DOI:10.19603/j.cnki.1000-1190.2025.01.012 文章编号:1000-1190(2025)01-0111-14

# 基于认知诊断 DINA 模型的 $Q$ 矩阵优化:一种结合样本筛选与假设检验的新策略

李 波,胡誉骞,章 勇,田 怡*

(华中师范大学数学与统计学学院,武汉 430079)

**摘 要**:该文创新性地提出了一种基于 DINA 模型的 $Q$ 矩阵修正策略——SHT 法,并借助蒙特卡洛模拟技术,将其与现有同类方法进行深入对比,以全面评估其可行性与精确性.该方法具有下述 3 个方面优势.1) 高效修正与稳健性验证:SHT 法在不同水平的作答错误率下均展现出了卓越的修正效能,显著提升了 $Q$ 矩阵的精确度;2) 小样本与大样本环境的双重优势:与国内外同类研究成果相比,SHT 法在小样本场景下尤为突出,其稳健性和性能优势在面临高作答失误率时更加显著;3) 复杂数据集下的显著优势:在实证数据分析中,SHT 法不仅提高了认知诊断模型的拟合能力,更在处理数据维度复杂、样本相对受限的数据集时,展现出相比其他算法更为明显的优势.

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

**中图分类号**:B841 **文献标志码**:A **开放科学(资源服务)标志码(OSID)**:

《教育部 2022 年工作要点》明确提出实施教育数字化战略行动[1],呼应习近平总书记关于深化教改、全面提升教育质量的指示[2].2021 年 7 月,教育部印发的《关于推进教育新型基础设施建设构建高质量教育支撑体系的指导意见》重点强调创新信息化评价工具,促进规模化在线考试,以客观数据分析学生能力,促进教育高质量发展[3].教育数字化转型促进了教育测评的变革,尤其在人工智能赋能下,教育更加关注以测评数据驱动的学习者个性化诉求[4].但在线考试环境下的新领域、新阶段和新试题面临着小样本实时测量的冷启动问题,如何在稀疏数据下精准实施个性化服务,是教育测量理论中面临的挑战.

与经典测量理论(classical test theory,CTT)和项目反应理论(item response theory,IRT)相比,认知诊断模型(cognitive diagnosis model,CDM)能提供更详细的诊断信息,因而广受国际学术界及实践者的瞩目[5-7].作为教育测量的关键工具,CDM 能够诊断出学生对各知识领域的学习状态.该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[8].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[9],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.

为解决 $Q$ 矩阵的主观性问题以提升其准确性,国内外学者提出了多种策略来估计和校正 $Q$ 矩阵.例如,de la Torre[10] 针对 DINA(deterministic inputs, noisy “and” gate)模型设计了 $\delta$ 法,以及在此基础上拓展而来的 $\xi^2$ 法[11],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.涂东波等[12] 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,结合 $g$ 和 $s$ 参数的经验判断,评判 $Q$ 矩阵的正确性并进行必要的调整($\gamma$ 法).参数化方法还包括基于似然比 $D^2$ 统计量的方法、残差分析的方法、ICC-IR(item consistency

收稿日期:2024-06-12.
基金项目:国家自然科学基金项目(62377019);湖北省高校省级教学研究项目(2022083);华中师范大学中央高校基本科研业务费(CCNU24JC004);湖北省数学-基础研究特区试点 2024 项目.
* 通信联系人. E-mail:tianyi@ccnu.edu.cn.

criterion based on ideal response)方法以及两阶段处理策略等[13-16].在非参数化方法领域,Chiu[17]提出的RSS(residual sum of squares)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,此法不仅计算简便,修正成效亦佳.然而,目前它尚不能直接用于$Q$矩阵的初步估计.此外,汪大勋等[18]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.

但是,现有的许多开发方法在实际应用中遭遇了挑战.比如,$\gamma$法、$\delta$法、基于残差的方法和基于距离的方法等,往往需要较大的样本容量才能保证修正结果的稳定性.而且,$\delta$法等技术因为复杂的计算流程和繁琐的操作步骤,导致耗时巨大.通过实验,可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,但它们会不慎将大量原本正确的$Q$矩阵条目误标为错误,这种反向修正的问题尤为突出,实非理想解决方案.在如今在线教育和在线考试背景下,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情况,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.面对冷启动问题,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.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样本筛选机制的假设检验$Q$矩阵修正策略(select hypothesis testing,SHT).

## 1 $Q$矩阵修正方法与思路

首先提出$Q$矩阵相关的理论和参数性质,然后介绍两种有效的$Q$矩阵修正算法,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.

### 1.1 $Q$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,它以结构简单且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,仍能正确回答题目的概率;而失误参数$s$则衡量考生已经掌握了题目所有相关知识点,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.

De la Torre[19]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,$g$和$s$参数较大可能是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$i$掌握了测试题目$j$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,则该受试者解答错误的可能性较大.简言之,受试者对题目所考察的属性的掌握程度直接影响其答题表现.

### 1.2 矩阵修正算法相关研究

近年来,学者们已提出不少$Q$矩阵修正的相关研究,其中,基于DINA模型的$Q$矩阵修正算法,以国外de la Torre的$\delta$法与国内涂东波的$\gamma$法最为著名,本节作简单介绍.

1.2.1 *$\delta$法* $\delta$法基于$g$,$s$与$Q$矩阵的关系提出了$\delta$指标,以修正题目$j$为例.

1)首先对题目$j$所有可能的考核模式穷举,共有$2^K-1$种($K$为考察属性的个数);

2)将这些考核模式分别代入原$Q$矩阵,DINA模型可估计出题目$j$的$g$,$s$参数;

3)最后计算使$\delta$值最大的考核模式,即$q_j=\arg\max(\delta_{jc})$,其中,$\delta_{jc}=1-s_{jc}-g_{jc}$,$c\in\{1,2,\cdots,2^K-1\}$.

因此$\delta$法实际上是在所有备选的考核模式下,选择使得$s$和$g$最小的考核模式为修正后的考核模式.关于最优考核模式的搜索,有两种搜索方法:穷举搜索法和顺序搜索法,后者效率更快,但仍存在计算消耗较大的问题.

1.2.2 *$\gamma$法* $\gamma$法基于$g$,$s$与$Q$矩阵的关系提出了效应大小指标,以$Q$矩阵中题目$j$属性$k$为例.

1)在DINA模型下估计题目$j$的$g$,$s$参数,以及学生对知识点$j$的掌握概率.

2)将学生对知识点$j$的掌握情况分为掌握组和未掌握组.通常而言,可认为掌握概率大于0.6的为掌握组,小于0.4的为未掌握组,介于0.4与0.6的无法判断.

3)根据两组学生计算效应大小$ES_{jk}=(\bar{x}_1-\bar{x}_2)/s_p$,其中,$s_p=\sqrt{\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$,$s_1$,$s_2$为掌握组和未掌握组的作答标准差.当$ES$较小时,说明两组学生对知识点回答情况差别很小,很可能未考察属性$k$,因此可能冗余,反之则可能缺失.

4）以 $ES_{jk}$ 值取 0.2 为临界值时，$Q$ 矩阵的修正规则如下：

a）当 $g_j > 0.2$ 且 $ES_{jk} < 0.2$ 时，$Q(j,k) = 1 \to 0$；

b）当 $s_j > 0.2$ 且 $ES_{jk} > 0.2$ 时，$Q(j,k) = 0 \to 1$.

基于以上公式可以发现，$\gamma$ 法的效果受到临界值的影响，临界值的选取并非使用所有实际场景，故 $\gamma$ 法在不同模型下的表现可能不够稳定.

### 1.3 修正 $Q$ 矩阵的假设检验理论

1.3.1 符号定义　$I$ 为被试数量，$J$ 为题目数量，$K$ 为属性数量. $R = (r_{ij})_{I\times J}$ 为作答矩阵，$r_{ij}$ 表示被试 $i$ 对题目 $j$ 的作答对错，正确为 1，错误为 0.

$Q$ 矩阵：

$$Q = (q_{jk})_{J\times K} = \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_J \end{pmatrix};$$

被试掌握情况矩阵：

$$\beta = (\beta_{ik})_{I\times K} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_I \end{pmatrix}.$$

$Q^{\omega}$ 为错误的 $Q$ 矩阵，$\widetilde{Q}$ 为修正后的 $Q$ 矩阵. $q_j = (q_{j1},\cdots,q_{jK})$ 为题目 $j$ 对 $K$ 个属性的考察情况，考察记为 1，未考察记为 0.

$\beta_i = (\beta_{i1},\cdots,\beta_{iK})$ 为被试 $i$ 对 $K$ 个属性的掌握情况，掌握记为 1，未掌握记为 0.

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

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

定义运算

$$Q\backslash Q_{[j;1,\cdots,K]} = \begin{pmatrix} q_1 \\ \vdots \\ q_{i-1} \\ q_{i+1} \\ \vdots \\ q_J \end{pmatrix}_{(J-1)\times 1},$$

表示 $Q$ 矩阵去除第 $j$ 行 $q_j$ 向量，记 $Q_{\backslash[j,:]} = Q\backslash Q_{[j;1,\cdots,K]}$；同理，$R\backslash R_{[1,\cdots,I;j]} = (r_1,\cdots,r_{j-1},r_{j+1},\cdots,r_J)$ 表示 $R$ 去除第 $j$ 列作答向量，记 $R_{\backslash[:,j]} = R\backslash R_{[1,\cdots,I;j]}$.

1.3.2 样本筛选机制　根据DINA概率模型，在已知被试的属性掌握情况 $\beta$ 和模型参数 $g,s$ 的估计基础上，设随机变量 $X,Y$ 分别为做对和做错题目 $j$ 的人数，则 $x$ 人做错，$y$ 人做对的概率如式(1) 和(2).

$$P(Y=y) = \sum_{c=1}^{C_I^y}\Big\{\prod_{i=1}^{y}P(r_{c_ij}=1)\prod_{i=y+1}^{I}[1-P(r_{c_ij}=1)]\Big\} =$$

$$\sum_{c=1}^{C_I^y}\prod_{i=1}^{y}(1-s_j)^{\eta_{c_i}}g_j^{1-\eta_{c_i}}\prod_{i=y+1}^{I}[1-(1-s_j)^{\eta_{c_i}}g_j^{1-\eta_{c_i}}],\eta_{c_i} = \prod_{k=1}^{K}\beta_{c_ik}^{q_{jk}}, \tag{1}$$

$$P(Y=x) = \sum_{c=1}^{C_I^x}\prod_{i=1}^{x}[1-(1-s_j)^{\eta_{c_i}}g_j^{1-\eta_{c_i}}]\prod_{i=x+1}^{I}(1-s_j)^{\eta_{c_i}}g_j^{1-\eta_{c_i}}, \tag{2}$$

其中，$C_I^y$ 表示对于题目 $j$，$I$ 个被试中有 $y$ 人做对的所有可能的情况种类，$C_I^x$ 同理；$r_{c_ij}$ 表示在第 $c$ 种情形下，被试 $c_i$ 的作答情况；$\eta_{c_i}$ 表示在第 $c$ 种情况下，被试 $c_i$ 是否掌握题目 $j$ 考察的所有属性. 直接采用该分布进行假设检验存在两个核心挑战：指数级增长的时间复杂度与参数估计引起的统计量偏差. 具体而言，该分布涉及大量样本的组合筛选、样本掌握模式的判断及其对应的概率计算，这些步骤时间开销巨大. 此外，参数 $g$ 和 $s$ 估计的微小偏差，在总体样本的计算下会累计放大，影响统计量的准确性，导致结果不确定性增加，可见统计量的分布是假设检验的核心部分. 本研究提出一种样本选择机制优化两个核心问题. 考虑题目 $j$，筛选方法如下：

$$T_{j1} = \{i:\beta_i \geq q_j\}, \tag{3}$$

$$T_{j2} = \{i:\beta_{ik} = 0\}, \tag{4}$$

$$T_j = \{i:\beta_{il} \geqslant q_{jl},\forall l \in [1,2,\cdots,k] \text{且} l \neq k\} \cap T_{j2}. \tag{5}$$

集合(3) 中的被试 $i$ 满足 $\forall k \in [1,2,\cdots,K]$，s.t. $\beta_{ik} \geqslant q_{jk}$，故子样本 $T_{j1}$ 可视作将总样本限制在 $\eta_i = \sum\beta_{ik}^{q_{jk}} = 1$ 情况的集合，因此公式(2) 可推导为下面公式(6)，同理，公式(1) 可推导为

$$P(Y=y) = C_{n_{T_{j1}}}^{y}(1-s_j)^y s_j^{n_{T_{j1}}-y}.$$

$$P(X=x) = \sum_{c=1}^{C_{n_{T_{j1}}}^{x}}\prod_{i=1}^{x}[1-(1-s_j)^{\eta_{c_i}}g_j^{1-\eta_{c_i}}]\prod_{i=x+1}^{n_{T_{j1}}}(1-s_j)^{\eta_{c_i}}g_j^{1-\eta_{c_i}} =$$

$$\sum_{c=1}^{C_{n_{T_{j1}}}^{x}}\prod_{i=1}^{x}s_j\prod_{i=x+1}^{n_{T_{j1}}}(1-s_j)(\text{其中}\prod\text{与样本 }i\text{ 无关})=$$

$$\sum_{c=1}^{C_{n_{T_{j1}}}^{x}}s_j^x(1-s_j)^{n_{T_{j1}}-x}(\text{其中}\sum\text{与样本选法 }c\text{ 无关})=$$

$$C_{n_{T_{j1}}}^{x}s_j^x(1-s_j)^{n_{T_{j1}}-x}. \tag{6}$$

集合(5)中的被试 $i$ 满足 $\forall l\in[1,\cdots,k-1,k+1,\cdots,K]$,s.t.$\beta_{il}\geqslant q_{jl}$ 且 $\beta_{ik}=0$,除属性 $k$ 外,被试 $i$ 至少掌握题目 $j$ 考察的所有其余的属性.若 $q_{jk}=1$,则样本 $T_j$ 限制在 $\eta_i=\sum\beta_{ik}^{q_{jk}}=0$ 的情况,故公式(1)可推导为下面公式(7),同理,公式(2)可推导为

$$P(X=x)=C_{n_{T_j}}^{x}(1-g_j)^x g_j^{n_{T_j}-x}.$$

$$P(Y=y)=\sum_{c=1}^{C_{n_{T_j}}^{y}}\prod_{i=1}^{y}(1-s_j)^{\eta_{c_i}}g_j^{1-\eta_{c_i}}\prod_{i=y+1}^{n_{T_j}}[1-(1-s_j)^{\eta_{c_i}}g_j^{1-\eta_{c_i}}]=$$

$$\sum_{c=1}^{C_{n_{T_j}}^{y}}\prod_{i=1}^{y}g_j\prod_{i=y+1}^{n_{T_j}}(1-g_j)(\text{其中}\prod\text{与样本 }i\text{ 无关})=$$

$$\sum_{c=1}^{C_{n_{T_j}}^{y}}g_j^y(1-g_j)^{n_{T_j}-y}(\text{其中}\sum\text{与样本选法 }c\text{ 无关})=$$

$$C_{n_{T_j}}^{y}g_j^y(1-g_j)^{n_{T_j}-y}. \tag{7}$$

通过(3)式和(5)式的样本筛选机制,统计量复杂的概率分布简化为如下二项分布(8)和(9):

$X\sim B(n_{T_{j1}},s_j)$,$Y\sim B(n_{T_{j1}},1-s_j)$,筛选机制为(3); (8)

$X\sim B(n_{T_j},1-g_j)$,$Y\sim B(n_{T_j},g_j)$,筛选机制为(5). (9)

采用细致的样本筛选策略所构建的概率分布能显著优化计算效率,提高统计量的稳定性.因该筛选机制选取具备特定属性的样本集,从而规避了对所有样本排列组合的遍历,无需逐一评估被试的掌握状态,加速了公式(1)、(2)的计算过程;从简化后的分布(8)、(9)可知,概率分布服从单参数的二项分布(仅含 $g$ 或 $s$),这种简化避免了多个估计参数偏差累加的风险,从而提高了统计量的稳定性.故本研究的样本筛选机制为假设检验的计算高效性和检验稳定性提供了可靠的支撑.

1.3.3 假设检验 本研究基于样本筛选机制和假设检验对 $Q$ 矩阵进行修正,以题目 $j$ 中的属性 $k$ 为例,其核心逻辑如下:当题目 $j$ 未考察属性 $k$ 时,理论上,那些掌握题目 $j$ 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题目,意味着错误出现的概率极低.如果出现了这种罕见情况,假设检验有充分理由判断题目 $j$ 实际上考察了属性 $k$;相反,若题目 $j$ 确实考察了属性 $k$,那么对于那些未掌握属性 $k$、但掌握了题目 $j$ 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,若观察到这类被试意外地频繁正确作答,假设检验有足够把握认为题目 $j$ 并未考察属性 $k$.

第一步:参数估计

由于 $q_{jk}$ 可能存在错误,因此为了降低 $q_j$ 向量错误所导致模型参数的估计误差,采用删除题目 $j$ 的数据.即 DINA 模型输入 $Q_{\setminus[j,:]}$ 与 $R_{\setminus[:,j]}$,输出 $J-1$ 道题目的猜测参数 $\tilde{g}$ 和失误参数 $\tilde{s}$,以及每个被试的掌握情况 $\tilde{\beta}$.

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:

$$H_0:q_{jk}=0\leftrightarrow H_1:q_{jk}=1. \tag{10}$$

根据题目 $j$ 的 $q_j$ 向量,以及被试掌握情况 $\tilde{\beta}$,使用公式(3)中选择方法构建样本 $T_j$,根据样本 $T_j$ 构建统计量,计算作答题目 $j$ 的错误数量 $X=n_{T_j}-\sum_{i\in T_j}r_{ij}$.

当原假设成立时($q_{jk}=0$),样本 $T_j$ 中被试的掌握模式 $\tilde{\beta}_i$,$i\in T_j$ 均可答对考察模式为 $q_j$ 的题目,则 $T_j$ 中被试对于题目 $j$ 的错误可以推断为非猜测性失误,其中错误概率可由 DINA 模型估计的失误参数定义.但需要注意的是,由于模型训练过程中排除了题目 $j$ 的数据,即输入数据为 $Q_{\setminus[j,:]}$ 与 $R_{\setminus[:,j]}$,导致题目 $j$ 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

均值来估计这一概率：$\bar{s}=\frac{1}{J-1}\sum_{l=1}^{J-1}\tilde{s}_l$. 因此，相应的统计量 $X$ 服从概率为 $\bar{s}$ 的二项分布：

$$X\sim B(n_{T_j},\bar{s}),P(X=x)=C_{n_{T_j}}^{x}\bar{s}^{x}(1-\bar{s})^{n_{T_j}-x}. \tag{11}$$

当原假设不成立时($q_{jk}=1$)，意味着题目 $j$ 考察了属性 $k$，然而样本 $T_j$ 中的所有被试均未掌握属性 $k$，即被试的掌握状态 $\eta_{ik}=0,i=1,2,\cdots T_j$. 因此 $T_j$ 中被试应以较大概率答错题目 $j$，此时统计量 $X$ 有偏大的趋势，则拒绝域形式为 $[c,+\infty]$，具体表述如下：

$$W_1=\{(r_{1j},r_{2j},\cdots,r_{n_{T_j}j}):X\geqslant c\}=\left\{n_{T_j}-\sum_{i\in T_j}r_{ij}\geqslant c\right\}. \tag{12}$$

最后，给定置信度 $\alpha$，得否定域为 $\{X\geqslant b_\alpha(n_{T_j},\bar{s})\}$ 水平为 $\alpha$ 的缺失属性二项分布检验. 若错误数量为 $x$ 且 $x\geqslant b_\alpha(n_{T_j},\bar{s})$，记 $P_{0\to1}=1-P(X\geqslant x)$，则有 $P_{0\to1}\geqslant1-\alpha$，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：

$$H_0:q_{jk}=1\leftrightarrow H_1:q_{jk}=0, \tag{13}$$

同理使用公式(5)的方法选择掌握模式集合 $T_j$，根据样本 $T_j$ 构建统计量，计算作答题目 $j$ 的正确数量 $Y=\sum_{i\in T_j}r_{ij}$.

原假设成立时($q_{jk}=1$)，样本 $T_j$ 中的被试均未掌握属性 $k$，因此他们无法正确作答题目 $j$，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 模型估计其它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：

$$Y\sim B(n_{T_j},\bar{g}),P(Y=y)=C_{n_{T_j}}^{y}\bar{g}^{y}(1-\bar{g})^{n_{T_j}-y}. \tag{14}$$

当原假设不成立时($q_{jk}=0$)，即题目 $j$ 未考察了属性 $k$，这意味着样本 $T_j$ 中的被试都应该正确回答题目 $j$，导致统计量 $Y$ 有偏大的趋势，拒绝域形式为 $[c,+\infty]$，具体为：

$$W_2=\{(r_{1j},r_{2j},\cdots,r_{n_{T_j}j}):Y\geqslant c\}=\left\{\sum\nolimits_{i\in T_j}r_{ij}\geqslant c\right\}. \tag{15}$$

最后，给定置信度 $\alpha$，得否定域为 $\{Y\geqslant b_\alpha(n_{T_j},\bar{g})\}$ 水平为 $\alpha$ 的冗余属性二项分布检验. 若正确数量为 $y$ 且 $y\geqslant b_\alpha(n_{T_j},\bar{g})$，记 $P_{1\to0}=1-P(Y\geqslant y)$，则有 $P_{1\to0}\geqslant1-\alpha$，拒绝原假设.

### 1.4　SHT 法实现步骤

本研究在 DINA 模型参数 $g,s$ 以及掌握状态 $\beta$ 的基础上，结合 1.3.2 节的样本选择机制和 1.3.3 节假设检验理论，提出了 $Q$ 矩阵的修正方法——基于样本筛选机制的假设检验法(SHT). 该方法基于两种基本的假设检验情形，对 $Q$ 矩阵元素逐题修正，并进行逐题验证，完整的 $Q$ 矩阵 SHT 法修正算法步骤见表 1 的伪代码 1.

算法具体分为三个模块：1) 认知诊断参数估计模块. 2) 属性修正模块. 3) 验证模块. 认知诊断模块用于估计参数，例如伪代码 2 ～ 4 行对 $g,s,\beta$ 分别估计；属性修正模块是对题目 $j$ 所有属性分别修正，如伪代码 7 ～ 20 行，其中，检验的属性可能同时拒绝缺失假设和冗余假设，因此需要取“更有可能”的情况；最后，鉴于 $q$ 向量可能修正为 0 向量，但题目至少考察一个属性，因此需要验证模块，取 $K$ 个属性中取缺失概率最大的属性进行考察.

## 2　SHT 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 SHT 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($4\times4\times3$)，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[20] 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. 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 $2^5-1=31$ 种所有可能的题目考核模式.

### 2.1　Monte 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、被试掌握模式的模拟以及作答记录的模拟，模拟方法参考 Leighton[20]、涂东波[12]、Nájera 等[21] 介绍的模拟方法.

1) 真值 $Q$ 矩阵生成

为了适应 $Q$ 矩阵的多样性变化，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. 题目总数设置参考 Nájera 等[21]，设定题目属性比为 8.

2) 错误 $Q$ 矩阵模拟

参考涂东波[12]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,但对于错误题目与错误属性的选择是完全随机的.

3) 被试掌握模式

被试掌握模式与 $Q$ 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1 000人,不同之处在于被试掌握模式有32种.

4) 被试作答矩阵

采用Leighton[20]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.

表1 基于样本筛选机制的假设检验算法伪代码1

Tab.1 Hypothesis testing algorithm base on sample selection pseudocode 1

基于样本筛选机制的假设检验算法(SHT)

输入:题目数 $J$,属性数 $K$,被试数 $I$,$Q=(q_{jk})$,作答矩阵 $R=(r_{ij})$,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

$Q_{\backslash[j,:]} \leftarrow Q\backslash Q_{[j;1,\cdots,K]}$; $R_{\backslash[:,j]} \leftarrow R\backslash R_{[1,\cdots,I;j]}$; \\ 输入数据去除题目 $j$

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

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

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

For $k=1$ to $K$ do \\ 遍历每个属性

根据公式(3)和(5)筛选样本分别为 $T_1$、$T_2$ \\ 样本筛选

$n_X=n_{T_1}-\sum_{i\in T_1}r_{ij}$; $n_Y=\sum_{i\in T_2}r_{ij}$ \\ 计算错误作答数、正确作答数

$p_{0\to1} \leftarrow P(X<n_X), p_{1\to0} \leftarrow P(Y<n_Y)$ \\ $X\sim B(n_{T_1},\bar{s}), Y\sim B(n_{T_2},\bar{g})$

If $p_{0\to1}\geqslant 1-\alpha$ and $p_{1\to0}\geqslant 1-\alpha$ then \\ 同时拒绝缺失假设和冗余假设

If $p_{0\to1}\geqslant p_{1\to0}$ then

$\tilde{q}_{jk} \leftarrow 1$ \\ 缺失的概率大于冗余的概率,判断为属性缺失

Else

$\tilde{q}_{jk} \leftarrow 0$ \\ 冗余的概率大于缺失的概率,判断为属性冗余

End If

Elif $p_{0\to1}\geqslant 1-\alpha$ then \\ 只拒绝一个假设

$\tilde{q}_{jk} \leftarrow 1$ \\ 缺失概率大于 $1-\alpha$,判断为属性缺失

Elif $p_{1\to0}\geqslant 1-\alpha$ then

$\tilde{q}_{jk} \leftarrow 0$ \\ 冗余概率大于 $1-\alpha$,判断为属性冗余

Else

Continue \\ 均不通过检验则不做修正

End If

End For

If $\|\tilde{q}_j\|=0$ then \\ 验证模块

$\tilde{q}_{jk^*} \leftarrow 1$,其中 $k^*=\arg\max p_{0\to1}^k$ \\ 出现0向量,则 $p$ 值最小的位置取1

End If

$\tilde{Q} \leftarrow \tilde{q}_j, j=1,2,\cdots,J$

End For

## 2.2 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SHT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 $PMR$、$AMR$、$TAR$ 及 $FAR$.

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

$$R_{PM}=\sum_{j=1}^{J}I(\tilde{q}_{jt}=q_{jt})/J. \tag{16}$$

2) 属性判准率(attribute match ratio,AMR):该指标从属性粒度评估修正方法的准确

性，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，其表达式：

$$R_{AM}=\sum_{j=1}^{J}\sum_{k=1}^{K}I(\tilde{q}_{jk}=q_{jk})/JK. \tag{17}$$

3）正确属性保留率（true attribute ratio，TAR）：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. $TAR$ 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：

$$R_{TA}=\frac{\sum_{j=1}^{J}\sum_{k=1}^{K}I(\tilde{q}_{jk}=q_{jk}\mid q_{jk}^{w}=q_{jk})}{\sum_{j=1}^{J}\sum_{k=1}^{K}I(q_{jk}^{w}=q_{jk})}, \tag{18}$$

其中，$q_{jk}$，$q_{jk}^{w}$，$\tilde{q}_{jk}$ 分别为真值 $Q$，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（false attribute ratio，FAR）：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

$$R_{FA}=\frac{\sum_{j=1}^{J}\sum_{k=1}^{K}I(\tilde{q}_{jk}=q_{jk}\mid q_{jk}^{w}\neq q_{jk})}{\sum_{j=1}^{J}\sum_{k=1}^{K}I(q_{jk}^{w}\neq q_{jk})}, \tag{19}$$

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.

### 2.3　试验结果

表 2～5 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，$Q^{w}R_{PM}$、$Q^{w}R_{AM}$ 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，$\tilde{Q}R_{PM}$、$\tilde{Q}R_{AM}$ 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. $R_{TA}$ 与 $R_{FA}$ 为修正后 $\tilde{Q}$ 矩阵的正确属性保留率和错误属性修正率.

表 2　SHT 法在作答失误为 5% 的情况下的 100 次试验平均结果

Tab. 2　Average results of 100 trials using the SHT method with a 5% error rate in response

| $Q$ 矩阵失误率 | 置信度 $\alpha$ | $Q^{w}R_{PM}$ | $\tilde{Q}R_{PM}$ | 提高率 | $Q^{w}R_{AM}$ | $\tilde{Q}R_{AM}$ | 提高率 | $R_{TA}$ | $R_{FA}$ |
|---|---|---|---|---|---|---|---|---|---|
| | 0.01 | 0.794 | 1.000 | 0.207 | 0.955 | 1.000 | 0.045 | 1.000 | 1.000 |
| 5% | 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 3　SHT 法在作答失误为 10% 的情况下的 100 次试验平均结果

Tab. 3　Average results of 100 trials using the SHT method with a 10% error rate in responses

| $Q$ 矩阵失误率 | 置信度 $\alpha$ | $Q^{w}R_{PM}$ | $\tilde{Q}R_{PM}$ | 提高率 | $Q^{w}R_{AM}$ | $\tilde{Q}R_{AM}$ | 提高率 | $R_{TA}$ | $R_{FA}$ |
|---|---|---|---|---|---|---|---|---|---|
| | 0.01 | 0.794 | 1.000 | 0.207 | 0.955 | 1.000 | 0.045 | 1.000 | 1.000 |
| 5% | 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 | 0.998 | 0.393 | 0.905 | 1.000 | 0.095 | 1.000 | 1.000 |
| 10% | 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 |
| | 0.01 | 0.459 | 0.991 | 0.532 | 0.855 | 0.998 | 0.143 | 0.998 | 0.998 |
| 15% | 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 |
| | 0.01 | 0.347 | 0.976 | 0.629 | 0.804 | 0.995 | 0.191 | 0.995 | 0.994 |
| 20% | 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4 SHT法在作答失误为15%的情况下的100次试验平均结果

Tab.4 Average results of 100 trials using the SHT method with a 15% error rate in responses

| Q矩阵失误率 | 置信度 $\alpha$ | $Q^{\omega}R_{PM}$ | $\widetilde{Q}R_{PM}$ | 提高率 | $Q^{\omega}R_{AM}$ | $\widetilde{Q}R_{AM}$ | 提高率 | $R_{TA}$ | $R_{FA}$ |
|---|---|---|---|---|---|---|---|---|---|
| | 0.01 | 0.794 | 1.000 | 0.206 | 0.955 | 1.000 | 0.045 | 1.000 | 1.000 |
| 5% | 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 |
| | 0.01 | 0.605 | 0.995 | 0.390 | 0.905 | 0.999 | 0.094 | 0.999 | 0.999 |
| 10% | 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 |
| | 0.01 | 0.459 | 0.983 | 0.525 | 0.855 | 0.997 | 0.142 | 0.997 | 0.997 |
| 15% | 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 |
| | 0.01 | 0.347 | 0.950 | 0.603 | 0.804 | 0.990 | 0.186 | 0.990 | 0.990 |
| 20% | 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5 SHT法在作答失误为20%的情况下的100次试验平均结果

Tab.5 Average results of 100 trials using the SHT method with a 20% error rate in responses

| Q矩阵失误率 | 置信度 $\alpha$ | $Q^{\omega}R_{PM}$ | $\widetilde{Q}R_{PM}$ | 提高率 | $Q^{\omega}R_{AM}$ | $\widetilde{Q}R_{AM}$ | 提高率 | $R_{TA}$ | $R_{FA}$ |
|---|---|---|---|---|---|---|---|---|---|
| | 0.01 | 0.794 | 0.995 | 0.202 | 0.955 | 0.999 | 0.044 | 0.999 | 0.999 |
| 5% | 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 |
| | 0.01 | 0.605 | 0.980 | 0.375 | 0.905 | 0.996 | 0.091 | 0.996 | 0.996 |
| 10% | 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 |
| | 0.01 | 0.459 | 0.942 | 0.484 | 0.855 | 0.988 | 0.133 | 0.989 | 0.986 |
| 15% | 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 |
| | 0.01 | 0.347 | 0.880 | 0.533 | 0.804 | 0.975 | 0.172 | 0.976 | 0.973 |
| 20% | 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 |

分析表2～5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SHT法均能有效修正$Q^{\omega}$,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,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3和表4),SHT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.

## 3 研究二:SHT法与其他算法的比较研究

初步分析SHT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SHT法的可行性与准确性,实验结果见表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,直观地展示了在样本少、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,SHT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

表 6　SHT 法与其他算法在不同参数下的 100 次试验平均结果

Tab. 6　SHT method vs. other algorithms: average outcomes of 100 trials under various parameters

| $Q$ 矩阵错误率 | 作答错误率 | 算法＼样本量 | $R_{PM}$（模式准确率） | | | | $R_{AM}$（属性准确率） | | | | $R_{TA}$（正确属性保留率） | | | | $R_{FA}$（错误属性修正率） | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 100 | 300 | 1 000 | 2 000 | 100 | 300 | 1 000 | 2 000 | 100 | 300 | 1 000 | 2 000 | 100 | 300 | 1 000 | 2 000 |
| 5% | 10% | $Q^{w}$ | 0.800 | 0.775 | 0.792 | 0.800 | 0.955 | 0.955 | 0.955 | 0.955 | | | | | | | | |
| | | SHT | **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 | | | | | | | | |
| | | SHT | **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 | | | | | | | | |
| | | SHT | **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 | | | | | | | | |
| | | SHT | **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 | | | | | | | | |
| | | SHT | **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 | | | | | | | | |
| | | SHT | **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 |
| 15% | 10% | $Q^{w}$ | 0.426 | 0.499 | 0.467 | 0.401 | 0.855 | 0.855 | 0.855 | 0.855 | | | | | | | | |
| | | SHT | **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 | | | | | | | | |
| | | SHT | **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 | | | | | | | | |
| | | SHT | **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 | | | | | | | | |
| | | SHT | **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 | | | | | | | | |
| | | SHT | **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 | | | | | | | | |
| | | SHT | **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 |

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,SHT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"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2 000,$Q$矩阵错误率为20%时),$\delta$法表现更优,然而,SHT法基于深厚的理论支撑和广泛的适用性,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,SHT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.其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的误判风险.

## 4 实证数据研究

为了验证SHT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,选取两个典型数据集:Tatsuoka分数减法数据[22]与TIMSS2007数据,分别从修正方法拟合度、题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和表8所示.

### 4.1 数据集

Tatsuoka数据简称FraSub,其涵盖了536名被试的测试表现,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[23]及de la Torre[10]等诸多学者深入测量与解析.而TIMSS2007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Wang等[24]、Lee等[25]、Park等[26]研究用于探索教育评估和认知诊断领域.

表7 FraSub数据集$Q$矩阵

Tab.7 $Q$ matrix of FraSub dataset

| Item | Attribute | | | | |
|---|---|---|---|---|---|
| | $A1$ | $A2$ | $A3$ | $A4$ | $A5$ |
| $T01$ | 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.

### 4.2 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(Akaike information criterion,AIC)和贝叶斯信息准则(Bayesian information criterion,BIC);第二类绝对拟合指标包括$M_2$检验、近似均方根误差(root mean square error of approximation,RMSEA)、标准均方根残差(standardized root mean squared residual,SRMSR)[27-29].

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[12].

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,SHT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 8　TIMSS2007 数据集 Q 矩阵

Tab. 8　Q matrix of TIMSS2007 dataset

| 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：属性编号的前缀 N，Number；G，Geometric Shapes & Measures；D，Data Display. 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；LM，Location and Movement；RI，Reading and Interpreting；OR，Organizing and Representing.

表 9　基于三种算法修正后 Q 矩阵的拟合指标(FraSub)

Tab. 9　Fitting metrics of Q matrix adjusted by three algorithms (FraSub)

| Q 矩阵 | 相对拟合指标 | | | 绝对拟合指标 | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | $M_2$ | | | *RMSEA* | *SRMSR* |
| | $-2LL$ | *AIC* | *BIC* | $M_2$ | $df$ | $p$ | | |
| 修正前 Q | 6 911.59 | 7 033.59 | 7 294.93 | 231.750 | 59 | <.001 | 0.070 | 0.112 |
| SHT 法 | **6 853.31** | **6 975.31** | **7 236.64** | **145.243** | 59 | <.001 | **0.052** | **0.076** |
| δ 法 | 6 890.29 | 7 012.29 | 7 273.62 | 178.978 | 59 | <.001 | 0.062 | 0.101 |
| γ 法 | 6 971.43 | 7 093.43 | 7 354.76 | 231.910 | 59 | <.001 | 0.074 | 0.091 |

表10 基于三种算法修正后Q矩阵的拟合指标(TIMSS2007)

Tab.10 Fitting metrics of Q matrix adjusted by three algorithms (TIMSS2007)

| Q矩阵 | 相对拟合指标 | | | 绝对拟合指标 |
|---|---|---|---|---|
| | $-2LL$ | $AIC$ | $BIC$ | $SRMSR$ |
| 修正前Q | 7 588.83 | 11 726.83 | 21 137.09 | 0.031 6 |
| SHT法 | **7 559.87** | **11 697.87** | **21 108.14** | **0.029 1** |
| $\delta$法 | 7 657.49 | 11 795.49 | 21 205.76 | 0.047 0 |
| $\gamma$法 | 7 603.33 | 11 741.33 | 21 151.59 | 0.033 0 |

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.

### 4.3 各修正算法题目拟合度评估

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;SHT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,SHT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SHT法有更优的拟合性能.

## 5 结论与讨论

### 5.1 结论

本研究提出了基于DINA模型的Q矩阵修正方法——SHT法,采用蒙特卡洛模拟与其他同类方法进行比较,验证SHT法的可行性和准确性,得到如下三个结论.

1) SHT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,SHT法在小样本环境中展现出更高的稳健性和优越性能、尤其在高作答失误率时,其优势更为显著;得益于统计理论支撑,在大样本环境中,SHT法对作答失误率的鲁棒性显著提高,其在保持强劲的竞争力同时,修正过程比$\delta$法更简单.

3) 在实证数据中,SHT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,展现出更显著的优势.

### 5.2 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.

表11 基于三种算法逐题修正后Q矩阵的拟合指标(TIMSS2007)

Tab.11 Fitting metrics of Q matrix question-wise adjusted by three algorithms(TIMSS2007)

| 题目 | Q矩阵 | 相对拟合指标 | | |
|---|---|---|---|---|
| | | $-2LL$ | $AIC$ | $BIC$ |
| 1 | $Q$ | 7 588.83 | 11 726.83 | 21 137.09 |
| | $h$ | 0.53 | 0.53 | 0.53 |
| | $\delta$ | 4.87 | 4.87 | 4.87 |
| | $\gamma$ | 0.00 | 0.00 | 0.00 |
| 2 | $Q$ | 7 588.83 | 11 726.83 | 21 137.09 |
| | $h$ | 0.00 | 0.00 | 0.00 |
| | $\delta$ | −0.87 | −0.87 | −0.86 |
| | $\gamma$ | 0.00 | 0.00 | 0.00 |
| 3 | $Q$ | 7 588.83 | 11 726.83 | 21 137.09 |
| | $h$ | 2.49 | 2.49 | 2.49 |
| | $\delta$ | 2.88 | 2.88 | 2.89 |
| | $\gamma$ | 3.18 | 3.18 | 3.18 |
| 4 | $Q$ | 7 588.83 | 11 726.83 | 21 137.09 |
| | $h$ | −21.99 | −21.99 | −21.99 |
| | $\delta$ | 19.31 | 19.31 | 19.31 |
| | $\gamma$ | 0.00 | 0.00 | 0.00 |
| 5 | $Q$ | 7 588.83 | 11 726.83 | 21 137.09 |
| | $h$ | 0.00 | 0.00 | 0.00 |
| | $\delta$ | −32.73 | −32.73 | −32.72 |
| | $\gamma$ | 0.00 | 0.00 | 0.00 |
| 6 | $Q$ | 7 588.83 | 11 726.83 | 21 137.09 |
| | $h$ | −15.23 | −15.23 | −15.22 |
| | $\delta$ | −18.88 | −18.88 | −18.87 |
| | $\gamma$ | 0.00 | 0.00 | 0.00 |
| 7 | $Q$ | 7 588.83 | 11 726.83 | 21 137.09 |
| | $h$ | −0.87 | −0.87 | −0.87 |
| | $\delta$ | −0.79 | −0.79 | −0.78 |
| | $\gamma$ | −1.22 | −1.22 | −1.21 |
| 9 | $Q$ | 7 588.83 | 11 726.83 | 21 137.09 |
| | $h$ | 0.00 | 0.00 | 0.00 |
| | $\delta$ | 1.83 | 1.83 | 1.84 |
| | $\gamma$ | 0.00 | 0.00 | 0.00 |
| 10 | $Q$ | 7 588.83 | 11 726.83 | 21 137.09 |
| | $h$ | 9.96 | 9.96 | 9.96 |
| | $\delta$ | 7.01 | 7.01 | 7.02 |
| | $\gamma$ | 13.71 | 13.71 | 13.71 |

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.

2）本研究为了避免较大的估计偏差，取其他题目的参数 $g$，$s$ 平均作为题目 $j$ 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.

## 参考文献：


[1] 教育部. 教育部 2022 年工作要点[EB/OL].(2022-02-08)[2023-05-11]. http://www.moe.gov.cn/jyb_sjzl/moe_164/202202/t20220208_597666.html.
Ministry of Education. Key work points of the Ministry of Education in 2022[EB/OL].(2022-02-08)[2023-05-11]. http://www.moe.gov.cn/jyb_sjzl/moe_164/202202/t20220208_597666.html.(Ch).

[2] 国务院关于深化教育改革全面提高义务教育质量的意见[EB/OL].(2019-06-23)[2023-05-11]. https://www.gov.cn/zhengce/2019-07/08/content_5407361.htm.
Opinions of the State Council on deepening education reform and comprehensively improving the quality of compulsory education[EB/OL].(2019-06-23)[2023-05-11]. https://www.gov.cn/zhengce/2019-07/08/content_5407361.htm.(Ch).

[3] 教育部，中央网信办，国家发展改革委，工业和信息化部，财政部，中国人民银行. 教育部等六部门关于推进教育新型基础设施建设构建高质量教育支撑体系的指导意见[J]. 中华人民共和国教育部公报，2021(9)：15-19.
Ministry of Education, Central Cyberspace Office, National Development and Reform Commission, Ministry of Industry and Information Technology, Ministry of Finance, People's Bank of China. Guiding opinions of the Ministry of Education and six other departments on promoting the construction of new education infrastructure and building a high quality education support system [J]. Bulletin of the Ministry of Education of the People's Republic of China, 2021(9): 15-19.(Ch).

[4] 杨华利，耿晶，胡盛泽，等. 人工智能时代的教育测评通用理论框架与实践进路[J]. 中国远程教育，2022(12)：68-77.
YANG H L, GENG J, HU S Z, et al. Educational assessment in the era of artificial intelligence: towards a universal theoretical framework and practice pathways[J]. Chinese Journal of Distance Education, 2022(12): 68-77.(Ch).

[5] 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.

[6] 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.

[7] DE LA TORRE J. DINA model and parameter estimation: a didactic[J]. Journal of Educational and Behavioral Statistics, 2009, 34(1): 115-130.

[8] TATSUOKA K K. Cognitive assessment: an introduction to the rule space method[M]. Oxfordshire: Taylor and Francis, 2009.

[9] 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.

[10] 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.

[11] de la TORRE J, CHIU C Y. A general method of empirical $Q$-matrix validation[J]. Psychometrika, 2016, 81: 253-273.

[12] 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.
TU D B, CAI Y, DAI H Q. A new method of $Q$-matrix validation based on DINA model[J]. Acta Psychologica Sinica, 2012, 44(4): 558-568.(Ch).

[13] 喻晓锋，罗照盛，高椿雷，等. 使用似然比 $D^2$ 统计量的题目属性定义方法[J]. 心理学报，2015，47(3)：417-426.
YU X F, LUO Z S, GAO C L, et al. An item attribute specification method based on the likelihood $D^2$ statistic[J]. Acta Psychologica Sinica, 2015, 47(3): 417-426.(Ch).

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

[15] 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，2018，41(2)：466-474.
WANG DX, GAO XL, CAI Y, et al. A new $Q$-matrix estimation method: ICC based on ideal response[J]. Journal of Psychological Science, 2018, 41(2): 466-474.(Ch).

[16] 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，2019，42(4)：988-996.
WANG D X, GAO X L, CAI Y, et al. A new general method for $Q$-matrix validation in cognitive diagnosis assessments[J]. Journal of Psychological Science, 2019, 42(4): 988-966.(Ch).

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

[18] 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)：180-188.
WANG D X, GAO X L, HAN Y T, et al. A simple and effective $Q$-matrix estimation method: from non-parametric perspective[J]. Applied Psychological Measurement, 2018, 41(1): 180-188.(Ch).

[19] 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.

[20] 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.

[21] 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.

[22] TATSUOKA K K. Toward an integration of item response theory and cognitive analysis[J]. Diagnostic Monitoring of Skill and Knowledge Acquisition, 1990: 543-588.

[23] TATSUOKA K K. Analysis of errors in fraction addition and subtraction problems. Final report [R]. 252 ERL, 103 S. Mathews St. Univ. of Illinois, Urbana, IL 61801,1984.

[24] WANG X Q, DING S L, LUO F. *Q* matrix and its applications in cognitive diagnosis [J]. Journal of Psychological Science, 2019 (3): 739-746.

[25] 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.

[26] PARK Y S, LEE Y S, XING K. Investigating the impact of item parameter drift for item response theory models with mixture distributions[J/OL]. Frontiers in Psychology, 2016, 7[2016-02-01]. https://doi.org/10.3389/fpsyg.2016.00255.

[27] 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.

[28] 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.

[29] 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.

# Optimization of the *Q*-matrix in cognitive diagnostic modeling base on DINA: a new approach combining sample selection and hypothesis testing

LI Bo, HU Yuqian, ZHANG Yong, TIAN Yi

(School of Mathematics and Statistics, Central China Normal University, Wuhan 430079, China)

**Abstract**: In this paper, an SHT method, a *Q*-matrix revision strategy based on the DINA model is proposed innovatively. The study employs Monte Carlo simulation techniques to conduct an in-depth comparison with existing similar approaches, comprehensively evaluating its feasibility and accuracy. This method has the following three advantages. 1) Efficient revision and robustness validation: the SHT method demonstrates remarkable revision efficacy under varying levels of error rates, significantly enhancing the precision of the *Q* matrix, thereby validating its robustness. 2) Dual benefits in small and large sample scenarios: compared with domestic and international counterparts, the SHT method particularly excels in small sample scenarios, with its robustness and performance advantages becoming more pronounced when confronted with high error rates. 3) Distinct advantages with complex datasets: in empirical data analysis, the SHT method not only enhances the fitting capacity of cognitive diagnostic models but also exhibits more conspicuous advantages over other algorithms when dealing with datasets of high dimensionality complexity and relatively limited samples.

**Key words**: cognitive diagnosis; *Q*-matrix validation; DINA model; hypothesis testing; the $\delta$ method; the $\gamma$ method