# 基于假设检验的认知诊断$Q$矩阵修正

**李波，胡誉骞**

（华中师范大学 数学与统计学学院，湖北 武汉 430079）

**摘要**：本文创新性地引入假设检验方法作为$Q$矩阵修正手段，该方法为判断并正确修正$Q$矩阵中的错误属性界定提供可靠的理论基础，在实践应用中具有独特的优势。本文采用Monte Carlo 模拟及实证分析与同类研究比较，研究发现假设检验法展现卓越的性能和广泛适用性：(1)在各种作答错误率与$Q$矩阵错误率条件下均能显著提高修正精度；(2)对超参数置信度的选取具有优异的稳健性，减少主观判断需求；(3)在小样本场景下，各性能指标超越$\delta$法与$\gamma$法，尤其在高作答失误率下，其优势更为显著；(4)在大样本环境中，依旧保持强劲的竞争力，相较于其他方法如$\delta$法，其修正流程更简洁高效；(5)假设检验法在实证研究中显著提升认知诊断模型拟合效果，尤其在高维属性、样本稀缺的复杂数据场景下，对比其他算法，其优越性更为突出。

**关键词：认知诊断；$Q$矩阵修正； DINA 模型；假设检验；$\delta$法；$\gamma$法**

项目基金：

**通讯作者：胡誉骞，E-mail：wujyuhin@163.com**

# 1 引言

认知诊断模型(CDM)作为一种关键工具，能够根据学生的答题情况深入分析他们在各知识领域的强项与弱点，因而广受国际学术界及实践者的瞩目[1]。模型的实施框架涵盖了两大核心环节：精准构建“$Q$矩阵”以明确试题与认知属性的关联，以及实施“诊断分类”来细化学生的能力分布(涂东波, 2012; Tatsuoka, 2009)[4]。值得注意的是$Q$矩阵构建是一个复杂任务，高度依赖领域专家的手动标注，这一过程不仅耗时耗力，还容易受限于专家个人见解的主观性，不同专家对相同试题的知识点映射可能持有相异观点，从而引致$Q$矩阵构建的不一致性，降低了诊断结论的精确性。这凸显了$Q$矩阵在认知诊断中的基础性和重要性，其建构的严谨性和准确性直接影响诊断的有效性。现有研究已证实，$Q$矩阵的错误界定会对模型输出结果产生显著的消极影响[5]。因此，发现并修正$Q$矩阵中的错误是确保精确评估学生学习情况的关键所在。

为了解决$Q$矩阵的主观性问题并提升其准确性，国内外学者们提出了多种策略来估计和校正$Q$矩阵。例如，De la Torre (2008)针对*DINA*模型设计了$\delta$法[6]及在此基础上拓展的$\varsigma^2$法[错误!未找到引用源。]，该方法通过顺序探索所有潜在的答题模式以最小化猜测和错误参数的估计值，这一过程涉及反复迭代以寻找最优解，因此计算负担较重且耗时较长。涂东波等人(2012)则依据熟练组与非熟练组应试者答题记录的差异性效应指标，并结合$g$和$s$参数的大小，来评判$Q$矩阵的正确性并进行必要的调整($\gamma$法)[错误!未找到引用源。]。尽管这种方法在计算上更为简便且效率较高，某些修正指标与$\delta$法相当，但要确保修正结果的稳定性，通常需要较大的样本容量。在参数化方法领域，还包括了基于似然比*D*2统计量的方法(喻晓锋, 2015)、残差分析的方法(Chen, 2017)、使用 ICC-IR 方法(汪大勋, 2018)以及两阶段处理策略(汪大勋, 2019)等[9]。而在非参数化方法领域，Chiu (2013)提出的 RSS 法通过量化理想回答与实际回答之间的差距来指导$Q$矩阵的修正[12]，此法不仅计算简便，修正成效亦佳，然而，它目前尚不能直接用于$Q$矩阵的初步估计。此外，汪大勋(2018)引入了一种基于海明距离的非参数方法，该方法简单且运行时间短，估计准确率尚可[13]。

许多现有的开发方法在实际应用中遭遇了挑战，比如$\gamma$法和基于残差的方法常常受到小样本启动难题的困扰[错误!未找到引用源。]。另一方面，$\delta$法等技术则因为复杂的计算流程和繁琐的操作步骤而导致耗时巨大。通过实验，我们观察到在数据量有限(特别是样本稀少)或问题质量不佳(即正确与错误答案的区分不明显)的条件下，尽管某些算法能相对有效地修正错误的$Q$矩阵，但它们也可能不慎将大量原本正确的$Q$矩阵条目误标为错误。鉴于正确识

别$Q$矩阵错误的机会本就不高，这种反向修正的问题尤为突出，实非理想解决方案。在探索新领域或构建新的知识框架时，通常会遇到数据稀缺且项目初期质量不高的现实，这进一步限制了现有算法在这些情境下有效修正$Q$矩阵的能力。面对冷启动环境下，如何高效且准确地处理少量且质量欠佳的样本中的$Q$矩阵修正，成为了亟待解决的关键问题。有鉴于此，本研究在借鉴前人研究成果的基础上，创新性地提出了一种基于假设检验的$Q$矩阵修正策略。

# 2 本研究$Q$矩阵修正方法与思路

本节首先提出$Q$矩阵相关的理论和参数性质，并给出假设检验在$Q$矩阵修改中的理论推导，最后提出完整的修正流程。

## 2.1 $Q$矩阵与认知诊断相关理论研究

*DINA*模型是当前广受认可的认知诊断工具之一，它以结构简单和诊断精确度高而著称。该模型核心的项目参数包括两个关键要素：猜测参数($g$)和失误参数($s$)。猜测参数$g$衡量了考生在未完全掌握项目所考查知识点的情况下，仍能正确回答项目的概率；而失误参数$s$则衡量了考生已经掌握了项目所有相关知识点，却给出错误答案的情况。这两个参数在某种程度上反映了诊断过程中的不确定性或“噪声”，它们的数值如果过高，可能会对诊断的准确性造成不利影响。

相关研究文献(如 Rupp & Templin, 2008; de la Torre, 2008)[15]指出，*DINA*模型中的猜测参数($g$)和失误参数($s$)能够有效揭示测验$Q$矩阵中存在的冗余与遗漏问题。具体而言，当考核的属性出现冗余时，猜测参数$g$往往会增高；相反，若属性有所缺失，则失误参数$s$会相应增大。这意味着$g$和$s$参数可作为衡量$Q$矩阵准确性的间接指标。当$g$和$s$参数较大，则可能由属性冗余和属性缺失引起的。

在认知诊断领域，非补偿性 DINA 模型认为，如果受试者$i$掌握了测试项目$j$所涉及的所有必要属性，则其解题正确的可能性较大；反之，若知识点有所遗漏，该受试者解答错误的可能性较大。简言之，受试者对项目所考察的属性的掌握程度直接影响其答题表现。

本研究采用了一种创新的方法，其核心在于如下逻辑：当项目$j$未考察属性$k$时，理论上，那些掌握项目$j$所有属性的被试除非失误(失误概率$s$)，否则应当几乎无误地完成项目，意味着错误出现的概率极低。如果出现了这种少见的高错误率情况，我们有充分理由相信项目$j$实际上涉及了属性$k$。相反，如果项目$j$确实考察了属性$k$，那么对于那些仅

未掌握属性$k$、但掌握了$j$其余所有属性的被试而言，除非猜测(猜测概率$g$)，否则他们做出正确回答是少见的情形。一旦观察到这类被试意外地频繁正确作答，这便有足够把握认为项目$j$并未考察属性$k$。因此本研究在结合$DINA$模型中项目参数($g,s$)以及掌握属性与否是否影响对项目作答两个因素的基础上，提出了$Q$矩阵的修正方法——假设检验法。

## 2.2 修正$Q$矩阵的假设检验理论

### 2.2.1 符号定义：

$I$为学生数量，$J$为题目数量，$K$为知识点数量；$R=\left(r_{ij}\right)_{I\times J}$为作答矩阵，$r_{ij}$表示第$i$个学生对第$j$道题目的作答对错，正确为1，错误为0；

$Q$矩阵：$Q=\left(q_{jk}\right)_{J\times K}=\begin{pmatrix}q_1\\q_2\\\vdots\\q_J\end{pmatrix}$；学生掌握情况矩阵：$\beta=\left(\beta_{ik}\right)_{I\times K}=\begin{pmatrix}\beta_1\\\beta_2\\\vdots\\\beta_I\end{pmatrix}$；

$Q^w$为错误的$Q$矩阵，$\tilde{Q}$为修正后的$Q$矩阵；

$q_j=\left(q_{j1},\ldots,q_{jK}\right)$为第$j$题对$K$个知识点的考察情况，考察记为1，未考察记为0；

$\beta_i=\left(\beta_{i1},\ldots,\beta_{iK}\right)$为第$i$个学生对$K$个知识点的掌握情况，掌握记为1，未掌握记为0；

$\beta_i\succeq q_j$表示$\beta_i$中的每个分量不小于$q_j$中的对应分量，即$\forall k\in[1,...,K]$都有$\beta_{ik}\geq q_{jk}$；

$Q_{\left[j_1,\ldots j_{n_j};k_1,\ldots,k_{n_j}\right]}$表示$Q$矩阵的子矩阵，特别的，$Q_{[j;1,\ldots,K]}$表示$Q$矩阵的第$j$行$q_j$向量

定义运算$Q\setminus Q_{[j;1,\ldots,K]}=\begin{pmatrix}q_1\\\vdots\\q_{i-1}\\q_{i+1}\\\vdots\\q_J\end{pmatrix}_{(J-1)\times 1}$：$Q$矩阵去除第$j$行$q_j$向量，记$Q_{\setminus[j,:]}=Q\setminus Q_{[j;1,\ldots,K]}$；

同理$R\setminus R_{[1,\ldots,I;j]}=\left(r_1,\ldots,r_{j-1},r_{j+1},\ldots r_J\right)$表示$R$去除第$j$列作答向量，记$R_{\setminus[:,j]}=R\setminus R_{[1,\ldots,I;j]}$。

### 2.2.2 样本选择分析

选取合适的样本对于假设检验至关重要，因为它允许我们从参与作答的被试中筛选出符合特定条件的学生，这样操作可以有效控制由这些样本构建的统计量的分布特性。以一个具体的筛选准则为例，准则可表述如(1)：

$$T=\left\{i:\beta_i \succeq q_j, \text{且} \beta_{ik}=0\right\} \tag{1}$$

$T$ 样本的筛选能够解释为这样一个过程：首先从全部样本中筛选出那些成功掌握了第 $j$ 题考察模式的学生，然后进一步筛选其中未掌握第 $k$ 个知识点的学生。

在此选择模式中，如果第 $j$ 题并未涉及第 $k$ 个知识点，理论上所有 $T$ 样本中的学生，凭借他们已掌握的模式，应当能够正确解答第 $j$ 题。仅在出现失误的情况下，他们才可能错解此题。相反，如果第 $j$ 题确实考核了第 $k$ 个知识点，鉴于 $T$ 样本中的学生恰恰未能掌握这一知识点，他们正常情况下应无法解答正确，除非是通过猜测才有偶然答对的可能。这样的样本筛选策略，确保了我们在分析数据时能够更准确地理解和解释结果，从而增强假设检验的有效性和准确性。

**2.2.2 假设检验**

本节将对 $Q$ 矩阵运用假设检验的逻辑进行分析，以第 $j$ 题中的第 $k$ 个知识点为例：

**第一步：参数估计**

由于 $q_{jk}$ 可能存在错误，因此为了降低第 $j$ 题 $q_j$ 向量错误导致模型参数的估计误差，采用删除了第 $j$ 题的数据。即 $DINA$ 模型输入 $Q_{[j,:]}$ 与 $R_{[:,j]}$，输出 $J-1$ 道题目的猜测参数 $\tilde{g}$ 和失误参数 $\tilde{s}$，以及每个学生的掌握情况 $\tilde{\beta}$ 。

**第二步：确定假设检验问题：**

当 $q_{jk}=0$，则可能存在属性缺失问题。当 $q_{jk}=1$，则可能存在属性冗余问题。

**第三步：分情况进行假设检验**

**情形一：属性缺失问题**

当 $q_{jk}=0$，为检验是否存在属性缺失问题，建立假设如下：

$$H_0: q_{jk}=0 \leftrightarrow H_1: q_{jk}=1 \tag{2}$$

根据第 $j$ 题的 $q_j$ 向量，以及学生掌握情况 $\tilde{\beta}$，使用公式 $(1)$ 中选择方法，选择能够做对第 $j$ 题但未掌握第 $k$ 个知识点的掌握模式集合 $T$ 。根据样本 $T$ 构建统计量，计算作答第 $j$ 题的错误数量 $X=n_T-\sum_{i\in T} r_{ij}$ 。

当原假设成立时( $q_{jk}=0$ )，样本 $T$ 中的学生都具有做对考察模式为 $q_j$ 的题目的掌握模

式 $\tilde{\beta}_i, i \in T$ 。他们对于第 $j$ 题的错误可以推断为非猜测性失误，因此错误概率可直接关联到 $DINA$ 模型估计的失误参数，但需要注意的是，由于模型训练过程中排除了第 $j$ 题的数据，即输入数据为 $Q_{[j,:]}$ 与 $R_{[:,j]}$，导致第 $j$ 题缺乏失误参数的估计。于是本文采取一个替代策略，即利用其他题目失误参数的平均指来估计这一概率：$\bar{s} = \frac{1}{J-1}\sum_{l=1}^{J-1}\tilde{s}_l$，在此基础上，相应的统计量 $X$ 服从概率为 $\bar{s}$ 的二项分布：

$$X \sim B\left(n_T, \bar{s}\right), P\left(X = x\right) = C_{n_T}^{x}\bar{s}^{x}\left(1-\bar{s}\right)^{n_T - x} \tag{3}$$

当原假设不成立时($q_{jk} = 1$)，意味着第 $j$ 题考察了第 $k$ 个知识点，然而样本 $T$ 中的所有学生均未掌握第 $k$ 个知识点。因此，在逻辑上预期这些学生会集体答错第 $j$ 题，在这种情况下，统计量 $X$ 有偏大的趋势，则拒绝域形式为 $\left[c, +\infty\right]$，具体表述如下：

$$W_1 = \left\{\left(r_{1j}, r_{2j}, \ldots, r_{n_T j}\right) : X \geq c\right\} = \left\{n_T - \sum_{i \in T} r_{ij} \geq c\right\} \tag{4}$$

给定置信度 $\alpha$，得否定域为 $\left\{X \geq b_\alpha\left(n_T, \bar{s}\right)\right\}$ 水平设定为 $\alpha$ 的缺失属性二项分布检验。若错误数量为 $x$ 且 $x \geq b_\alpha\left(n_T, \bar{s}\right)$，记 $P_{0\to 1} = 1 - P\left(X \geq x\right)$，则有 $P_{0\to 1} \geq 1-\alpha$，拒绝原假设。

**情形二：属性冗余问题**

当 $q_{jk} = 1$，问题变为检验是否存在属性冗余，建立假设如下：

$$H_0 : q_{jk} = 1 \leftrightarrow H_1 : q_{jk} = 0 \tag{5}$$

同理使用公式(1)的方法选择掌握模式集合 $T$，根据样本 $T$ 构建统计量，计算作答第 $j$ 题的正确数量 $Y = \sum_{i \in T} r_{ij}$。

原假设成立时($q_{jk} = 1$)，此时样本 $T$ 中的学生均未掌握第 $k$ 个知识点，因此他们无法正确作答第 $j$ 题，其正确答案仅能归因于随机猜测。与情形一类似，猜测概率采用 $DINA$ 模型估计其他题目猜测参数的平均值 $\bar{g} = \frac{1}{J-1}\sum_{l=1}^{J-1}\tilde{g}_l$，基于这样的猜测机制，统计量 $Y$ 服从概率为 $\bar{g}$ 的二项分布：

$$Y \sim B\left(n_T, \bar{g}\right), P\left(Y = y\right) = C_{n_T}^{y}\bar{g}^{y}\left(1-\bar{g}\right)^{n_T - y} \tag{6}$$

当原假设不成立时($q_{jk} = 0$)，即第 $j$ 题未考察了第 $k$ 个知识点，这意味着样本 $T$ 中的

学生均未掌握第$k$个知识点，因此这些学生都应该正确回答第$j$题，导致统计量$Y$有偏大的趋势，拒绝域形式为$[c,+\infty]$，具体为：

$$W_2=\left\{\left(r_{1j},r_{2j},\ldots,r_{n_Tj}\right):Y\geq c\right\}=\left\{\sum_{i\in T}r_{ij}\geq c\right\} \tag{7}$$

给定置信度$\alpha$，则临界值$c$满足$P\left(Y\geq c\right)=\sum_{x=c}^{n_T}P\left(Y=y\right)=\alpha$。取$c=b_\alpha\left(n_T,\bar{g}\right)$，得到水平$\alpha$的冗余属性二项分布，其否定域为$\left\{Y\geq b_\alpha\left(n_T,\bar{g}\right)\right\}$。

给定置信度$\alpha$，得否定域为$\left\{Y\geq b_\alpha\left(n_T,\bar{g}\right)\right\}$水平设定为$\alpha$的冗余属性二项分布检验。若正确数量为$y$且$y\geq b_\alpha\left(n_T,\bar{g}\right)$，记$P_{1\to0}=1-P\left(Y\geq y\right)$，则有$P_{1\to0}\geq1-\alpha$，拒绝原假设。

## 2.3 假设检验方法实现步骤

基于 2.2.2 两种基本的假设检验情形，对$Q$矩阵元素按图 1 流程逐题修正：认知诊断模块用于估计参数；属性修正模块是对第 $j$ 题所有属性分别修正，其中，属性可能同时拒绝缺失和冗余假设，因此需要取“更有可能”的情况；最后，每个属性可能同时修正为 0，因此需要验证模块，由于至少考察一个属性，故在$K$个属性中取缺失概率最大的属性。

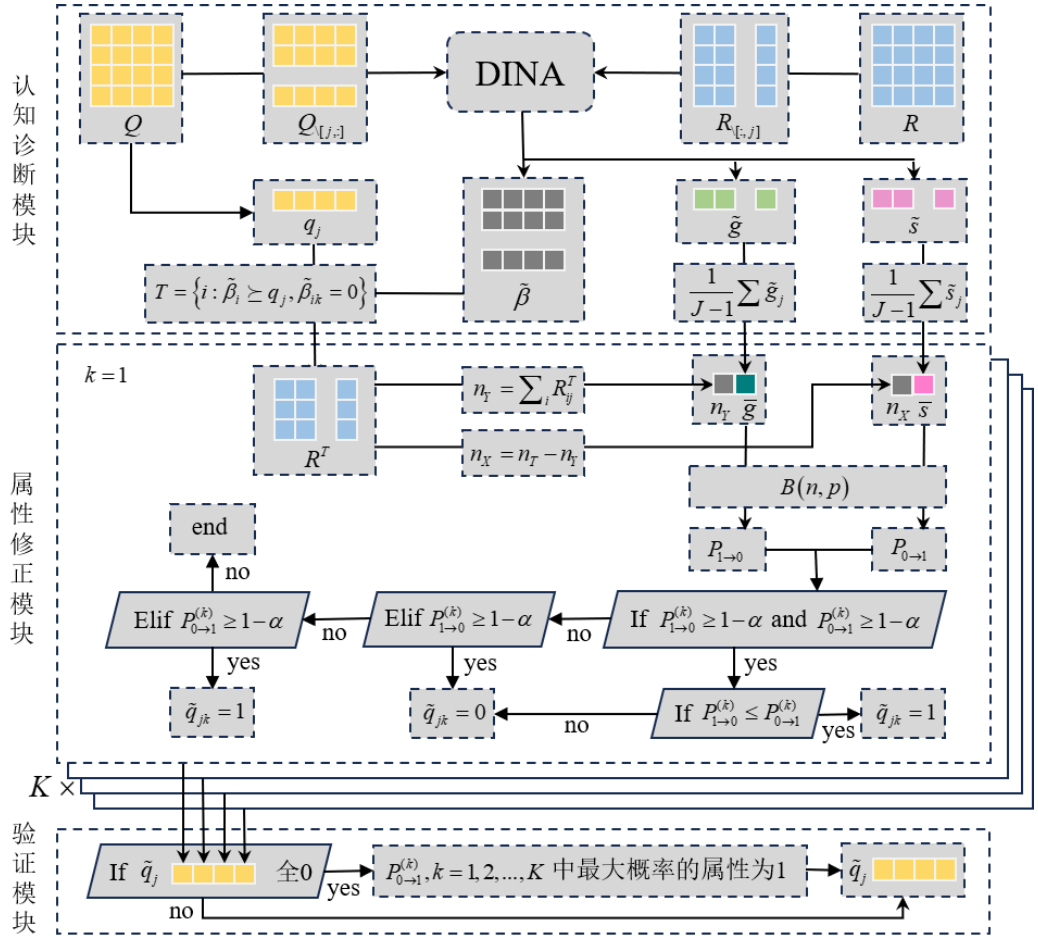


图 1 $Q$ 矩阵的第 $j$ 题修正流程图

给出完整的基于假设检验方法的 $Q$ 矩阵修正算法步骤，伪代码如下：

假设检验算法

输入：题目数 $J$，属性数 $K$，学生数 $I$，$Q=\left(q_{jk}\right)$，作答矩阵 $R=\left(r_{ij}\right)$，置信度 $\alpha$

输出：修正后的 $\tilde{Q}$ 矩阵

**For** $j=1$ **to** $J$ **do** \\ 遍历每道题目

  $Q_{[j,:]} \leftarrow Q \setminus Q_{[j;1,...,K]}$；$R_{[:,j]} \leftarrow R \setminus R_{[1,...,I;j]}$；\\ 输入数据去除第 $j$ 题

  $\tilde{g},\tilde{s},\tilde{\beta} \leftarrow DINA(Q_{[j,:]},R_{[:,j]})$；\\ 使用 DINA 模型估计参数

  $\bar{g}=\sum_{l=1}^{J-1}\tilde{g}_l/(J-1)$；$\bar{s}=\sum_{l=1}^{J-1}\tilde{s}_l/(J-1)$；\\ 对第 $j$ 题的参数估计

  $\tilde{q}_j \leftarrow copy(q_j), j=1,2,...,J$；$\tilde{Q} \leftarrow copy(Q)$ \\ 存储修正结果

  **For** $k=1$ **to** $K$ **do** \\ 遍历每个属性

    $T=\left\{i:\tilde{\beta}_i \succeq q_j, \text{且} \tilde{\beta}_{ik}=0\right\}$ \\ 样本筛选

    $n_X=n_T-\sum_{i\in T} r_{ij}$；$n_Y=\sum_{i\in T} r_{ij}$ \\ 计算错误作答数、正确作答数

    $p_{0\to1} \leftarrow P\left(X<n_X\right), p_{1\to0} \leftarrow P\left(Y<n_Y\right)$ \\ $X\sim B\left(n_T,\bar{s}\right), Y\sim B\left(n_T,\bar{g}\right)$

10. **If** $p_{0\to1} \geq 1-\alpha$ **and** $p_{1\to0} \geq 1-\alpha$ **then**
11. **If** $p_{0\to1} \geq p_{1\to0}$ **then**
12. $\tilde{q}_{jk} \leftarrow 1$ \\ 缺失的概率大于冗余的概率，判断为属性缺失
13. **Else**
14. $\tilde{q}_{jk} \leftarrow 0$
15. **Elif** $p_{0\to1} \geq 1-\alpha$ **then**
16. $\tilde{q}_{jk} \leftarrow 1$ \\ 缺失概率大于$1-\alpha$，判断为属性缺失
17. **Elif** $p_{1\to0} \geq 1-\alpha$ **then**
18. $\tilde{q}_{jk} \leftarrow 0$ \\冗余概率大于$1-\alpha$，判断为属性冗余
19. **Else**
20. Continue // 均不通过检验则不做修正
21. **If** $\| \tilde{q}_j \| = 0$ **then**
22. $\tilde{q}_{jk^*} \leftarrow 1$,其中 $k^* = \arg\max p_{0\to1}^k$ // 出现 0 向量，则 $p$ 值最小的位置取 1
23. $\tilde{Q} \leftarrow \tilde{q}_j, j = 1,2,...,J$

# 3 研究一：假设检验法中置信度$\alpha$选择及其可行性和准确性

为了考察本文提出的假设检验法中的置信度$\alpha$及其可行性和对$Q$矩阵修正的准确性，采用 3 因素实验设计(4×4×3)，分别为被试作答失误率(5%,10%,15%,20%)，$Q$矩阵错误率(5%,10%,15%,20%)和置信度$\alpha$ (0.01,0.05,0.1)。其中被试作答失误率指在 Leighton 等人(2004)介绍的理想模式作答的情况下(即无猜测和失误)，根据一定的失误比率模拟被试的作答反应，生成被试得分矩阵[16]。在实验中指定属性数量为 5 个，属性间不可补偿且相互独立，共$2^5-1=31$种所有可能的项目考核模式。

## 3.1Monte Carlo 模拟过程

模拟过程具体包括四个步骤，包括$Q$矩阵真值、错误的$Q$矩阵的生成、学生掌握模式的模拟以及作答记录的模拟，模拟方法参考 Leighton(2004)、涂东波(2012)、Nájera(2021)等人介绍的模拟方法[16][7][17]。

(1)真值$Q$矩阵生成

为了适应$Q$矩阵的多样性变化，本文依据指定考察属性数量的考察模式占比来生成相应考察模式，以确保评估的全面性。例如，5 个属性存在 31 种不同的考察模式，而掌握 1

个属性的模式共有 $C_5^1$ 种，其占全部模式的比例为 5/31=16.12%，则按这一比例生成相应数量和类型的考察模式，以维持模拟试验均衡性和代表性。项目总数设置参考 Nájera(2020)，设定项目属性比为 8。

(2)错误 $Q$ 矩阵模拟

参考涂东波(2012)的设计，根据(1)中的 $Q$ 矩阵，分别生成错误率为 5%,10%,15%,20% 下的错误 $Q$ 矩阵，其中错误类型具体分为属性冗余和属性缺失，且这两种情况错误率相等，但对于错误项目与错误属性的选择是完全随机的。

(3)学生掌握模式

学生掌握模式与 $Q$ 矩阵真值生成模式相似，对指定掌握属性数量的掌握模式占比来生成相应的掌握模式，共模拟 1000 人，不同在于学生掌握模式有 32 种。

(4) 学生作答矩阵

采用 Leighton(2004)的模拟方法，即根据(3)的掌握模式计算理想作答，对作答矩阵分别模拟 5%、10%、15%、20%的错误比例。

## 3.2 评价指标

本节将生成的错误 $Q$ 矩阵(记为 $Q^w$ 矩阵)和作答矩阵 $R$ 进行 *DINA* 模型诊断分析，根据诊断结果采用假设检验法对 $Q^w$ 进行修正(修正后的 $Q$ 矩阵记为 $\tilde{Q}$)，并进一步计算修改前后的正确率(模式判准率、属性判准率)，同时计算知识结构的保真度和纠错度(正确属性保留率、错误属性修正率)，从而全方位评估假设检验法修正的可行性、准确性和可靠性，所有试验均重复 100 次，然后再计算 100 次试验的平均 PMR、AMR、TPR 及 FPR。

(1)模式判准率(PMR)：该指标从项目粒度评估修正方法的修正准确性，即修正后的项目 $j$ 考核模式 $\tilde{q}_j$ 是否和该项目的真实考核模式 $q_j$ 相同，其表达式如(8)：

$$PMR = \sum_{j=1}^{J} I(\tilde{q}_{jt} = q_{jt}) \Big/ J \tag{8}$$

(2)属性判准率(AMR)：该指标从属性粒度评估修正方法的准确性，其表示对整个 $Q$ 矩阵而言，修正后的元素 $\tilde{q}_{jk}$ 是否和真实元素 $q_{jk}$ 一致，表达式如(9)：

$$AMR = \sum_{j=1}^{J}\sum_{k=1}^{K} I(\tilde{q}_{jk} = q_{jk}) \Big/ JK \tag{9}$$

(3)正确属性保留率(TPR)：该指标衡量了在 $Q$ 矩阵元素正确的情况下，经过修正后仍然保持其正确的比例。TPR 值越高，意味着模型在保护既有的知识结构方面表现越佳，同

时展现出更高的稳定性和可靠性，以及更精准的预测能力。其表达式如(10)：

$$TPR=\frac{\sum_{j=1}^{J}\sum_{k=1}^{K}I\left(\tilde{q}_{jk}=q_{jk}\mid q_{jk}^{w}=q_{jk}\right)}{\sum_{j=1}^{J}\sum_{k=1}^{K}I\left(q_{jk}^{w}=q_{jk}\right)} \tag{10}$$

其中$q_{jk}$，$q_{jk}^{w}$，$\tilde{q}_{jk}$分别真值$Q$，错误$Q^{w}$与修正后$\tilde{Q}$的元素。$q_{jk}^{w}=q_{jk}$表示$Q^{w}$矩阵$(j,k)$元素无误，$\tilde{q}_{jk}=q_{jk}\mid q_{jk}^{w}=q_{jk}$表示在$Q^{w}$矩阵$(j,k)$元素正确情况下该元素不修正。

(4)错误属性修正率(FPR)：该指标表示在$Q$矩阵中所有错误元素修正回来的比例。其表达式为(11)，其中$\tilde{q}_{jk}=q_{jk}\mid q_{jk}^{w}\neq q_{jk}$表示在$Q$矩阵$(j,k)$元素有误时该元素正确修正。

$$FPR=\frac{\sum_{j=1}^{J}\sum_{k=1}^{K}I\left(\tilde{q}_{jk}=q_{jk}\mid q_{jk}^{w}\neq q_{jk}\right)}{\sum_{j=1}^{J}\sum_{k=1}^{K}I\left(q_{jk}^{w}\neq q_{jk}\right)} \tag{11}$$

## 3.3 试验结果

表 1 至表 4 是在作答失误分别是 5%，10%，15%和 20%下假设检验法对$Q$矩阵的修正情况。其中$Q^{w}$PMR、$Q^{w}$AMR 表示错误的$Q^{w}$相较于真值$Q$的模式准确率与属性准确率，$\tilde{Q}$PMR、$\tilde{Q}$AMR 指采用假设检验法修正后矩阵的$\tilde{Q}$的模式准确率和属性准确率，提高率为修正后$\tilde{Q}$准确率与错误$Q^{w}$的准确率之差。TPR 与 FPR 为修正后$\tilde{Q}$矩阵的正确属性保留率和错误属性修正率。

分析表 1 至表 4 的数据可看出，无论在何种预设的作答失误率水平上，采用的假设检验法均能有效修正$Q^{w}$，修正后的准确率明显提高，且随着作答失误率降低，其判别能力越精确，准确率趋于理想的 100%；在设置置信水平分别为 0.01、0.05 及 0.1 的情况下，经过严谨的假设检验分析，我们可以观察到该方法的校正性能表现出高度的一致性，意味着该算法对于超参数置信度的选取具有优异的稳健性，这不仅验证了统计理论的可靠性，亦展现了广泛的适应性；在$Q$矩阵错误率较低时(尤其在错误率小于 10%时，见表 1、表 2)，假设检验法的属性准确率基本能达到 100%，表明该方法具有较强的识别能力。

表格 1 假设检验法在作答失误为 5%的情况下的 100 次试验平均结果

| Q 矩阵失误率 | 置信度 $\alpha$ | $Q^{w}$PMR | $\tilde{Q}$PMR | 提高率 | $Q^{w}$AMR | $\tilde{Q}$AMR | 提高率 | TPR | FPR |
|---|---|---|---|---|---|---|---|---|---|
| 5% | 0.01 | 0.794 | 1.000 | 0.207 | 0.955 | 1.000 | 0.045 | 1.000 | 1.000 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0.05 | 0.794 | 1.000 | 0.207 | 0.955 | 1.000 | 0.045 | 1.000 | 1.000 |
| | 0.1 | 0.794 | 1.000 | 0.207 | 0.955 | 1.000 | 0.045 | 1.000 | 1.000 |
| | 0.01 | 0.605 | 0.999 | 0.394 | 0.905 | 1.000 | 0.095 | 1.000 | 1.000 |
| 10% | 0.05 | 0.605 | 0.999 | 0.394 | 0.905 | 1.000 | 0.095 | 1.000 | 1.000 |
| | 0.1 | 0.605 | 0.999 | 0.394 | 0.905 | 1.000 | 0.095 | 1.000 | 1.000 |
| | 0.01 | 0.459 | 0.993 | 0.535 | 0.855 | 0.999 | 0.144 | 0.999 | 0.999 |
| 15% | 0.05 | 0.459 | 0.993 | 0.534 | 0.855 | 0.999 | 0.144 | 0.999 | 0.999 |
| | 0.1 | 0.459 | 0.993 | 0.534 | 0.855 | 0.999 | 0.144 | 0.999 | 0.999 |
| | 0.01 | 0.347 | 0.981 | 0.635 | 0.804 | 0.996 | 0.192 | 0.996 | 0.997 |
| 20% | 0.05 | 0.347 | 0.981 | 0.635 | 0.804 | 0.996 | 0.192 | 0.996 | 0.997 |
| | 0.1 | 0.347 | 0.981 | 0.634 | 0.804 | 0.996 | 0.192 | 0.996 | 0.997 |

表格 2 假设检验法在作答失误为 10%的情况下的 100 次试验平均结果

| Q 矩阵失误率 | 置信度 $\alpha$ | $Q^w$ PMR | $\tilde{Q}$ PMR | 提高率 | $Q^w$ AMR | $\tilde{Q}$ AMR | 提高率 | TPR | FPR |
|---|---|---|---|---|---|---|---|---|---|
| 5% | 0.01 | 0.794 | 1.000 | 0.207 | 0.955 | 1.000 | 0.045 | 1.000 | 1.000 |
| | 0.05 | 0.794 | 1.000 | 0.207 | 0.955 | 1.000 | 0.045 | 1.000 | 1.000 |
| | 0.1 | 0.794 | 1.000 | 0.207 | 0.955 | 1.000 | 0.045 | 1.000 | 1.000 |
| 10% | 0.01 | 0.605 | 0.998 | 0.393 | 0.905 | 1.000 | 0.095 | 1.000 | 1.000 |
| | 0.05 | 0.605 | 0.998 | 0.393 | 0.905 | 1.000 | 0.095 | 1.000 | 1.000 |
| | 0.1 | 0.605 | 0.998 | 0.393 | 0.905 | 1.000 | 0.095 | 1.000 | 1.000 |
| 15% | 0.01 | 0.459 | 0.991 | 0.532 | 0.855 | 0.998 | 0.143 | 0.998 | 0.998 |
| | 0.05 | 0.459 | 0.991 | 0.532 | 0.855 | 0.998 | 0.143 | 0.998 | 0.998 |
| | 0.1 | 0.459 | 0.991 | 0.532 | 0.855 | 0.998 | 0.143 | 0.998 | 0.998 |
| 20% | 0.01 | 0.347 | 0.976 | 0.629 | 0.804 | 0.995 | 0.191 | 0.995 | 0.994 |
| | 0.05 | 0.347 | 0.976 | 0.630 | 0.804 | 0.995 | 0.191 | 0.995 | 0.994 |
| | 0.1 | 0.347 | 0.976 | 0.630 | 0.804 | 0.995 | 0.191 | 0.995 | 0.994 |

表格 3 假设检验法在作答失误为 15%的情况下的 100 次试验平均结果

| Q 矩阵失误率 | 置信度 $\alpha$ | $Q^w$ PMR | $\tilde{Q}$ PMR | 提高率 | $Q^w$ AMR | $\tilde{Q}$ AMR | 提高率 | TPR | FPR |
|---|---|---|---|---|---|---|---|---|---|
| 5% | 0.01 | 0.794 | 1.000 | 0.206 | 0.955 | 1.000 | 0.045 | 1.000 | 1.000 |
| | 0.05 | 0.794 | 1.000 | 0.206 | 0.955 | 1.000 | 0.045 | 1.000 | 1.000 |
| | 0.1 | 0.794 | 1.000 | 0.206 | 0.955 | 1.000 | 0.045 | 1.000 | 1.000 |
| 10% | 0.01 | 0.605 | 0.995 | 0.390 | 0.905 | 0.999 | 0.094 | 0.999 | 0.999 |
| | 0.05 | 0.605 | 0.995 | 0.390 | 0.905 | 0.999 | 0.094 | 0.999 | 0.999 |
| | 0.1 | 0.605 | 0.995 | 0.390 | 0.905 | 0.999 | 0.094 | 0.999 | 0.999 |
| 15% | 0.01 | 0.459 | 0.983 | 0.525 | 0.855 | 0.997 | 0.142 | 0.997 | 0.997 |
| | 0.05 | 0.459 | 0.983 | 0.524 | 0.855 | 0.997 | 0.142 | 0.997 | 0.997 |
| | 0.1 | 0.459 | 0.983 | 0.524 | 0.855 | 0.997 | 0.142 | 0.997 | 0.997 |
| 20% | 0.01 | 0.347 | 0.950 | 0.603 | 0.804 | 0.990 | 0.186 | 0.990 | 0.990 |
| | 0.05 | 0.347 | 0.950 | 0.603 | 0.804 | 0.990 | 0.186 | 0.990 | 0.991 |
| | 0.1 | 0.347 | 0.951 | 0.604 | 0.804 | 0.990 | 0.186 | 0.990 | 0.991 |

表格 4 假设检验法在作答失误为 20%的情况下的 100 次试验平均结果

| Q 矩阵失误率 | 置信度 $\alpha$ | $Q^w$ PMR | $\tilde{Q}$ PMR | 提高率 | $Q^w$ AMR | $\tilde{Q}$ AMR | 提高率 | TPR | FPR |
|---|---|---|---|---|---|---|---|---|---|
| 5% | 0.01 | 0.794 | 0.995 | 0.202 | 0.955 | 0.999 | 0.044 | 0.999 | 0.999 |
| | 0.05 | 0.794 | 0.995 | 0.202 | 0.955 | 0.999 | 0.044 | 0.999 | 0.999 |
| | 0.1 | 0.794 | 0.995 | 0.202 | 0.955 | 0.999 | 0.044 | 0.999 | 0.999 |
| 10% | 0.01 | 0.605 | 0.980 | 0.375 | 0.905 | 0.996 | 0.091 | 0.996 | 0.996 |
| | 0.05 | 0.605 | 0.980 | 0.375 | 0.905 | 0.996 | 0.091 | 0.996 | 0.996 |
| | 0.1 | 0.605 | 0.980 | 0.375 | 0.905 | 0.996 | 0.091 | 0.996 | 0.996 |
| 15% | 0.01 | 0.459 | 0.942 | 0.484 | 0.855 | 0.988 | 0.133 | 0.989 | 0.986 |
| | 0.05 | 0.459 | 0.942 | 0.484 | 0.855 | 0.988 | 0.133 | 0.989 | 0.986 |
| | 0.1 | 0.459 | 0.942 | 0.484 | 0.855 | 0.988 | 0.133 | 0.989 | 0.986 |
| 20% | 0.01 | 0.347 | 0.880 | 0.533 | 0.804 | 0.975 | 0.172 | 0.976 | 0.973 |
| | 0.05 | 0.347 | 0.880 | 0.533 | 0.804 | 0.975 | 0.172 | 0.976 | 0.974 |
| | 0.1 | 0.347 | 0.879 | 0.532 | 0.804 | 0.975 | 0.171 | 0.976 | 0.974 |

# 4 研究二：假设检验法与其他算法的比较研究

在初步剖析了假设检验法的校正性能后，本研究深化探索，与其他同类算法对比分析——特别是基于 DINA 模型框架下的$\delta$法和$\gamma$法，旨在全面评估假设检验法的可行性与准确性，实验结果见表 5 与表 6。

**(1)小样本挑战下的修正性能比较：稳健性与偏差分析**

在小样本情形下，$\delta$法与$\gamma$法会面临"逆向修正"问题，即修改后$Q$矩阵中的错误非减反增，特别在高作答错误率时这一现象尤为突出，直观的展示了在样本少、题目质量较差的条件下，这些方法容易陷入"欠拟合"困境。根本原因在于少量样本导致的效应大小(effect size, ES)、猜测参数$g$与失误参数$s$估计偏差，严重影响了模式与属性的有效识别，从而引起误判，这是不可接受的。相比之下，假设检验法在模式和属性准确率上有出色的表现，这得益于其遵循小概率原理，审慎对待原假设的拒绝，力图维护原本正确的属性不变，仅在有足够的把握时才实施修正。因此，相较于$\delta$法与$\gamma$法，假设检验法受样本量的影响较小，有较强的稳健性。

**(2)大样本情景中的精度评估与适应性**

随着样本量增加，$\delta$法与$\gamma$法的"欠拟合"的现象缓解，准确率显著提升；虽然假设检验法与它们的差异会变小，甚至在特定条件下(如样本量为 2000，$Q$矩阵错误率为 20%时)，$\delta$法表现更优，然而，假设检验法基于深厚的理论支撑和广泛地适用性，在模式和属性准确率上始终维持着高水平和高度的稳定性；在大样本情形下，所有算法对作答失误率均表现出一定鲁棒性，表明题目质量并不是影响修正效能的关键。

**(3)属性处理优势及整体效能评价：保护正确和精准修正**

$Q$矩阵错误率一般较低，这意味着大多数属性是正确的，而错误的属性仅占少数。假设检验展现出显著的优势：它能极好地保留原本正确的属性，正如表 6 所示，其正确属性保持率名列前茅。尽管在修正错误属性方面，其表现可能并非总是最佳，但鉴于 $Q$矩阵的特点，这种偏重于保护正确信息的方法使得其整体效能依然领先。

总的来说，假设检验方法在修正 $Q$矩阵的准确率方面展现出独特的优势，这一特性在处理样本量相对有限且数据质量问题较为突出的情形下显得尤为重要。假设检验法的核心不仅在于提供了系统性的框架判断观测结果是否纯粹由随机变异引起，更在于其能有效地减少由样本不足或项目噪声导致地误判风险。

表格 5 假设检验法与其他算法在不同参数下的 100 次试验平均结果

| $Q$ 矩阵错误率 | 作答错误率 | 算法 \ 样本量 | PMR(模式准确率) | | | | AMR(属性准确率) | | | | TPR(正确属性保留率) | | | | FPR(错误属性修正率) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 100 | 300 | 1000 | 2000 | 100 | 300 | 1000 | 2000 | 100 | 300 | 1000 | 2000 | 100 | 300 | 1000 | 2000 |
| 5% | 10% | $Q^w$ | 0.800 | 0.775 | 0.792 | 0.800 | 0.955 | 0.955 | 0.955 | 0.955 | | | | | | | | |
| | | 假设检验 | **0.951** | **0.987** | **1.000** | **0.975** | **0.990** | **0.998** | **1.000** | **0.990** | **0.995** | **0.997** | **1.000** | 0.989 | 0.890 | **1.000** | **1.000** | **1.000** |
| | | $\delta$ | 0.701↓ | 0.887 | 0.966 | 0.951 | 0.935↓ | 0.970 | 0.993 | **0.990** | 0.932 | 0.968 | 0.993 | 0.990 | **0.999** | **1.000** | **1.000** | **1.000** |
| | | $\gamma$ | 0.875 | 0.925 | 0.917 | 0.949 | 0.975 | 0.985 | 0.983 | **0.990** | 0.984 | 0.995 | 0.995 | **1.000** | 0.776 | 0.774 | 0.741 | 0.777 |
| | 15% | $Q^w$ | 0.776 | 0.800 | 0.800 | 0.775 | 0.955 | 0.955 | 0.955 | 0.955 | | | | | | | | |
| | | 假设检验 | **0.901** | **0.975** | **0.976** | **0.999** | **0.980** | **0.995** | **0.995** | **1.000** | **0.990** | 0.995 | 0.995 | **1.000** | 0.779 | **1.000** | **1.000** | 0.999 |
| | | $\delta$ | 0.676↓ | 0.813 | 0.852 | **0.999** | 0.920↓ | 0.960 | 0.951↓ | **1.000** | 0.916 | 0.958 | 0.954 | **1.000** | **0.999** | 0.999 | 0.891 | **1.000** |
| | | $\gamma$ | 0.800 | 0.924 | 0.901 | 0.950 | 0.960 | 0.985 | 0.980 | 0.990 | 0.979 | **0.997** | **1.000** | **1.000** | 0.556 | 0.720 | 0.560 | 0.776 |
| | 20% | $Q^w$ | 0.800 | 0.812 | 0.788 | 0.788 | 0.955 | 0.955 | 0.955 | 0.955 | | | | | | | | |
| | | 假设检验 | **0.961** | **0.938** | **1.000** | **0.963** | **0.987** | **0.988** | **1.000** | **0.986** | **1.000** | **0.987** | **1.000** | 0.985 | 0.721 | **0.999** | **1.000** | **1.000** |
| | | $\delta$ | 0.623↓ | 0.726↓ | 0.937 | 0.921 | 0.875↓ | 0.916↓ | 0.977 | 0.977 | 0.876 | 0.912 | 0.979 | 0.976 | **0.834** | 0.996 | 0.944 | **1.000** |
| | | $\gamma$ | 0.664↓ | 0.886 | 0.926 | 0.947 | 0.888↓ | 0.963 | 0.973 | 0.985 | 0.896 | 0.966 | 0.979 | **0.994** | 0.723 | 0.884 | 0.834 | 0.801 |
| 10% | 10% | $Q^w$ | 0.600 | 0.625 | 0.592 | 0.576 | 0.905 | 0.905 | 0.905 | 0.905 | | | | | | | | |
| | | 假设检验 | **0.901** | **1.000** | **1.000** | **0.951** | **0.980** | **1.000** | **1.000** | 0.980 | **1.000** | **1.000** | **1.000** | 0.984 | 0.791 | **1.000** | **1.000** | 0.948 |
| | | $\delta$ | 0.751 | 0.950 | 0.950 | **0.951** | 0.945 | 0.982 | 0.988 | **0.990** | 0.939 | 0.981 | 0.987 | 0.989 | **1.000** | **1.000** | **1.000** | **1.000** |
| | | $\gamma$ | 0.799 | 0.849 | 0.850 | 0.924 | 0.960 | 0.970 | 0.968 | 0.985 | 0.983 | 0.997 | 0.996 | **1.000** | 0.736 | 0.709 | 0.703 | 0.841 |
| | 15% | $Q^w$ | 0.599 | 0.600 | 0.625 | 0.551 | 0.905 | 0.905 | 0.905 | 0.905 | | | | | | | | |
| | | 假设检验 | **0.826** | **0.939** | **0.951** | **0.949** | **0.965** | **0.985** | **0.990** | **0.980** | **0.983** | **0.989** | **0.989** | 0.983 | 0.789 | **0.948** | **1.000** | 0.948 |
| | | $\delta$ | 0.702 | 0.801 | 0.902 | 0.926 | 0.935 | 0.952 | 0.975 | 0.956 | 0.945 | 0.953 | 0.978 | 0.951 | **0.844** | 0.947 | 0.948 | **1.000** |
| | | $\gamma$ | 0.675 | 0.788 | 0.826 | 0.826 | 0.930 | 0.953 | 0.960 | 0.965 | 0.956 | 0.986 | 0.984 | **0.995** | 0.682 | 0.633 | 0.738 | 0.686 |
| | 20% | $Q^w$ | 0.674 | 0.625 | 0.625 | 0.609 | 0.905 | 0.905 | 0.905 | 0.905 | | | | | | | | |
| | | 假设检验 | **0.850** | **0.900** | **0.986** | **0.963** | **0.960** | **0.980** | **0.997** | **0.988** | **0.989** | **0.978** | **1.000** | 0.988 | 0.684 | **0.998** | **0.973** | **0.982** |
| | | $\delta$ | 0.611↓ | 0.650 | 0.838 | 0.889 | 0.892↓ | 0.898↓ | 0.928 | 0.965 | 0.892 | 0.887 | 0.923 | 0.966 | **0.896** | 0.996 | **0.973** | 0.957 |
| | | $\gamma$ | 0.638↓ | 0.762 | 0.861 | 0.887 | 0.880↓ | 0.928 | 0.955 | 0.973 | 0.901 | 0.938 | 0.978 | **0.991** | 0.685 | 0.836 | 0.737 | 0.806 |

表格 6 假设检验法与其他算法在不同参数下的 100 次试验平均结果(续表)

| $Q$ 矩阵错误率 | 作答错误率 | 算法 \ 样本量 | PMR(模式准确率) | | | | AMR(属性准确率) | | | | TPR(正确属性保留率) | | | | FPR(错误属性修正率) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 100 | 300 | 1000 | 2000 | 100 | 300 | 1000 | 2000 | 100 | 300 | 1000 | 2000 | 100 | 300 | 1000 | 2000 |
| 15% | 10% | $Q^w$ | 0.426 | 0.499 | 0.467 | 0.401 | 0.855 | 0.855 | 0.855 | 0.855 | | | | | | | | |
| | | 假设检验 | **0.875** | **1.000** | **0.992** | 0.827 | **0.965** | **1.000** | **0.998** | 0.950 | **1.000** | **1.000** | **0.998** | 0.954 | 0.759 | **1.000** | **1.000** | 0.932 |
| | | $\delta$ | 0.700 | 0.912 | 0.941 | **0.926** | 0.925 | 0.975 | 0.986 | **0.985** | 0.936 | 0.979 | 0.988 | 0.983 | **0.863** | 0.949 | 0.976 | **1.000** |
| | | $\gamma$ | 0.651 | 0.662 | 0.758 | 0.775 | 0.920 | 0.922 | 0.950 | 0.945 | 0.971 | 0.974 | 0.992 | **0.994** | 0.622 | 0.619 | 0.701 | 0.656 |
| | 15% | $Q^w$ | 0.376 | 0.462 | 0.425 | 0.450 | 0.855 | 0.855 | 0.855 | 0.855 | | | | | | | | |
| | | 假设检验 | **0.823** | **0.938** | **0.927** | **0.973** | **0.960** | **0.988** | **0.985** | **0.990** | **0.994** | **0.988** | **0.994** | **0.988** | 0.759 | **0.983** | **0.933** | 0.999 |
| | | $\delta$ | 0.626 | 0.812 | 0.829 | 0.925 | 0.905 | 0.957 | 0.937 | 0.955 | 0.929 | 0.956 | 0.943 | 0.948 | **0.762** | 0.965 | 0.899 | **1.000** |
| | | $\gamma$ | 0.525 | 0.675 | 0.701 | 0.800 | 0.875 | 0.922 | 0.921 | 0.960 | 0.936 | 0.973 | 0.971 | **0.988** | 0.517 | 0.621 | 0.623 | 0.792 |
| | 20% | $Q^w$ | 0.500 | 0.512 | 0.525 | 0.459 | 0.855 | 0.855 | 0.855 | 0.855 | | | | | | | | |
| | | 假设检验 | **0.675** | **0.815** | **0.948** | **0.912** | **0.908** | **0.958** | **0.990** | **0.977** | **0.974** | **0.957** | **0.991** | 0.979 | 0.517 | 0.965 | **0.982** | **0.959** |
| | | $\delta$ | 0.500 | 0.528 | 0.863 | 0.847 | 0.830⬇ | 0.875 | 0.957 | 0.958 | 0.842 | 0.855 | 0.953 | 0.961 | **0.761** | **0.997** | **0.982** | 0.937 |
| | | $\gamma$ | 0.464⬇ | 0.614 | 0.764 | 0.799 | 0.835⬇ | 0.869 | 0.935 | 0.949 | 0.880 | 0.905 | 0.971 | **0.988** | 0.571 | 0.657 | 0.726 | 0.719 |
| 20% | 10% | $Q^w$ | 0.326 | 0.424 | 0.334 | 0.326 | 0.800 | 0.805 | 0.803 | 0.805 | | | | | | | | |
| | | 假设检验 | **0.800** | **0.950** | **0.893** | 0.875 | **0.950** | **0.988** | 0.977 | 0.960 | **0.988** | **0.985** | **0.981** | 0.963 | 0.800 | **1.000** | 0.958 | 0.949 |
| | | $\delta$ | 0.725 | 0.839 | 0.891 | **0.950** | 0.940 | 0.965 | **0.978** | **0.990** | 0.950 | 0.969 | **0.981** | **0.988** | **0.900** | 0.949 | **0.965** | **0.999** |
| | | $\gamma$ | 0.575 | 0.588 | 0.592 | 0.773 | 0.885 | 0.890 | 0.908 | 0.940 | 0.957 | 0.969 | 0.967 | 0.987 | 0.600 | 0.564 | 0.669 | 0.743 |
| | 15% | $Q^w$ | 0.276 | 0.387 | 0.275 | 0.325 | 0.800 | 0.805 | 0.800 | 0.805 | | | | | | | | |
| | | 假设检验 | **0.626** | **0.900** | **0.901** | **0.923** | **0.915** | **0.980** | **0.980** | 0.970 | **0.981** | **0.987** | **0.981** | 0.962 | 0.650 | **0.949** | **0.975** | **0.999** |
| | | $\delta$ | 0.553 | 0.714 | 0.852 | **0.923** | 0.876 | 0.925 | 0.961 | **0.984** | 0.906 | 0.932 | 0.957 | **0.987** | **0.754** | 0.899 | **0.975** | 0.975 |
| | | $\gamma$ | 0.500 | 0.649 | 0.573 | 0.600 | 0.850 | 0.922 | 0.890 | 0.905 | 0.925 | **0.987** | 0.975 | 0.963 | 0.551 | 0.653 | 0.550 | 0.666 |
| | 20% | $Q^w$ | 0.300 | 0.411 | 0.375 | 0.326 | 0.805 | 0.805 | 0.803 | 0.803 | | | | | | | | |
| | | 假设检验 | **0.513** | **0.675** | **0.823** | **0.847** | **0.860** | **0.932** | **0.962** | **0.961** | **0.956** | **0.935** | 0.965 | 0.963 | 0.462 | **0.921** | 0.949 | **0.955** |
| | | $\delta$ | 0.450 | 0.536 | 0.799 | 0.810 | 0.803⬇ | 0.865 | 0.932 | 0.949 | 0.804 | 0.854 | 0.928 | 0.949 | **0.796** | 0.908 | **0.950** | 0.950 |
| | | $\gamma$ | 0.364 | 0.488 | 0.637 | 0.681 | 0.810 | 0.846 | 0.905 | 0.924 | 0.879 | 0.893 | **0.975** | **0.985** | 0.527 | 0.652 | 0.623 | 0.675 |

# 5 研究三：实证数据研究

为了验证假设检验法在实证数据中的效果及与$\delta$法和$\gamma$法进行实证对比分析，本研究选取两个典型数据集：Tatsuoka(1990)分数减法数据[错误!未找到引用源。]与 TIMSS2007 数据，分别从修正方法拟合度、项目拟合度两个方面探讨本研究在实际应用中的性能。本研究中两个数据对应的$Q$矩阵如表 7 与表 8。

## 5.1 数据集

Tatsuoka 数据简称 FraSub，其涵盖了 536 名学生的测试表现，他们在涉及 5 个认知领域的 15 项分数减法任务中接受了评估，该数据此前已被 Tatsuoka(1984)及 de la Torre(2008)等诸多学者深入测量与解析[6][19]。而 TIMSS 2007 则针对奥地利 698 名四年级生，通过 25 题覆盖 15 个认知维度，且被 Lee, Park, & Taylan (2011), Park & Lee (2014), 以及 Soo(2018)等多篇研究用于探索教育评估和认知诊断领域[错误!未找到引用源。]。

表格 7 FraSub 数据集$Q$矩阵

| | Attribute | | | | |
|---|---|---|---|---|---|
| Item | A1 | A2 | A3 | A4 | A5 |
| T01 | 1 | 0 | 0 | 0 | 0 |
| T02 | 1 | 1 | 1 | 1 | 0 |
| T03 | 1 | 0 | 0 | 0 | 0 |
| T04 | 1 | 1 | 1 | 1 | 1 |
| T05 | 0 | 0 | 1 | 0 | 0 |
| T06 | 1 | 1 | 1 | 1 | 0 |
| T07 | 1 | 1 | 1 | 1 | 0 |
| T08 | 1 | 1 | 0 | 0 | 0 |
| T09 | 1 | 0 | 1 | 0 | 0 |
| T10 | 1 | 0 | 1 | 1 | 1 |
| T11 | 1 | 0 | 1 | 0 | 0 |
| T12 | 1 | 0 | 1 | 1 | 0 |
| T13 | 1 | 1 | 1 | 1 | 0 |
| T14 | 1 | 1 | 1 | 1 | 1 |
| T15 | 1 | 1 | 1 | 1 | 0 |

注：A1 进行基本的分数减法运算，A2 化简和约简，A3 从分数中分离整数，A4 从整数借 1 到分数，A5 将整数化为分数。

表格 8 TIMSS2007 数据集 $Q$ 矩阵

| Item | Attribute | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NWN | | | | NFD | | NNS | NPR | GLA | GTT | | GLM | DRI | | DOR |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| M041052 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M041056 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M041069 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M041076 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M041281 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M041164 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| M041146 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| M041152 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| M041258A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| M041258B | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| M041131 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| M041275 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| M041186 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| M041336 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| M031303 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M031309 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M031245 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M031242A | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M031242B | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| M031242C | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| M031247 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M031219 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| M031173 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M031085 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| M031172 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |

注 1: 属性编号的前缀 N, Number; G, Geometric Shapes & Measures; D, Data Display.

注 2: 属性编号后缀 WN, Whole Numbers; FD, Fractions and Decimals; NS, Number Sentences with Whole Numbers; PR, Patterns and Relationships; LA, Lines and Angles; TT, Two- and Three-dimensional Shapes; L M, Location and Movement; RI, Reading and Interpreting; OR, Organizing and Representing.

## 5.2 各修正算法整体拟合度评估

常用反映模型整体拟合的指标有两类、第一类相对拟合指标包括偏差(-2LogLikehood, -2LL)、赤池信息准则(AIC)和贝叶斯信息准则(BIC)；第二类绝对拟合指标包括 $M_2$ 检验、近似均方根误差(RMSEA)、标准均方根残差(SRMSR)。[23]

如表 9 和表 10 所示，针对两种不同的数据集，三种算法在修正 $Q$ 矩阵后，依据 DINA 模型进行拟合得到评估结果。关于超参数设置，假设检验 $\alpha$ 取 0.05； $\delta$ 法 $\varepsilon$ 取 0.01；而 $\gamma$ 法中猜测参数 $g$ 、失误参数 $s$ 、效应大小 $ES$ 的阈值均设定为 0.2[8]。

总体而言，在 FraSub 数据集上，由于只涉及 5 个属性，数量较少，因此 536 个样本能够使 $\delta$ 法达到一定程度的拟合水平，与此同时 $\gamma$ 法也初显欠拟合迹象；转至 TIMSS2007 数

据集，其考察属性数量大幅增加至 15 个，仅有的 698 个样本难以令$\delta$法和$\gamma$法得到较好的拟合效果，甚至产生拟合度不降反增的情况。对比之下，假设检验方法在所有拟合指标上表现出最佳性能，突显了其在小样本数据修正任务中的稳健性和准确性优势。

表格 9 基于三种算法修正后$Q$矩阵的拟合指标(FraSub)

| $Q$矩阵 | 相对拟合指标 | | | 绝对拟合指标 | | | | |
|---|---|---|---|---|---|---|---|---|
| | -2LL | AIC | BIC | $M_2$ | | | RMSEA | SRMSR |
| | | | | $M_2$ | $df$ | $p$ | | |
| 修正前$Q$ | 6911.59 | 7033.59 | 7294.93 | 231.750 | 59 | <.001 | 0.070 | 0.112 |
| 假设检验 | **6853.31** | **6975.31** | **7236.64** | **145.243** | 59 | <.001 | **0.052** | **0.076** |
| $\delta$法 | 6890.29 | 7012.29 | 7273.62 | 178.978 | 59 | <.001 | 0.062 | 0.101 |
| $\gamma$法 | 6971.43 | 7093.43 | 7354.76 | 231.910 | 59 | <.001 | 0.074 | 0.091 |

表格 10 基于三种算法修正后$Q$矩阵的拟合指标(TIMSS2007)

| $Q$矩阵 | 相对拟合指标 | | | 绝对拟合指标 |
|---|---|---|---|---|
| | -2LL | AIC | BIC | SRMSR |
| 修正前$Q$ | 7588.83 | 11726.83 | 21137.09 | 0.0316 |
| 假设检验 | **7559.87** | **11697.87** | **21108.14** | **0.0291** |
| $\delta$法 | 7657.49 | 11795.49 | 21205.76 | 0.0470 |
| $\gamma$法 | 7603.33 | 11741.33 | 21151.59 | 0.0330 |

注：TIMSS2007 的维度较高，估计参数较多而样本不足，导致自由度受限，无法进行有效地进行绝对拟合度指标的统计检验，故只计算相对拟合指标。

## 5.3 各修正算法项目拟合度评估

表 11 表明，$\gamma$法仅降低了题目 7 的拟合度，却在题目 3 和题目 10 中拟合不良；假设检验法与$\delta$法对某些题目均具有较好的拟合度，但前者仅 3 题欠拟合，少于$\delta$法的 5 个，印证了模拟实验中假设检验相对而言不轻易修正的特点，并且在相同欠拟合的题目上，假设检验法欠拟合程度相对较低，优于$\delta$法；除此之外，特别值得注意的是题目 4，其中$\delta$法出现显著的欠拟合情况，反而假设检验法有更优的拟合性能。

# 6 小结与讨论

## 6.1 结论

本研究提出了基于 DINA 模型的$Q$矩阵修正方法——假设检验法，采用蒙特卡洛模拟与其他同类方法进行比较，验证假设检验法的可行性和准确性，通过本文三个研究发现：

(1)假设检验法证明了各种作答错误率下的高效修正能力，显著提升$Q$矩阵准确率；算法对于超参数置信度的选取具有优异的稳健性。

(2)与国内外同类研究相比，假设检验法在小样本环境中展现出更高的稳健性和优越性

能、尤其在面临高作答失误率时，其优势更为显著；在大样本环境中，作答失误率对修正效果的影响力显著降低，得益于统计理论支撑，假设检验法依旧维持保持强劲的竞争力，相较于$\delta$法具有更简单的修正过程。

(3)在实证数据中，假设检验法不仅能增强认知诊断模型的拟合效能，而且在面对属性维度增多、样本量相对有限的复杂数据集时，相较于其他算法，其展现出更为显著的优势。

## 6.2 讨论

(1)从本研究提出的修正方法是基于 DINA 非补偿型诊断模型和独立的属性关系，因此未来还可考虑补偿情形的认知诊断模型，甚至进一步考虑直线型、收敛型或分支型等属性层级关系结构；

(2) 本研究为了避免较大的估计偏差，取其他题目的参数$g,s$平均作为第$j$题的估计，但仍然可能由样本的分布不同产生一定偏差，有待进一步研究。

(3) 尽管在拟合度上有所降低，但本研究主要目的是为$Q$矩阵修正提供辅助支持，在实际应用中还需结合相关学科领域专家的意见，确定修正的$Q$矩阵是否合理；

表格 11 基于三种散发逐题修正后$Q$矩阵的拟合指标(TIMSS2007)

| 题目 | Q矩阵 | 相对拟合指标 | | | 题目 | Q矩阵 | 相对拟合指标 | | |
|---|---|---|---|---|---|---|---|---|---|
| | | -2LL | AIC | BIC | | | -2LL | AIC | BIC |
| 1 | $Q$ | 7588.83 | 11726.83 | 21137.09 | 6 | $Q$ | 7588.83 | 11726.83 | 21137.09 |
| | $h$ | 0.53 | 0.53 | 0.53 | | $h$ | -15.23 | -15.23 | -15.22 |
| | $\delta$ | 4.87 | 4.87 | 4.87 | | $\delta$ | -18.88 | -18.88 | -18.87 |
| | $\gamma$ | 0.00 | 0.00 | 0.00 | | $\gamma$ | 0.00 | 0.00 | 0.00 |
| 2 | $Q$ | 7588.83 | 11726.83 | 21137.09 | 7 | $Q$ | 7588.83 | 11726.83 | 21137.09 |
| | $h$ | 0.00 | 0.00 | 0.00 | | $h$ | -0.87 | -0.87 | -0.87 |
| | $\delta$ | -0.87 | -0.87 | -0.86 | | $\delta$ | -0.79 | -0.79 | -0.78 |
| | $\gamma$ | 0.00 | 0.00 | 0.00 | | $\gamma$ | -1.22 | -1.22 | -1.21 |
| 3 | $Q$ | 7588.83 | 11726.83 | 21137.09 | 9 | $Q$ | 7588.83 | 11726.83 | 21137.09 |
| | $h$ | 2.49 | 2.49 | 2.49 | | $h$ | 0.00 | 0.00 | 0.00 |
| | $\delta$ | 2.88 | 2.88 | 2.89 | | $\delta$ | 1.83 | 1.83 | 1.84 |
| | $\gamma$ | 3.18 | 3.18 | 3.18 | | $\gamma$ | 0.00 | 0.00 | 0.00 |
| 4 | $Q$ | 7588.83 | 11726.83 | 21137.09 | 10 | $Q$ | 7588.83 | 11726.83 | 21137.09 |
| | $h$ | -21.99 | -21.99 | -21.99 | | $h$ | 9.96 | 9.96 | 9.96 |
| | $\delta$ | 19.31 | 19.31 | 19.31 | | $\delta$ | 7.01 | 7.01 | 7.02 |
| | $\gamma$ | 0.00 | 0.00 | 0.00 | | $\gamma$ | 13.71 | 13.71 | 13.71 |
| 5 | $Q$ | 7588.83 | 11726.83 | 21137.09 | 8 | | | | |
| | $h$ | 0.00 | 0.00 | 0.00 | | | | | |
| | $\delta$ | -32.73 | -32.73 | -32.72 | 11 | | | | |
| | $\gamma$ | 0.00 | 0.00 | 0.00 | | | | | |

注：针对题 8 与题 11，由于$\delta$法修正后的$Q$矩阵存在某一属性未被任何题目考察，故未纳入对比。

# 参考文献


[1]. Huebner A, Wang C. A Note on Comparing Examinee Classification Methods for Cognitive Diagnosis Models[J]. Educational and Psychological Measurement, 2011, 71 (2): 407-419.

[2]. DeCarlo, L.T. On the Analysis of Fraction Subtraction Data: The DINA Model, Classification, Latent Class Sizes, and the Q-Matrix[J]. Applied Psychological Measurement, 2011, 35 (1): 8-26.

[3]. De La Torre J. DINA Model and Parameter Estimation: A Didactic[J]. Journal of Educational and Behavioral Statistics, 2009, 34 (1): 115-130.

[4]. Tatsuoka K K. Cognitive Assessment: An Introduction to the Rule Space Method[M]. 2009.

[5]. Rupp A A , Templin J. The Effects of Q-Matrix Misspecification on Parameter Estimates and Classification Accuracy in the DINA Model[J]. Educational and Psychological Measurement, 2008, 68(1): 78-96.

[6]. De La Torre J. An Empirically Based Method of Q-matrix Validation for the DINA Model: Development and Applications[J]. Journal of Educational Measurement, 2008, 45 (4): 343-362.

[7]. De La Torre J, Chiu C Y. A general method of empirical Q-matrix validation[J]. Psychometrika, 2016, 81: 253-273.

[8]. 涂冬波, 蔡艳, 戴海琦. 基于 DINA 模型的 Q 矩阵修正方法[J]. 心理学报, 2012, 44(4): 558-568.

[9]. 喻晓锋, 罗照盛, 高椿雷, 等. 使用似然比 D2 统计量的题目属性定义方法[J]. 心理学报, 2015, 47(3): 417.

[10]. Chen J. A residual-based approach to validate Q-matrix specifications[J]. Applied Psychological Measurement, 2017, 41(4): 277-293.

[11]. 汪大勋, 高旭亮, 蔡艳, 等. 一种非参数化的 Q 矩阵估计方法: ICC-IR 方法开发[J]. 心理科学, 2018 (2): 466.

[12]. 汪大勋, 高旭亮, 蔡艳, 等. 一种广义的认知诊断 Q 矩阵修正新方法[J]. 心理科学, 2019 (4): 988.

[13]. Chiu C Y. Statistical refinement of the Q-matrix in cognitive diagnosis[J]. Applied Psychological Measurement, 2013, 37(8): 598-618.

[14]. 汪大勋, 高旭亮, 韩雨婷, 等. 一种简单有效的 Q 矩阵估计方法开发: 基于非参数化方法视角[J]. 心理科学, 2018, 41(1): 180.

[15]. Rupp A A, Templin J. The Effects of Q-Matrix Misspecification on Parameter Estimates and Classification Accuracy in the DINA Model[J]. Educational and Psychological Measurement, 2008, 68(1): 78-96.

[16]. Leighton J P, Gierl M J, Hunka S M. The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach[J]. Journal of Educational Measurement, 2004, 41 (3): 205-237.

[17]. Nájera P, Sorrel M A, de la Torre J, et al. Balancing fit and parsimony to improve Q-matrix validation[J]. British Journal of Mathematical and Statistical Psychology, 2021, 74: 110-130.

[18]. Tatsuoka K K. Toward an integration of item response theory and cognitive analysis[J]. Diagnostic monitoring of skill and knowledge acquisition, 1990: 543-588.

[19]. Tatsuoka K K. Analysis of Errors in Fraction Addition and Subtraction Problems. Final Report[J]. 1984.

[20]. Shu-Liang D. Q matrix and its applications in cognitive diagnosis[J]. Journal of Psychological Science, 2019 (3): 739.

[21]. Lee Y S, Park Y S, Taylan D. A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007[J]. International Journal of Testing, 2011, 11(2): 144-177.

[22]. Park Y S, Lee Y S, Xing K. Investigating the impact of item parameter drift for item response theory models with mixture distributions[J]. Frontiers in Psychology, 2016, 7: 179776.

[23]. Chen J, Jimmy D L T, Zhang Z. Relative and Absolute Fit Evaluation in Cognitive Diagnosis Modeling[J]. Journal of Educational Measurement, 2013, 50 (2): 123-140.

[24]. Liu Y, Tian W, Xin T. An application of M 2 statistic to evaluate the fit of cognitive diagnostic models[J]. Journal of Educational and Behavioral Statistics, 2016, 41(1): 3-26.

[25]. Nájera P, Sorrel M A, de la Torre J, et al. Balancing fit and parsimony to improve Q‐matrix validation[J]. British Journal of Mathematical and Statistical Psychology, 2021, 74: 110-130.