (excluding the rl method) under the condition of saturated G-DINA data generation, $J = 16$, $N = 500$. Based on the LL_{best} metric in Table 1, the best-performing criterion among these is the comprehensive method mcomp8 under the new mCDM framework with convergence tolerance 10^{-8}. Regarding convergence assessment methods, under the same convergence tolerance, whether in GDINA or mCDM framework, the best-performing method is comp, followed by dp; ip performs similarly to dp but slightly worse, mainly because dp uses model parameters while ip uses combinations of parameters. Regarding convergence tolerance, under the same model parameter estimation framework and convergence assessment method, as convergence tolerance becomes smaller, convergence criterion performance improves. Taking the comp method as an example, as convergence tolerance changes from 10^{-4} to 10^{-6}, Gcomp shows similar LL_{sd} but improved performance on LL_{best}, LL_{mean}, LL_{max}, and LL_{min} metrics; mcomp shows similar patterns. Additionally, it should be noted that when convergence tolerance changes from 10^{-6} to 10^{-8}, metrics such as LL_{best}, LL_{mean}, LL_{max}, and LL_{min} show almost no noticeable change, but Itr_{mean} and Itr_{max} increase substantially. Regarding model parameter estimation framework, it is evident that each convergence criterion performs better under the mCDM

framework than under the GDINA framework. A clear example is that mcomp8 outperforms Gcomp8 on the LL_{best} metric. Moreover, t_{mean} and λ_{ext} under the mCDM framework are also significantly better than under the GDINA framework. The Itr_{max} metric shows that the maximum actual iteration counts for both GDINA and mCDM exceed 30000. This indicates that default maximum iteration settings in some CDM parameter estimation software are unreasonable and may produce erroneous non-convergence conclusions.

Based on results in Table 1, when convergence tolerance equals 10^{-4} across convergence assessment methods, performance on the LL_{best} metric is not good; although ip methods perform similarly to dp methods, ip performance is relatively worse. Therefore, under completely correct model specification, results for convergence tolerance 10^{-4} and convergence assessment method ip are no longer presented.

Table 2 presents simulation results for convergence criteria under saturated CDM data generation with $J = 16$ and sample sizes $N = 1000$ and 4000. At $N = 1000$, the best-performing convergence criterion is again mcomp8; when $N = 4000$ and convergence tolerance is 10^{-8}, all convergence criteria in Table 2 show good performance. Comparing Tables 1 and 2 comprehensively reveals that as sample size increases: (1) dp, ll, and Gcomp methods with convergence tolerances 10^{-6} and 10^{-8} show improved performance on LL_{best}, LL_{mean}, LL_{max}, and LL_{min} metrics, but performance is better with convergence tolerance 10^{-8}; (2) Itr_{mean}, Itr_{max}, and λ_{ext} metrics all decrease, with λ_{ext} even equaling 0 at $N = 4000$; (3) at $N = 500$ and 1000, convergence criteria under the mCDM framework outperform those under the GDINA framework, but at $N = 4000$, most convergence criteria under the GDINA framework perform similarly to those under the mCDM framework.

Table 3 presents simulation study results under saturated model data generation with $J = 32$. Since no extreme values occurred in any replications, the λ_{ext} column is not presented. Additionally, results from Tables 1 and 2 show that performance at convergence tolerance 10^{-8} is significantly better than at 10^{-6}, so dp and ll method results are no longer presented. Table 3 also shows that mcomp8 performs best on the LL_{best} metric. That is, under completely correct model specification, mcomp8 performs best among all convergence criteria examined in this study.

Comparing performance of each convergence criterion across different sample sizes in Table 3 reveals: (1) At $N = 500$, most convergence assessment methods under the mCDM framework outperform those under the GDINA framework on the LL_{best} metric; at $N = 1000$ and 4000, performance of convergence criteria under the GDINA framework improves, and both mCDM and GDINA frameworks show good performance when convergence tolerance is 10^{-8}. (2) Regarding Itr_{mean} and Itr_{max} metrics, these decrease as sample size increases. This indicates that when $J = 32$, required iteration counts decrease as sample size increases.

4.4.2 Performance of Convergence Criteria with HCDM Data Generation Tables 4 to 6 present simulation results under conditions where response data are generated by HCDM (with linear hierarchical relationships among the first three attributes) but model parameters are estimated using saturated CDM.

According to results in Table 4, among all convergence criteria, mcomp8 under the mCDM framework performs best, with mdp showing performance close to mcomp under the same convergence tolerance. Based on LL_{best}, LL_{mean}, LL_{max}, and LL_{min} metrics, convergence criteria under the mCDM framework far outperform those under the GDINA framework. Regarding convergence tolerance, under the three tolerance levels in the mCDM framework, performance on LL_{best}, LL_{mean}, LL_{max}, and LL_{min} metrics shows clear differences. As convergence tolerance becomes smaller, performance of each method improves, with best performance at 10^{-8}.

Combining with Table 1, we can see that differences in performance across convergence tolerance levels 10^{-6} and 10^{-8} for each convergence criterion in Table 4 are more pronounced. Regarding iteration counts, the maximum iteration counts for the best-performing comp8 under both mCDM and GDINA frameworks exceed 10000, indicating that under small sample conditions (e.g., $N = 500$), setting iteration count too small (e.g., less than 10000) may cause the model parameter estimation program to output erroneous non-convergence conclusions. The λ_{ext} metric in Table 4 shows that among 500 replications, Gcomp8 has 591 parameters with extreme values, while the mCDM framework has 483 extreme values. This indicates that although the mCDM framework can effectively reduce extreme value counts, comparing with extreme value counts in Table 1 reveals that boundary value problems have relatively negative impacts on model parameters under both frameworks.

Combining Tables 1 and 4, under conditions where boundary values exist in the model, we similarly find: (1) dp and ip methods show high consistency in performance, with dp performance comparable to or better than ip; (2) performance of convergence assessment methods at convergence tolerance 10^{-4} is not better than at 10^{-6} or 10^{-8}. Therefore, subsequent results no longer present ip method results or convergence criteria results at tolerance 10^{-4} under conditions where boundary values exist in the model.

Simulation results for $N = 1000$ and 4000 in Table 5 show that the best-performing convergence criterion on the LL_{best} metric remains mcomp8, followed by mdp8. Regarding convergence tolerance, each convergence criterion shows clear differences on LL_{best} and LL_{mean} metrics between 10^{-6} and 10^{-8} levels, with better performance at 10^{-8}. Model parameter estimation framework has a clear impact on convergence criteria performance; overall, across all metrics used in this study, convergence criteria under the mCDM framework outperform those under the GDINA framework. Sample size also affects the λ_{ext} metric for each convergence criterion when $J = 16$ and boundary values exist in the model. Combining Tables 4 and 5 reveals that as sample size increases, λ_{ext} values corresponding to each convergence criterion decrease. Re-

garding $Itr_{\max}$, when boundary values exist in the model, the iteration counts required for better-performing model convergence criteria under both mCDM and GDINA frameworks are very large. For example, under $N = 4000$, mcomp8 requires a maximum iteration count of 12714, and Gcomp8 requires 12509, far exceeding default iteration counts in CDM or GDINA software packages.

Table 6 presents simulation results when boundary values exist in the model and $J = 32$. Based on results in Tables 4 and 5 showing that convergence criteria perform better at convergence tolerance 10^{-8}, Table 6 no longer presents complete simulation results under GDINA framework and convergence tolerances 10^{-4} and 10^{-6}, including only Gcomp8 and mcomp6 for comparison. Table 6 shows that the best-performing convergence criterion across the three sample size levels $N = 500$, 1000, and 4000 is mcomp8, with mdp8 and mll8 performing relatively well. Regarding model parameter estimation framework: (1) convergence criteria under the mCDM framework outperform the same criteria under the GDINA framework on LL_{best}, LL_{mean}, t_{mean}, and λ_{ext} metrics; (2) under the mCDM framework, convergence tolerance values have clear effects on convergence method performance, with smaller convergence tolerance values yielding better performance on LL_{best} and LL_{mean} metrics for the same assessment method. Sample size has a clear impact on the λ_{ext} metric for each convergence criterion, with larger sample sizes yielding smaller λ_{ext} values for the same convergence criterion. Comparing performance of the same convergence criteria on Itr_{mean} and $Itr_{\max}$ metrics between $J = 16$ (Tables 4 and 5) and $J = 32$ levels reveals that these metrics decrease as the number of items increases; however, it should be noted that even under $J = 32$ conditions, $Itr_{\max}$ in Gcomp8 and mcomp8 convergence criteria may still exceed 3000.

5 Empirical Data Analysis

The data come from Yuan et al.'s (2022) cognitive diagnostic study on elementary school students' fraction operations. This dataset includes responses from 817 examinees to 56 items. Based on literature review, expert consultation, examinee interviews, and verbal report methods, Yuan et al. (2022) defined five cognitive attributes: basic operations (α_1), reduction (α_2), common denominator (α_3), mixed number splitting (α_4), and borrowing (α_5). Their research proposed a possible path for fraction operation cognitive processes: mastering α_1 is a prerequisite for mastering α_2, α_3, and α_5; since attribute α_4 only involves separating integer and fraction parts, it does not require prior mastery of α_1. Figure 4 [Figure 4: see original paper] presents the cognitive attribute hierarchical relationship diagram. Yuan et al. (2022) used likelihood ratio statistics to compare differences in log-likelihood function values between saturated CDM and HCDM under logit link function, initially confirming the existence of the hierarchical relationship presented in Figure 4 in the elementary school fraction operation dataset.

[Figure 4: see original paper] Cognitive attribute hierarchical relationship for elementary school fraction operations defined by Yuan et al. (2022)

This paper uses the elementary school fraction operation dataset as an example to explore the performance of 30 convergence criteria composed of 5 convergence assessment methods (dp, ip, ll, rl, comp) and 3 convergence tolerances (10^{-4}, 10^{-6}, 10^{-8}) under GDINA and mCDM model parameter estimation frameworks when boundary values exist in CDM models. Table 7 presents the log-likelihood function values (abbreviated as LL), model parameter estimation time in seconds (t), iteration counts, and λ_{ext} corresponding to these 30 convergence criteria; for ease of interpretation, LL values are retained to four decimal places.

According to maximum likelihood theory of model parameter estimation, larger LL values corresponding to a convergence criterion indicate better performance and higher reliability of point estimates of model parameters.

The results show: (1) The factor with the greatest impact on LL values is the model parameter estimation framework; LL values corresponding to convergence criteria under the newly developed mCDM framework are much larger than those under the GDINA framework. (2) Among all convergence criteria, the best-performing are mdp8 and mcomp8, which not only have the largest likelihood functions but also have no extreme values in item parameters. (3) Regarding the three convergence tolerances, 10^{-4} performs worst and 10^{-8} performs best under both mCDM and GDINA frameworks; although 10^{-6} performance is similar to 10^{-8} in some convergence criteria, the former lacks universal applicability. These three findings are highly consistent with conclusions from the simulation study.

6 Discussion and Outlook

This paper demonstrates through theoretical analysis and simulation studies that point estimates of psychometric models can have reliability issues in some contexts, and that the newly developed model parameter estimation framework and convergence criteria can improve the reliability of model parameter estimates.

6.1 Discussion

First, through preliminary research the authors believe that setting the maximum iteration count too low may cause model parameter non-convergence problems (e.g., 3000 or below, see GDINA and CDM software packages). Therefore, this study sets the maximum iteration count to 50000. Simulation studies show that under all experimental condition combinations in this paper, both mCDM and GDINA model parameter estimation frameworks converge. Simulation studies reveal that under some specific conditions (see Table 1), the maximum iteration counts for both mCDM and GDINA exceed 30000, meaning that if the maximum convergence count is set to 3000, model parameter non-convergence problems will occur. Therefore, this paper argues that increasing the maximum iteration count of model parameter estimation programs helps solve model parameter non-convergence problems.

Second, to address potential boundary value and extreme item parameter value problems in CDMs, this paper developed a new CDM model parameter estimation framework mCDM. By comparing the performance of mCDM and GDINA frameworks in simulation studies and empirical data analysis, we find that the mCDM framework performs better than or at least comparable to the GDINA framework; and the mCDM framework effectively reduces the number of extreme item parameter values. Therefore, this paper suggests that mCDM may be a better choice when estimating CDM model parameters. One reason for boundary value problems in CDMs is the existence of hierarchical relationships among attributes, making some parameters in saturated CDM approximately equal to 0. Researchers have developed some attribute hierarchical relationship exploration or verification methods based on saturated CDM (Gu & Xu 2019; Liu et al., 2022; Templin & Bradshaw, 2014). We recommend that researchers further use existing methods or develop new methods to study attribute hierarchical relationships under the mCDM framework. When there is sufficient evidence for hierarchical relationships, using HCDM to analyze data under the mCDM framework may improve the reliability of point estimates of model parameters.

Third, this paper newly proposes the comprehensive convergence assessment method comp and compares the performance of dp, ip, ll, rl, and comp methods composed of 30 convergence criteria under 2 parameter estimation frameworks (mCDM and GDINA) and 3 convergence tolerances (10^{-4}, 10^{-6}, 10^{-8}). Regarding the three convergence tolerances examined in this study, 10^{-8} performs best, while 10^{-4} performs worse than 10^{-6} and 10^{-8}; smaller convergence tolerance values yield better convergence criterion performance, especially under the mCDM framework. Regarding the five convergence assessment methods dp, ip, ll, rl, and comp, comp performs best while rl performs worst; this is most evident under the mCDM framework. Therefore, this paper argues that when estimating model parameters, the comp method with convergence tolerance 10^{-8} under the mCDM framework has higher reliability.

6.2 Outlook

This paper uses the saturated G-DINA model under a single link as an example to examine the performance of various convergence criteria under mCDM and GDINA frameworks. Although this study preliminarily solves the problem of how to select appropriate convergence criteria in CDM model parameter estimation, the authors believe several issues require further exploration.

The first concerns the applicable convergence tolerance for the rl method. This paper finds that compared to other criteria, rl performs poorly on LL_{best} and LL_{mean} metrics when using convergence tolerance values 10^{-4}, 10^{-6}, and 10^{-8}. Analyzing the rl calculation formula $\text{abs}\{[\ell(\gamma^{(\text{rep}+1)} \mid \mathbf{Y}) - \ell(\gamma^{(\text{rep})} \mid \mathbf{Y})]/\ell(\gamma^{(\text{rep}+1)} \mid \mathbf{Y})\}$ in combination with the ll method can reveal the reason for this problem. Taking simulation results presented in Table 1 as an example, we can see that under this experimental condition combination, the ll method

performs relatively better under the mCDM framework with convergence criterion value equal to 10^{-8}. According to the definition of mll8, at this time abs$\{-2[\ell(\gamma^{(\text{rep}+1)} \mid \mathbf{Y}) - \ell(\gamma^{(\text{rep})} \mid \mathbf{Y})]\} < 10^{-8}$; based on the LL_{mean} value of -4948.008 in the table, we can approximately obtain the mll value at this time and calculate mrl:

$$mrl = \text{abs}\left\{\frac{\ell(\gamma^{(\text{rep}+1)} \mid \mathbf{Y}) - \ell(\gamma^{(\text{rep})} \mid \mathbf{Y})}{\ell(\gamma^{(\text{rep}+1)} \mid \mathbf{Y})}\right\} < \frac{10^{-8}}{2 \times 4948.008}$$

This means that if mrl wants to achieve effects similar to mll8, the convergence tolerance for the mrl method should be approximately 10^{-11}. Therefore, the authors suggest that subsequent researchers can continue exploring the performance of the rl method along this line.

The second concerns the mCDM framework and its applications. The main purpose of developing the mCDM framework in this study is to provide a more reasonable CDM model parameter estimation framework that minimizes the impact of model parameter non-convergence, boundary value problems, and extreme item parameter values on convergence criteria performance. It should be noted that when the maximum iteration count was set to 50000 in simulation experiments, parameter estimation converged in all replications under both parameter estimation frameworks. Therefore, in this study, the mCDM framework only functions when boundary value problems and extreme item parameter values exist. When boundary values exist in the model, although the number of extreme item values under the mCDM framework is less than under the GDINA framework, the frequency of extreme values under the mCDM framework remains high even at $N = 4000$. Therefore, this study believes it is necessary to continue exploring issues such as model parameter non-convergence, boundary value problems, and extreme item parameter values based on the mCDM framework.

Third, the performance of various convergence criteria under different link functions requires further exploration. This paper examines the performance of different convergence criteria using the saturated G-DINA model under a single link. However, two other widely used links exist in CDMs: logit link and log link (de la Torre, 2009, 2011; Templin & Bradshaw, 2014). One main difference among these three link functions is the different expression of the relationship between item parameters and item correct response probabilities. Given that dp performance is slightly better than ip in most cases, this study suggests that subsequent research could further explore the performance of various convergence criteria under different link functions.

References

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). Washington.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 553, 452–454.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391), 531–533.

Chiu, C. Y., Köhn, H. F., & Ma, W. (2022). Commentary on "Extending the Basic Local Independence Model to Polytomous Data" by Stefanutti, de Chiusole, Anselmi, and Spoto. *Psychometrika*. Advance online publication. https://doi.org/10.1007/s11336-022-09873-7

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics*, 40(5), 454–476.

DeCarlo, T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-Matrix. *Applied Psychological Measurement*, 35(1), 8–26.

DeCarlo, T. (2019). Insights from reparameterized DINA and beyond. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 549–572). Springer.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1), 115–130.

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

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–22.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*, 74(2), 1-24.

Gu, Y., & Xu, G. (2019). Learning attribute patterns in high-dimensional structured latent attribute models. *Journal of Machine Learning Research*, 20(115), 1–58.

Gu, Y., & Xu, G. (2020). Partial identifiability of restricted latent class models. *The Annals of Statistics*, 48(4), 2082–

Hu, C., Wang, F., Guo, J., Song, M., Sui, J., & Peng. K. (2014). The replication crisis in psychological research. *Advances in Psychological Science*, 24(9), 1504–1518. [胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平. (2016). 心理学研究中的可重复性问题: 从危机到契机. 心理科学进展, 24(9), 1504–1518.]

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5), 640–648.

Khorramdel, L., Shin, H. J., & von Davier, M. (2019). GDM software mdltm Including parallel EM algorithm. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models* (pp. 603–628). Springer.

Liu, R. (2018). Misspecification of attribute structure in diagnostic measurement. *Educational and psychological measurement*, 78(4), 605–634.

Liu, Y. (2022). A simple and effective new method of Q-matrix validation. *Acta Psychologica Sinica*, 54(8), 996–

Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics*, 41(1), 3–26.

Liu, Y., Xin, T., & Jiang, Y. (2022). Structural parameter standard error estimation method in diagnostic classification models: Estimation and application. *Multivariate Behavioral Research*, 57(5), 784-803.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, 69(3), 253–275.

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, 93(14), 1–26.

Ma, W., de la Torre, J., Sorrel, M., & Jiang, Z. (2022). GDINA: The generalized DINA model framework. R package version 2.9.3. https://CRAN.R-project.org/package=GDINA

Ma, W., & Guo, W. (2019). Cognitive diagnosis models for multiple strategies. *British Journal of Mathematical and Statistical Psychology*, 72(2), 370–392.

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.

Paek, I., & Cai, L. (2013). A comparison of item parameter standard error estimation procedures for unidimensional and multidimensional item response theory modeling. *Educational and Psychological Measurement*, 74(1), 58-

Paulsen, J., & Valdivia, D. S. (2022). Examining cognitive diagnostic modeling in classroom assessment conditions. *The Journal of Experimental Education*, 90(4), 916-933.

Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics*, 43(1), 88–115.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2022). CDM: Cognitive Diagnosis Modeling. R package version 8.2-6. http://CRAN.R-project.org/package=CDM

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: theory, methods, and applications*. Guilford.

Rupp, A. A., & van Rijn, P. W. (2018). GDINA and CDM packages in R. *Measurement: Interdisciplinary Research and Perspectives*, 16(1), 71-77.

Sen, S., & Terzi, R. (2020). A comparison of software packages available for dina model estimation. *Applied Psychological Measurement*, 44(2), 150–164.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgment test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, 19(3), 506–532.

Tajika, A., Ogawa, Y., Takeshima, N., Hayasaka, Y., & Furukawa, T. A. (2015). Replication and contradiction of highly cited research papers in psychiatry: 10-year follow-up. *The British Journal of Psychiatry*, 207(4), 357–

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, 79(2), 317–339.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, 32(2), 37–50.

Tian, W., Xin, T., & Kang, C. (2014). The data-augmentation techniques in item response modeling: current approaches and new developments. *Advances in Psychological Science*, 22(6), 1036–1046. [田伟, 辛涛, 康春花. (2014). 项目反应理论中潜在心理特质“填补”的参数估计方法及其演变. 心理科学进展, 22(6), 1036–1046.]

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61(2), 287–307.

Wu, Z., Deloria-Knoll, M., & Zeger, S. L. (2017). Nested partially latent class models for dependent binary data; estimating disease etiology. *Biostatistics*, 18(2), 200–213.

Xu, X., & von Davier, M. (2008). Fitting the structured general diagnostic model to NAEP data (RR-08–27). Princeton, NJ: Educational Testing Service.

Yamaguchi, K. (2022). On the boundary problems in diagnostic classification models. *Behaviormetrika*. Advance online publication. https://doi.org/10.1007/s41237-022-00187-7

Yamaguchi, K., & Okada, K. (2020). Variational bayes inference algorithm for the saturated diagnostic classification model. *Psychometrika*, 85(4), 973-995.

Yamaguchi, K., & Templin, J. (2022). Direct estimation of diagnostic classification model attribute mastery profiles via a collapsed Gibbs sampling algorithm. *Psychometrika*, 87(4), 1390-1421.

Yuan, L., Liu, Y., Chen, P., & Xin, T. (2022). Development of a new learning progression verification method based on the hierarchical diagnostic classification model: taking grade 5 students' fractional operations as an example. *Educational Measurement: Issues and Practice*, 41(3), 69-82.

Zeng, Z., Gu, Y., & Xu, G. (2022). A Tensor-EM method for large-scale latent class analysis with binary responses. *Psychometrika*. Advance online publication. https://doi.org/10.1007/s11336-022-09887-1

On the reliability of point estimation of model parameter: taking the CDMs as an example

LIU Yanlou[1],[2], CHEN Qishan[3],[4], WANG Yiming[2], JIANG Xiaotong[2]
([1] Academy of Big Data for Education; [2] School of Psychology, Qufu Normal University, Jining 273165, China)
([3] Philosophy and Social Science Laboratory of Reading and Development in Children and Adolescents (South China Normal University), Ministry of Education; [4] School of Psychology, South China Normal University, Guangzhou 510631, China)

Abstract

Cognitive diagnostic models (CDMs) are psychometric models which have received increasing attention within the field of psychological, educational, social, biological, and many other disciplines. It has been argued that an inappropriate convergence criterion for MLE-EM (maximum likelihood estimation using the expectation maximization) algorithm could result in unpredictably distorted model parameter estimates, and thus may yield unstable and misleading conclusions drawn from the fitted CDMs. Although several convergence criteria have been developed, it remains an unexplored question, how to specify the appropriate convergence criterion for the fitted CDMs.

A comprehensive method for assessing convergence is proposed in this study. To minimize the impact by the model parameter estimation framework, a new framework adopting the multiple starting values strategy mCDM is introduced. To examine the performance of the convergence criterion for MLE-EM in CDMs, a simulation study under various conditions was conducted. Five convergence assessment methods were examined: the maximum absolute change in model parameters, the maximum absolute change in item endorsement probabilities and structural parameters, the absolute change in log-likelihood, the relative log-likelihood, and the comprehensive method. The data generating models were the saturated CDM and the hierarchical CDM. The number of items was set to $J = 16$ and 32. Three levels of sample sizes were considered: 500, 1000, and 4000. Three convergence tolerance value conditions were: 10^{-4}, 10^{-6}, and

10^{-8}.

The simulated response data were fitted by the saturated CDM using the mCDM and the R package GDINA. And the maximum number of iterations was set to 50000.

Simulation results suggest that: (1) The saturated CDM converged under all conditions. However, the actual number of iterations exceeded 30000 under some conditions, which implies that when predefined maximum iteration number is less than 30000, the MLE-EM algorithm might mistakenly stop. (2) The model parameter estimation framework affected the performance of the convergence criteria. The performance of the convergence criteria under the mCDM framework was comparable or superior to that of the GDINA framework. (3) Regarding the convergence tolerance values considered in this study, 10^{-8} consistently had the best performance in providing the maximum value of the log-likelihood and 10^{-4} had the worst as suggested by the higher log-likelihood value. Compared to all other convergence assessment methods, the comprehensive method in general had the best performance, especially under the mCDM framework. The performance of the maximum absolute change in model parameters was similar to the comprehensive method, however, its good performance was not guaranteed. On the contrary, the relative log-likelihood had the worst performance under the mCDM or GDINA framework.

The simulation results showed that, the most appropriate convergence criterion for MLE-EM in CDMs was the comprehensive method with tolerance 10^{-8} under the mCDM framework. Results from the real data analysis also demonstrated the good performance of the proposed comprehensive method and mCDM framework.

Key words: model parameter estimation, point estimation, convergence criterion, cognitive diagnostic model

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