基于类别水平的多级计分认知诊断 Q 矩阵修正：相对拟合统计量视角*

汪大勋[1] 高旭亮[2] 蔡艳[1] 涂冬波[1]

（[1]江西师范大学心理学院，南昌，330022）（[2]贵州师范大学心理学院，贵阳，550000）

摘　要 多级计分认知诊断模型的开发对认知诊断的发展具有重要作用，但对于多级计分模型下的 Q 矩阵修正还有待研究。本研究尝试对多级计分认知诊断 Q 矩阵修正进行研究，并聚焦更具诊断价值的基于项目类别水平的 Q 矩阵修正。将相对拟合统计量应用于多级计分认知诊断 Q 矩阵修正，并与已有方法 stepwise 方法（Ma & de la Torre, 2019）进行比较。研究表明：BIC 方法对多级计分认知诊断模型的 Q 矩阵修正具有较高的模式判准率和属性判准率，其对 Q 矩阵的恢复率也高于 stepwise 方法，BIC 方法修正后的 Q 矩阵与数据更加拟合；在复杂模型中，相对拟合指标 BIC 比 AIC 和-2LL 表现更好，在实践中，使用者可以选择 BIC 法进行测验 Q 矩阵修正；Q 矩阵修正效果受到被试人数的影响，增加被试人数可以提高 Q 矩阵修正的正确率。总之，本研究为多级计分认知诊断 Q 矩阵修正提供了重要的方法支持。

关键词 认知诊断；Q 矩阵；seq-GDINA；BIC

1 引言

传统心理与教育测验通过对学生的能力进行评估和排序，来评价学生的学习效果或者进行选拔，而对分数背后的心理加工过程和认知技能无法提供详细的信息。随着测评技术的发展，人们希望测验能够提供更详细的诊断信息，从而进行针对性的补救或因材施教。认知诊断作为认知心理学和心理测量学的结合，可以实现对人的内部心理加工过程和认知技能的诊断，从而为针对性地补救和教师因材施教提供依据(Chang, 2015; Chen，2017；张华华，汪文义, 2016)。为此，研究者们开发了许多具有认知诊断功能的计量模型（Cognitive Diagnosis Models, CDMs），常见的有 DINA（Haertel, 1984;）、NIDA（Maris, 1999;）、DINO（Templin

收稿日期：2018-12-14
* 国家自然科学基金（31660278,31760288,31960186）、江西省教育厅研究生创新基金（YC2018-B025）和江西师范大学研究生境内外访学项目的资助。
高旭亮为共同第一作者。
通讯作者：涂冬波，Email：tudongbo@aliyun.com

& Henson, 2006）、R-RUM（Hartz & Roussos, 2008）、A-CDM 和 G-DINA（de la Torre, 2011）等，这些模型均适用于 0-1 计分的测验情境。为了适应多级评分的测验情景，研究者们也开发出了能用于多级计分的认知诊断模型，如多级计分的 GDM (von Davier, 2008)，P- DINA 模型（涂冬波，蔡艳，戴海琦，丁树良，2010）、多级的 LCDM（Hansen, 2013）、seq-GDINA（Ma & de la Torre, 2016）等。与其他的多级计分模型不同，seq-GDINA 模型可以从项目水平和得分类别水平分别定义 Q 矩阵，Ma 和 de la Torre（2016）将基于项目水平的 Q 阵定义为非约束化的 Q 阵（Unrestricted Q），而基于类别水平的 Q 阵定义为约束化的 Q 阵（Restricted Q）。如对于算式 $\sqrt{45/3-6}$ 的求解分为三个步骤：即 step 1 计算 45/3=15，step 2 计算 15-6=9，step 3 计算 $\sqrt{9}=3$。三个步骤分别考察了三个属性 A1(除法)、A2(减法)、A3(根号运算)。项目水平的 Q 矩阵将该题测量 Q 阵定义为 $q_j=[111]$，即该题测量了三个属性；而基于得分类别的 Q 矩阵则需要对该题的每一步骤（或每一个得分类别）进行 Q 矩阵标定，将该题基于类别水平的 Q 矩阵定义为 $q_j=\begin{bmatrix}100\\010\\001\end{bmatrix}$，即 step 1 测量了 A1，step 2 测量了 A2，step 3 测量了 A3。相比较而言，基于得分类别的 Q 矩阵可以更加深入地探测出学生的解题过程，从而能提高被试分类的准确性（Ma & de la Torre, 2016）。关于项目水平 Q 矩阵和类别水平 Q 矩阵的详细介绍可以参考 Ma 和 de la Torre (2016)的文献。总体来讲，基于得分类别的 Q 矩阵能更准确地反应出被试每个步骤需要的属性。而在实际中得分类别的 Q 矩阵的标定比项目水平 Q 矩阵的标定更为复杂，为每个类别标定 Q 矩阵也增加了专家的工作负担。在 Q 矩阵的修正上，对得分类别的 Q 矩阵修正也比项目水平 Q 矩阵修正的难度更大，因为对得分类别 Q 矩阵的修正需要考察所有题目每个类别的 Q 矩阵，而项目水平的 Q 矩阵则只需考虑整个项目整体测量了哪些属性，因此相对容易。

在对 Q 矩阵的修正上，国内外研究者提出许多种方法。其中一部分方法只适用于简化的认知诊断模型（如 DINA 和 DINO 模型），如：γ法（涂冬波，蔡艳，戴海崎，2012）、海明距离的方法（汪大勋, 高旭亮, 韩雨婷, 涂冬波，2018）、ICC-IR 法（汪大勋，高旭亮，蔡艳，涂冬波，2018）、δ法（de la Torre，2008）、*RSS* 法（Chiu，2013）。除此之外，研究者还提出一些适用于饱和认知诊断模型的 Q 矩阵修正方法，如 GDI 法(de la Torre & Chiu, 2016)、基于似然的方法（Xu & Shang，2018）和基于残差的方法 (Chen, 2017)。相比较而言，后三种方法的适用范围更广，既适用于饱和的认知诊断模型也适用于简化的认知诊断模

型。而三种可以适用于饱和模型的方法中，GDI 方法计算相对复杂，且需要设定一个截断值（PVAF=0.95）。并且通过预研究发现，该方法受样本量影响较大，在小样本量（N=500）中该方法表现不理想。而基于残差的方法虽然可以在测验层面考察测验属性是否多余或缺失，但该方法对题目层面的属性多余不够敏感(Chen, 2017)。而 Xu 和 Shang（2018）的方法采用 TLP (truncated L1 penalty function)的正则化算法，由估计的项目参数稀疏矩阵来推断题目的 q 向量，并结合信息指标（BIC）来进行 Q 矩阵估计或修正，他们的研究不仅进行了理论证明，Monte Carlo 实验也表明其方法具有较好的效果。此外，Chen，de la Torre 和 Zhang（2013）将-2LL、AIC 和 BIC 指标用于对不同 Q 矩阵的鉴别。研究发现在 DINA 模型中，-2LL 指标表现较好；而在饱和模型中，-2LL 倾向于选择在原有 Q 矩阵基础上增加属性的 Q 矩阵，而 BIC 指标的表现是最出色的。以上方法均是在 0-1 模型中的 Q 矩阵修正方法，而对于多级计分模型中的 Q 矩阵修正的研究，Ma 和 de la Torre (2019)提出了 GDI 和 wald 检验相结合的 stepwise 方法进行 seq-GDINA 模型的 Q 矩阵修正。该方法先选择单属性 q 向量中具有最大 GDI 值的 q 向量作为基础，再用 wald 检验是否显著来决定是否增加或删除属性，并通过计算 wald 检验以后 q 向量的 GDI 值来决定是否终止。该方法在确定每个类别的 q 向量时，需要进行多次的 wald 检验，并计算标准误，计算相对复杂。此外，该方法是从属性的角度来考察属性是否缺失或冗余，而对于 Q 矩阵修正后模型的整体拟合并没有考虑。

因此本研究尝试使用模型的相对拟合统计量对多级计分认知诊断模型 Q 矩阵进行修正，并聚焦更具诊断价值的基于项目类别水平的 Q 矩阵修正方法。具体来讲，本研究将模型拟合统计量中的-2LL、AIC（Akaike's Information Criterion）和 BIC（Bayesian Information Criteria）指标用于多级计分认知诊断模型的 Q 矩阵修正。本文采用的方法与 Xu 和 Shang (2018)方法有相似之处，都是需要对模型参数进行估计，并使用信息指标来进行 Q 矩阵修正。并且在修正 Q 矩阵时都是在其余题目 Q 矩阵保持不变的情况下，逐题（或类别）对 q 向量进行确定。而两种方法的区别在于，Xu 和 Shang （2018）的方法采用 TLP (truncated L1 penalty function)的正则化算法，由估计的项目参数稀疏矩阵来推断题目的 q 向量，并结合信息指标（BIC）来进行 Q 矩阵修正，因此并不需要对所有可能的 q 向量进行估计。而本文则在所有可能的 q 向量中通过拟合统计量挑选出最优的 q 向量。此外 Xu 和 Shang （2018）的方法是用于二级计分的 Q 矩阵估计或修正，而本文则是对多级计分 Q 矩阵修正进行研究。通过 Monte Carlo 模拟研究和实证数据分析来验证本文的方法并与 Ma 和 de la Torre (2019)提出的 stepwise 方法进行比较，从而为实际工作者在多级计分认知诊断中 Q 矩阵的修正与标定提供方法支持。

2 seq-GDINA 模型介绍

前已述及，在众多多级计分 CDMs 中，seq-GDINA 模型能从得分类别标定 Q 矩阵，从而更深入细致地探查被试的解题过程。此外，该模型使用 G-DINA 模型作为每个类别上的链接函数，在不同的假设条件下，seq-GDINA 模型可以转化为不同的多级计分模型（如 seq-DINA 和 seq-RRUM），因此该模型更加灵活。因此本研究采用 Ma 和 de la Torre（2016）开发的 seq-GDINA 模型进行研究。对该模型的介绍如下：

对于属性掌握模式为 α_c 的被试，其在项目 j 上 h 类别的正确作答概率为 $S_j(h|\alpha_c)$。那么，

$$S_j(h|\alpha_c)=\begin{cases}1, & \text{if } h=0\\ 0, & \text{if } h=H_j+1\end{cases} \tag{1}$$

其中 H_j 为题目 j 的类别总和，则该被试在项目 j 上得 h 分的概率为：

$$p\left(X_j=h|\alpha_c\right)=\left[1-S_j\left(h+1|\alpha_c\right)\right]\prod_{x=0}^{h}S_j\left(x|\alpha_c\right) \tag{2}$$

则对于每一种掌握模式的被试，其在项目 j 上得各种分数的概率之和为 1。

$$\sum_{h=0}^{H_j}P\left(X_j=h|\alpha_c\right)=1\ \forall c \tag{3}$$

被试在题目 j 上每个类别的得分概率受到题目 j 每个类别测量属性的影响。如在一个属性个数为 K 的测验中，设定 K_j^* 为题目 j 测量的属性个数，K_{jh}^* 为项目 j 类别 h 所测量的属性个数。$L=2^{K_{jh}^*}$ 为所有简化后的掌握模式，α_{ljh}^* 表示第 l 种掌握模式。对于掌握模式 α_{ljh}^*，其在项目 j 上 h 类别的概率函数表示为：

$$S_j\left(h|\alpha_{ljh}^*\right)=\phi_{jh0}+\sum_{k=1}^{K_{jh}^*}\phi_{jhk}\alpha_{lk}+\sum_{k'=k+1}^{K_{jh}^*}\sum_{k=1}^{K_{jh}^*-1}\phi_{jhkk'}\alpha_{lk}\alpha_{lk'}+\cdots+\phi_{jh12\cdots K_{jh}^*}\prod_{k=1}^{K_{jh}^*}\alpha_{lk} \tag{4}$$

其中 ϕ_{jh0} 是截距参数，ϕ_{jhk} 是属性 α_{lk} 的主效应，$\phi_{jhkk'}$ 是属性 α_{lk} 和 $\alpha_{lk'}$ 的交互效应，$\phi_{jh12\ldots k_j^*}$ 是所有属性的交互效应。与 GDINA 模型相似（de la Torre, 2011），在不同约束条件下 seq-GDINA 模型可以转换为 seq-DINA、seq-RRUM 等模型。

3 多级计分认知诊断 Q 矩阵修正

在认知诊断中定义题目 j 的 q 向量时，在其他题目 Q 矩阵不变的情况下，在所有可能的

q 向量中能使模型相对拟合更好的 q 向量应该为题目 j 的 q 向量。而认知诊断中常用的模型相对拟合指标包括-2LL、AIC 和 BIC。在约束模型（如 DINA）中，题目 q 向量标定错误会导致题目猜测参数和失误参数增加，从而降低模型的似然值，因此在约束模型中-2LL 指标可以挑选出恰当的 q 向量。而在复杂模型中，已有研究（Chen et al.，2013；Chen，2017）表明，若在原有 Q 矩阵基础上增加属性（overspecified）会产生更大的模型似然（由于模型参数个数增多）。所以在复杂模型中，-2LL 指标通常会挑选全为 1 的 q 向量($\boldsymbol{q}=[1111]$)作为题目 j 的 q 向量。因此在复杂模型中对题目 j 的 q 向量进行标定时需要对模型的参数个数进行惩罚，而 AIC 和 BIC 指标则是在-2LL 的基础上对模型的参数个数进行了惩罚。

3.1 -2LL 方法

在 0-1 计分模型 Q 矩阵修正中，如果一个测验包含了 J 个题目，K 个属性，用 Q_r（reduced Q-matrix）来表示所有可能的属性模式的集合，当属性之间没有关系时 Q_r 里包含了 2^K-1 种属性模式。使用-2LL 方法来进行 Q 矩阵标定时，是将最小-2LL 所对应的属性模式作为题目 j 的属性模式。即分别将 Q_r 里的属性模式作为题目 j 的属性模式（其余 J-1 个题目的 Q 矩阵不变），与其余 J-1 个题目一起进行参数估计，并计算-2LL。将最小-2LL 所对应的属性模式作为题目 j 的属性模式，公式表示为：

$$\hat{q}_j=\arg\min_{q_l\in Q_r}\ -2\ln(L(X|\hat{\beta},Q_{jl}^T)) \tag{5}$$

其中 $L(X|\hat{\beta},Q_{jl}^T)=\prod_{i=1}^{N}\prod_{j=1}^{J}\sum_{\alpha_c\in Q_s}L\left(X_{ij}\middle|\alpha_c,\hat{\beta}_j,q_j\right)\pi(\alpha_c)$，$L\left(X_{ij}\middle|\alpha_c,\hat{\beta}_j,q_j\right)$是当被试 i 的掌握模式为 α_c 时，题目 j 的似然，$\pi(\alpha_c)$是掌握模式为 α_c 的后验概率，Q_s 是所有可能掌握模式的集合。Q_{jl}^T 是当题目 j 的属性模式为第 l 种属性模式时，整个测验的 Q 矩阵。

与 0-1 计分情况下不同，在多级计分模型中，需要依次对每个题目的每个类别 q 向量进行验证与修正。同样可以将最小-2LL 所对应的属性模式作为题目 j 第 h 类别的属性模式。分别将 Q_r 里的属性模式作为题目 j 第 h 类别的属性模式（题目 j 的其余类别和其他 J-1 个题目的所有类别 Q 矩阵均保持不变），与其余题目一起进行参数估计，并计算-2LL。最小-2LL 所对应的属性模式作为题目 j 第 h 类别的属性模式，表示为：

$$\hat{q}_{jh}=\arg\min_{q_l\in Q_r}\ -2\ln(L(X|\hat{\beta},Q_{jhl}^T)) \tag{6}$$

其中 Q_{jhl}^T 是当题目 j 第 h 类别的属性模式为第 l 种属性模式 q_l 时，整个测验的 Q 矩阵。

使用-2LL 方法来修正 Q 矩阵时，需要对所有题目的所有类别依次循环验证，由此可进行 Q 矩阵修正。

3.2 AIC 方法

AIC（Akaike's Information Criterion）指标是由 Akaike（1974）开发的信息指标，是心理测量领域常用的测验相对拟合指标，可用于比较模型的相对拟合程度。与-2LL 方法相似，在使用 AIC 指标对题目 j 第 h 类别的 q 向量进行标定时，将 Q_r 里的所有属性模式依次作为题目 j 第 h 类别的属性模式，与其余题目一起进行参数估计，并计算模型的 AIC 指标，将 AIC 指标最小的属性模式作为题目 j 第 h 类别的属性模式。AIC 指标的计算公式如下：

$$\mathrm{AIC} = -2LL + 2d \tag{7}$$

其中 L 为模型的边际似然，其计算方法与-2LL 方法相同，d 为需要估计参数的个数。相对于-2LL 方法，AIC 指标考虑了参数个数的影响，参数个数多的属性模式将会受到惩罚。

3.3 BIC 方法

BIC（Bayesian Information Criteria）指标是由 Schwarz（1978）开发的信息指标，通常与 AIC 指标一起用于模型比较。与 AIC 指标相比，BIC 指标还考虑了样本量对模型拟合度的影响。BIC 指标的计算公式如下：

$$\mathrm{BIC} = -2LL + d \times ln(N) \tag{8}$$

其中 N 为被试样本量，L 和 d 分别为模型的边际似然和参数个数。使用 BIC 指标来标定题目 j 第 h 类别的属性模式的方法与 AIC 指标相同，最后选择具有最小 BIC 指标的属性模式作为题目 j 第 h 类别的属性模式。

3.4 **穷尽算法**（exhaustive search algorithm）**和顺序算法**（sequential search algorithm）

假设 $q^{(h)}$ 为 Q_r 中第 h 个 q 向量，$K^{(h)}$ 为 $q^{(h)}$ 的属性个数，$q^{(h')} \le q^{(h)}$ 表示 $q^{(h')}$ 嵌套于 $q^{(h)}$（de la Torre, 2011）。用 S_j 表示 q 向量的集合。

穷尽算法是分别将 Q_r 中的属性模式作为题目 j 第 h 类别的属性模式，计算出相对拟合指标后挑选出最优的属性模式，即集合 $S_j = \{q^{(h)} | q^{(h)} \in Q_r\}$ 中拟合最好的 q 向量作为题目 j 第 h 类别的 q 向量。这样的方法计算比较耗时，如当 K=5 时，对每个类别的 q 向量的确定，穷尽算法需要估计 $2^5 - 1 = 31$ 次。

顺序算法包括：（1）增加属性算法（forward search algorithm）是先从单属性 q 向量中

挑选出拟合最好的 q 向量，记为 $q^{(y)}$，再比较集合 $S_j=\left\{q^{(h)}\middle|q^{(h)}\in Q_r,q^{(y)}\leq q^{(h)},K^{(h)}=K^{(y)}+1\right\}$中的 q 向量与 $q^{(y)}$ 的拟合。如果集合中拟合最好的 q 向量的拟合指标优于 $q^{(y)}$，则用该 q 向量更新 $q^{(y)}$。重复该步骤直到 S_j 中没有 q 向量拟合优于 $q^{(y)}$ 或 $q^{(y)}$ 包含了所有属性（即 $\boldsymbol{q}^{(y)}=\boldsymbol{1}$）。（2）删除属性算法（backward search algorithm）是从全为 1 的 q 向量出发，即 $\boldsymbol{q}^{(y)}=\boldsymbol{1}$。然后比较集合 $S_j=\left\{q^{(h)}\middle|q^{(h)}\in Q_r,q^{(h)}\leq q^{(y)},K^{(h)}=K^{(y)}-1\right\}$中的 q 向量与 $q^{(y)}$ 的拟合，如果集合中拟合最好的 q 向量的拟合指标优于 $q^{(y)}$，则用该 q 向量更新 $q^{(y)}$。重复该步骤直到 S_j 中没有 q 向量拟合优于 $q^{(y)}$ 或 $q^{(y)}$ 只测量了一个属性（即 $K^{(y)}=1$）。（3）先增加属性后删除属性算法（forward-then-backward search algorithm）是将原始 Q 矩阵中专家给定的 q 向量作为 $q^{(y)}$，在此基础上先进行增加属性算法，然后进行删除属性算法。（4）先删除属性后增加属性算法（backward-then-forward search algorithm）也将原始 Q 矩阵中专家给定的 q 向量作为 $q^{(y)}$，在此基础上先进行删除属性算法，然后进行增加属性算法。

顺序算法中增加属性算法（forward search algorithm）和删除属性算法（backward search algorithm）并没有利用到专家给定的 Q 矩阵信息，而后两种算法则是在专家给定 q 向量的基础上进行的搜索算法。此外，后两种顺序算法在每个类别上需要估计的次数会根据该类别 q 向量的错误程度变化，但相对于穷尽算法，顺序算法可以大大减少计算次数。

3.5 Q 矩阵修正步骤

将需要修正的原始 Q 矩阵定义为 $Q^{(0)}$，这个 Q 矩阵通常是由专家界定。假设一个测验包含 J 个题目，每个题目包含 H_j 个类别（每个题目类别数可能不同），则该测验共有 $H=\sum_{j=1}^{J}H_j$ 个类别，将所有类别的集合定义为 $S^{(0)}=\{1,\ldots\ldots,H\}$。具体步骤如下：

步骤 1：从测验（$S^{(0)}$）中抽取出第一个题目 j，对其第 1 个类别的 q 向量进行验证。其他 $J-1$ 个题目以及题目 j 其他类别的 Q 矩阵保持不变。

步骤 2：使用顺序算法根据-2LL 指标（如果修正方法为 AIC 或 BIC 方法，则分别计算 AIC 和 BIC 指标）挑选出题目 j 第 1 个类别最优-2LL 指标（或 AIC、BIC 指标）所对应的 q

向量。

步骤 3： 重复步骤 1-步骤 2，确定第 j 题其他类别的最优 q 向量，方法与题目 j 第 1 个类别的方法相同。

步骤 4： 根据步骤 1-步骤 3，则完成对题目 j 所有类别最优 q 向量的确定。重复步骤 1-步骤 3，对剩余题目所有类别的最优 q 向量进行确定。直到确定完所有类别的最优 q 向量。

步骤 5： 在所有类别中，挑选修改后相对拟合指标（-2LL、AIC、BIC）达到最优的 q 向量进行修改，并将该类别从 $S^{(0)}$ 中移除。删除类别后的集合表示为 $S^{(1)}$，修改后的 Q 矩阵表示为 $Q^{(1)}$。

步骤 6： 验证 $Q^{(1)}$ 与 $Q^{(0)}$ 是否相同。如果 $Q^{(0)} \neq Q^{(1)}$，则将 $Q^{(0)}$ 替换为 $Q^{(1)}$，$S^{(0)}$ 替换为 $S^{(1)}$，重复步骤 1-步骤 5。当 $S^{(0)} = \varnothing$ 或 $Q^{(0)} = Q^{(1)}$，则算法停止。

为了探讨不同方法在多级计分认知诊断中 Q 矩阵修正的效果。模拟研究考查了不同方法在不同样本量、Q 矩阵错误类型以及不同的多级计分认知诊断模型下的效果，并将其与 Ma 和 de la Torre（2019）的 stepwise 方法进行比较。具体为：研究一：不同方法在简化多级计分认知诊断模型（seq-DINA 和 seq-RRUM）的效果及其比较研究；研究二：不同方法在饱和多级计分认知诊断模型（seq-GDINA）的效果及其比较研究。

4 研究一：不同方法在 seq-DINA 和 seq-RRUM 模型中的比较研究

4.1 研究一实验设计

4.1.1 Q 矩阵

本研究采用的 Q 矩阵（Ma & de la Torre，2016）如下，共包含了 21 个题目，5 个属性，Q 矩阵见表 1。

表 1 测验 Q 矩阵

题目	类别	A1	A2	A3	A4	A5	题目	类别	A1	A2	A3	A4	A5
1	1	1	0	0	0	0	11	1	1	1	0	0	0
1	2	0	1	0	0	0	11	2	0	0	0	0	1
2	1	0	0	1	0	0	12	1	1	1	1	0	0
2	2	0	0	0	1	0	12	2	0	0	0	1	1
3	1	0	0	0	0	1	13	1	1	1	0	0	0
3	2	1	0	0	0	0	13	2	0	0	1	1	1
4	1	0	0	0	0	1	14	1	1	0	1	0	0
4	2	0	0	0	1	0	14	2	0	0	0	1	0
5	1	0	0	1	0	0	14	3	0	0	0	0	1

5	2	0	1	0	0	0	15	1	0	0	0	0	1
6	1	1	0	0	0	0	15	2	0	0	1	1	0
6	2	0	1	1	0	0	15	3	0	1	0	0	0
7	1	0	0	1	0	0	16	1	1	0	0	0	0
7	2	0	0	0	1	1	16	2	0	1	0	0	0
8	1	0	0	0	0	1	16	3	0	0	1	1	0
8	2	1	1	0	0	0	17	1	1	0	0	0	0
9	1	0	0	0	1	1	18	1	0	1	0	0	0
9	2	0	0	1	0	0	19	1	0	0	1	0	0
10	1	0	1	0	1	0	20	1	0	0	0	1	0
10	2	1	0	0	0	0	21	1	0	0	0	0	1

4.1.2 认知诊断模型、被试参数和题目参数模拟

研究一使用的模型为 seq-DINA 和 seq-RRUM 模型。被试掌握模式由多元正态分布（multidimensional normal distribution）$MVN(\mathbf{0},\boldsymbol{\Sigma})$ 产生，参考已有研究（chen, 2017；Liu, Xin, Andersson, & Tian, 2019）属性间相关设置为 0.5。样本量分别为 500、1000 和 2000 人，代表小样本、中等样本和大样本。题目参数模拟方法为掌握项目 j 第 h 类别全部属性的被试得 h 分的概率从[.75~1]中随机产生,即 $S_j\left(h\middle|\alpha^*_{ljh}=1\right)=U[0.75,1]$，未掌握项目 j 第 h 类别任何属性的被试得 h 分的概率从[0~.25]中随机产生，即 $S_j\left(h\middle|\alpha^*_{ljh}=0\right)=U[0,0.25]$。对于 seq-RRUM 模型，其他掌握模式的被试得 h 分的概率从 $\left[S_j\left(h\middle|\alpha^*_{ljh}=0\right),\ S_j\left(h\middle|\alpha^*_{ljh}=1\right)\right]$ 中随机产生且服从单调性约束，即掌握属性个数多的被试在题目 j 上得 h 分的概率大于掌握属性个数少的被试，即当 $\alpha_l \geq \alpha'_l$，$S_j\left(h\middle|\alpha_l\right)\geq S_j\left(h\middle|\alpha'_l\right)$。

4.1.3 Q 矩阵错误模拟

参考已有研究（Chen et al., 2013；Liu, Tian, & Xin, 2016；Chen，2017；Liu et al., 2019），分别考察 Q 矩阵中有属性冗余、属性缺失、属性既缺失又冗余等情况，分别设置了以下 6 种 Q 矩阵错误类型。Q1 为随机挑选 5 个测量了一个属性的类别，随机将每个类别中一个为“0”的属性改为“1”。Q2 为随机挑选 5 个测量了 2 个属性以上的类别，随机将每个类别中一个为“1”的属性改为“0”。Q3 为随机挑选 5 个测量了 2 个属性以上的类别，随机将每个类别中一个为“0”的属性改为“1”，将其中一个为“1”的属性改为“0”。Q4 则包含了前三个 Q 向量的所有错误。Q5 和 Q6 则分别模拟了 10%和 20%的随机错误，但保证每个类别测量了最多 3 个属性最少 1 个属性。

表 2　Q 矩阵错误类型

Q	Q 矩阵错误模拟规则	调整的类别	调整的属性个数	备注
Q1	$q_{jk}=0\rightarrow q_{jk}=1$	$K^*_{jh}=1$ 的类别	5	属性冗余
Q2	$q_{jk}=1\rightarrow q_{jk}=0$	$K^*_{jh}>2$ 的类别	5	属性缺失
Q3	$q_{jk}=0\rightarrow q_{jk}=1, q_{jk'}=1\rightarrow q_{jk'}=0$	$K^*_{jh}>2$ 的类别	10	属性既冗余又缺失
Q4	$q_{jk}=0\rightarrow q_{jk}=1$ $q_{jk}=1\rightarrow q_{jk}=0$ $q_{jk}=0\rightarrow q_{jk}=1, q_{j'k}=1\rightarrow q_{j'k}=0$	分别为 Q1、Q2 和 Q3 的类别	20	Q1、Q2 和 Q3 的组合
Q5	10% 随机调整	随机	20	调整后 $1<K^*_{jh}<3$
Q6	20% 随机调整	随机	40	调整后 $1<K^*_{jh}<3$

4.1.4 被试作答模拟

根据模拟的被试参数和题目参数分别计算被试 i 在题目 j 上所有类别的得分概率 $P_{ij}=\left[P(X_j=0|\alpha^*_{ij}),\cdots,P(X_j=H_j|\alpha^*_{ij})\right]$，以 P_{ij} 为概率在类别分布（Categorical distribution）中产生被试 i 在题目 j 上的作答反应得分，即 $X_{ij}=Cat(P_{ij})$。

4.1.5 评价指标

计算每次修正后的 Q 矩阵与真实 Q 矩阵每个类别属性模式的一致性作为模式判准率（pattern match ratio，PMR）。计算每次修正后的 Q 矩阵与真实 Q 矩阵属性的一致性作为属性判准率（attribute match ratio，AMR）。以及 FPR（False Positive Rate）和 TPR（True Positive Rate）分别代表错误标定的属性未被修改的比例和正确标定的属性未被修改的比例。所有实验均重复 200 次，然后再计算 200 次实验的平均 PMR、AMR、FPR 以及 TPR。

$$PMR=\frac{\sum_{j=1}^{J}\sum_{h=1}^{H_j}n_{jh\_correct}}{\sum_{j=1}^{J}H_j} \tag{9}$$

$$AMR=\frac{\sum_{j=1}^{J}\sum_{h=1}^{H_j}\sum_{k=1}^{K}n_{jhk\_correct}}{K\times\sum_{j=1}^{J}H_j} \tag{10}$$

公式 9 和 10 中，J 为题目个数，H_j 为第 j 题的类别数量，$n_{jh\_correct}$ 为修正后的第 j 题第 h 类别的 q 向量是否与真实 Q 矩阵中第 j 题第 h 类别一致，完全一致则为 1，否则为 0。公式（10）中，K 为属性个数，$n_{jhk\_correct}$ 表示修正后的第 j 题第 h 类别的第 k 个属性（为 0 或者 1）是否与真实 Q 矩阵中第 j 题第 h 类别第 k 个属性一致，如果一致则为 1，否则为 0。

为了比较修正前后 Q 矩阵的优劣，分别计算 Q 矩阵修正前后的绝对拟合指标 RMSEA（Liu et al., 2016），并计算 200 次试验的平均值。

在复杂模型（seq-RRUM 和 seq-GDINA）中，三种模型相对拟合统计量在所有实验条件下均是 BIC 指标的 Q 矩阵恢复率最高，且-2LL 和 AIC 指标修正后 Q 矩阵的 RMSEA 指标均不如 BIC 方法修正后的结果。而在简化模型（seq-DINA）中，AIC、BIC 指标与-2LL 指标是等价的。同时，四种顺序算法中先增加属性后删除属性的算法（forward-then-backward search algorithm）与先删除属性后增加属性的算法（backward-then-forward search algorithm）的表现几乎一致，而增加属性算法（forward search algorithm）和删除属性算法（backward search algorithm）在一些实验条件下略低于前两种算法。而相对于穷尽算法，先增加属性后删除属性的算法并不会降低 Q 矩阵修正的正确率，穷尽算法与先增加属性后删除属性的算法之间属性判准率差异不超过 1%。因此为了报告的简洁性，本文只报告先增加属性后删除属性算法的 BIC 方法和 stepwise 方法 Q 矩阵修正后的结果及 RMSEA 指标。

4.2 研究一实验结果

表 3 和表 4 分别呈现了 BIC 方法和 stepwise 方法在 seq-DINA 模型以及 seq-RRUM 模型中的实验结果。根据表 3 的结果可知，在 seq-DINA 模型下，使用 BIC 方法进行 Q 矩阵修正具有很好的效果。在所有试验条件下，BIC 方法修正 Q 矩阵的平均模式判准率和属性判准率分别为 83.0%和 96.9%；stepwise 方法的平均模式判准率和属性判准率为 78.1%和 95.7%。总体上，BIC 方法的模式判准率和属性判准率略高于 stepwise 方法，大多数条件下两者属性判准率差异不超过 1%。

在不同 Q 矩阵错误类型上，BIC 方法对 Q1-Q5 的恢复率相当，属性判准率在 95%-98%之间；而对于 Q6 的修正结果略差于前 5 个 Q 矩阵，属性判准率在 93%-95%之间。Stepwise 方法在不同错误 Q 矩阵上表现也相近，属性判准率均在 92%以上。因此在 seq-DINA 模型下，

不同 Q 矩阵的错误类型对 BIC 方法和 stepwise 方法的整体修正效果影响不大。

对于 FPR 和 TPR 指标，BIC 方法的 TPR 指标在所有实验条件下均能达到 95%左右，表明 BIC 方法不会轻易更改正确标定的属性。而 BIC 方法的 FPR 指标，在 Q2 时低于其他 Q 矩阵，这也许是由于 DINA 模型的特性造成的，即需要掌握题目测量的所有属性才能答对，因此 BIC 方法倾向于将缺失的属性修改过来。而 stepwise 方法的 TPR 指标与 BIC 方法相差不大，而 FPR 指标在 Q2-Q4 下比其他错误 Q 矩阵更高，说明 stepwise 方法对属性缺失不够敏感。

在样本量对 Q 矩阵修正效果的影响上，样本量越大，两种方法对 Q 矩阵的恢复率越高。当 N=500 时，BIC 方法和 stepwise 方法的平均属性判准率 95.6%和 94.6%；当 N=2000 时，BIC 方法和 stepwise 方法的平均属性判准率为 97.9%和 96.7%。因此增加样本量可以提高两种方法的 Q 矩阵的修正效果。

在修正前后 Q 矩阵的拟合上来看，两种方法修正后的 Q 矩阵的 RMSEA 值均低于修正前的 Q 矩阵，说明修正后的 Q 矩阵与数据更加拟合。在所有实验条件下，修正前的 Q 矩阵平均 RMSEA 值为 0.048，BIC 方法和 stepwise 方法修正以后 Q 矩阵的平均 RMSEA 值为 0.007 和 0.017。BIC 方法修正后的 Q 矩阵比 stepwise 方法修正后的 Q 矩阵拟合更好，平均差异为 0.01。此外，样本量越大，BIC 方法修正后的 Q 矩阵的 RMSEA 值更小，如在 Q1-Q5 中且 N=2000 时，BIC 方法修正后的 RMSEA 值在 0.003~0.004 左右。

根据表 4 的结果，在 seq-RRUM 模型中，总体上 BIC 方法表现优于 stepwise 方法。在所有实验条件下，stepwise 方法和 BIC 方法的模式判准率分为 78.1%和 87.5%，属性判准率分别为 96%和 98%。

对于样本量的影响，两种方法对 Q 矩阵修正的模式判准率和属性判准率随着被试人数的增加而增加。当 N=500 时，stepwise 方法和 BIC 方法的平均属性判准率为 94.8%和 97.4%；当 N=2000 时，stepwise 方法和 BIC 方法的平均属性判准率为 96.9%和 98.6%。

在不同 Q 矩阵错误类型上，stepwise 方法和 BIC 方法对 Q 矩阵的整体恢复率受 Q 矩阵错误类型的影响不大。如在 Q1-Q5 中，stepwise 方法和 BIC 方法的属性判准率均在 96%和 98%左右波动。而在 Q6 中，两种方法的属性判准率有所降低，但是降幅不大。

对于 FPR 和 TPR 指标，在所有 Q 矩阵错误类型下，两种方法的 TPR 指标相近，均在 95%以上，而两种方法的 FPR 指标在 Q2-Q6 中略高于 Q1 中，说明两种方法对属性冗余更加敏感。这同样也说明 Q 矩阵中包含属性缺失对两种方法的影响更大。

对于修正前后 Q 矩阵的绝对拟合，与 seq-DINA 模型中略有不同，即在 Q1 条件下，两

种方法修正后的 Q 矩阵与修正前的 Q 矩阵的 RMSEA 指标几乎一致。这是由于 Q1 为属性冗余，而在复杂模型中，属性冗余并不会导致拟合变差。而在 Q2-Q6 中，修正前的 Q 矩阵的平均 RMSEA 分别为 0.037，stepwise 方法和 BIC 方法修正后的 Q 矩阵平均 RMSEA 分别为 0.007 和 0.005，说明两种方法修正后的 Q 矩阵与数据更加拟合。而 BIC 方法修正后的 Q 矩阵平均拟合要优于 stepwise 方法修正后的 Q 矩阵。同样，随着样本量的增加，stepwise 方法和 BIC 方法修正后的 Q 矩阵具有更好的拟合值，如当 N=2000 时，stepwise 方法和 BIC 方法修正后的 Q 矩阵的平均 RMSEA 为 0.006 和 0.003。

表 3 BIC 方法和 stepwise 方法在 seq-DINA 模型中 200 次实验的平均结果

Q-matrix	N	PMR		AMR		FPR		TPR		RMSEA		
		Stepwise	BIC	Stepwise	BIC	Stepwise	BIC	Stepwise	BIC	Q_W	$Q_{stepwise}$	Q_{BIC}
Q1	500	0.795	0.788	0.957	0.963	0.118	0.157	0.958	0.965	0.017	0.015	0.007
	1000	0.879	0.863	0.975	0.977	0.065	0.074	0.975	0.978	0.018	0.009	0.005
	2000	0.918	0.911	0.984	0.986	0.048	0.049	0.985	0.986	0.019	0.005	0.003
Q2	500	0.763	0.790	0.953	0.962	0.367	0.021	0.958	0.962	0.017	0.016	0.007
	1000	0.826	0.856	0.967	0.975	0.257	0.004	0.971	0.975	0.016	0.011	0.005
	2000	0.865	0.903	0.976	0.984	0.219	0.002	0.980	0.984	0.017	0.008	0.003
Q3	500	0.758	0.786	0.952	0.962	0.339	0.126	0.963	0.966	0.033	0.016	0.006
	1000	0.815	0.861	0.964	0.976	0.251	0.089	0.972	0.979	0.034	0.010	0.005
	2000	0.856	0.910	0.974	0.985	0.180	0.065	0.980	0.987	0.035	0.009	0.004
Q4	500	0.680	0.776	0.938	0.961	0.363	0.110	0.962	0.966	0.041	0.020	0.007
	1000	0.721	0.853	0.950	0.975	0.288	0.064	0.968	0.978	0.041	0.015	0.005
	2000	0.745	0.905	0.956	0.984	0.251	0.040	0.972	0.986	0.042	0.013	0.003
Q5	500	0.760	0.777	0.951	0.959	0.112	0.068	0.956	0.961	0.075	0.020	0.008
	1000	0.835	0.851	0.968	0.975	0.082	0.041	0.972	0.976	0.073	0.011	0.004
	2000	0.874	0.903	0.975	0.984	0.065	0.022	0.978	0.984	0.076	0.013	0.004
Q6	500	0.629	0.687	0.924	0.933	0.184	0.105	0.943	0.940	0.100	0.035	0.015
	1000	0.656	0.744	0.931	0.942	0.173	0.097	0.949	0.949	0.102	0.038	0.017
	2000	0.687	0.793	0.935	0.953	0.163	0.081	0.951	0.959	0.102	0.037	0.010

注：Q_w（Q_wrong） 为修正前的 Q 矩阵，下同。

表 4　BIC 方法和 stepwise 方法在 seq-RRUM 模型中 200 次实验的平均结果

Q-matrix	N	PMR		AMR		FPR		TPR		RMSEA		
		Stepwise	BIC	Stepwise	BIC	Stepwise	BIC	Stepwise	BIC	Q_W	$Q_{stepwise}$	Q_{BIC}
Q1	500	0.750	0.841	0.952	0.975	0.083	0.022	0.952	0.975	0.006	0.007	0.006
	1000	0.823	0.884	0.968	0.982	0.037	0.041	0.968	0.983	0.005	0.005	0.005
	2000	0.864	0.915	0.976	0.987	0.029	0.020	0.977	0.987	0.004	0.005	0.004
Q2	500	0.746	0.839	0.953	0.975	0.332	0.199	0.958	0.978	0.027	0.008	0.007
	1000	0.819	0.890	0.968	0.983	0.264	0.153	0.972	0.986	0.026	0.006	0.005
	2000	0.843	0.919	0.974	0.988	0.252	0.117	0.978	0.990	0.026	0.005	0.003
Q3	500	0.734	0.847	0.949	0.976	0.300	0.121	0.959	0.980	0.022	0.008	0.006
	1000	0.794	0.877	0.963	0.981	0.241	0.086	0.971	0.983	0.023	0.006	0.005
	2000	0.843	0.914	0.973	0.987	0.171	0.057	0.979	0.989	0.023	0.005	0.003
Q4	500	0.714	0.832	0.946	0.974	0.275	0.123	0.963	0.981	0.030	0.008	0.007
	1000	0.770	0.881	0.959	0.982	0.215	0.085	0.973	0.987	0.031	0.006	0.005
	2000	0.796	0.917	0.966	0.987	0.195	0.058	0.978	0.991	0.032	0.005	0.003
Q5	500	0.751	0.841	0.952	0.975	0.098	0.047	0.956	0.976	0.039	0.008	0.006
	1000	0.807	0.880	0.965	0.982	0.073	0.038	0.968	0.983	0.035	0.005	0.005
	2000	0.849	0.914	0.974	0.987	0.053	0.021	0.976	0.987	0.039	0.005	0.004
Q6	500	0.686	0.817	0.941	0.968	0.134	0.063	0.953	0.973	0.063	0.014	0.008
	1000	0.726	0.848	0.948	0.973	0.127	0.058	0.960	0.978	0.070	0.012	0.007
	2000	0.748	0.896	0.953	0.982	0.120	0.032	0.966	0.984	0.063	0.009	0.004

5 研究二：不同方法在 seq-GDINA 模型中的比较研究

研究一中多级计分的认知诊断模型在每个类别上的链接函数具有一定的约束，可以由 seq-GDINA 模型转换而来。而 seq-GDINA 是饱和的模型，因此具有更广的适用性。研究二则是对不同方法在 seq-GDINA 模型中的效果进行验证及比较。

5.1 研究二实验设计

研究二的实验设计与研究一的实验设计相似，不同的是研究二使用的是 seq-GDINA 模型。其余实验条件请见研究一。

5.2 研究二实验结果

表 5 呈现了 BIC 方法和 stepwise 方法在 seq-GDINA 模型中的实验结果。从表 5 可以看出，与研究一的结果相似，总体上 BIC 方法略优于 stepwise 方法，BIC 方法和 stepwise 方法的平均模式判准率分别为 90.5%和 84.5%，属性判准率分别为 98.6%和 97.1%。两种方法的 Q 矩阵恢复率随着被试人数的增加而逐渐增加，如当 N=500 时，BIC 方法的平均模式判准率和平均属性判准率分别为 86%和 97.9%，stepwise 方法的平均模式判准率和属性判准率分别为 78%和 95.9%；当 N=2000 时，BIC 方法的平均模式判准率和平均属性判准率分别为 94.8%和 99.3%，stepwise 方法的平均模式判准率和属性判准率分别为 90.8%和 98.5%。不同 Q 矩阵错误类型下两种方法对 Q 矩阵的整体恢复率差异不大。在修正前后 Q 矩阵的绝对拟合上，在 Q1 条件下，两种方法修正后 Q 矩阵的 RMSEA 指标几乎与修正前 Q 矩阵一致。这与 seq-RRUM 模型中结果一样，这是由于 Q1 矩阵中属性冗余导致的。而在 Q2-Q6 中，修正前 Q 矩阵的平均 RMSEA 为 0.036，stepwise 方法和 BIC 方法修正后 Q 矩阵的平均 RMSEA 分别为 0.007 和 0.006，说明修正后的 Q 矩阵与数据更加拟合。随着样本量的增加，两种方法修正后的 Q 矩阵具有更好的拟合值。

表 5　BIC 方法和 stepwise 方法在 seq-GDINA 模型中 200 次实验的平均结果

Q-matrix	N	PMR		AMR		FPR		TPR		RMSEA		
		Stepwise	BIC	Stepwise	BIC	Stepwise	BIC	Stepwise	BIC	Q_W	$Q_{stepwise}$	Q_{BIC}
Q1	500	0.795	0.861	0.961	0.979	0.075	0.006	0.962	0.979	0.007	0.007	0.007
	1000	0.875	0.913	0.978	0.987	0.032	0.004	0.978	0.987	0.005	0.006	0.005
	2000	0.919	0.950	0.986	0.993	0.020	0.001	0.986	0.993	0.004	0.005	0.004
Q2	500	0.799	0.864	0.964	0.980	0.209	0.211	0.967	0.983	0.029	0.008	0.009
	1000	0.877	0.916	0.979	0.988	0.108	0.110	0.980	0.990	0.030	0.006	0.006
	2000	0.915	0.948	0.986	0.993	0.093	0.067	0.987	0.994	0.030	0.004	0.004
Q3	500	0.794	0.867	0.961	0.980	0.236	0.125	0.969	0.984	0.022	0.007	0.008
	1000	0.854	0.910	0.974	0.987	0.170	0.069	0.979	0.989	0.025	0.006	0.006
	2000	0.904	0.949	0.984	0.993	0.106	0.038	0.988	0.994	0.025	0.005	0.003
Q4	500	0.775	0.863	0.958	0.979	0.193	0.106	0.970	0.985	0.033	0.009	0.008
	1000	0.840	0.912	0.972	0.987	0.136	0.062	0.981	0.991	0.033	0.007	0.006
	2000	0.884	0.945	0.981	0.992	0.112	0.040	0.989	0.994	0.034	0.005	0.004
Q5	500	0.786	0.866	0.959	0.980	0.086	0.040	0.962	0.981	0.034	0.009	0.009
	1000	0.859	0.911	0.975	0.987	0.053	0.023	0.977	0.988	0.036	0.006	0.006
	2000	0.912	0.949	0.985	0.993	0.031	0.011	0.987	0.993	0.041	0.005	0.004
Q6	500	0.731	0.838	0.948	0.973	0.122	0.059	0.960	0.978	0.061	0.015	0.009
	1000	0.784	0.885	0.959	0.980	0.104	0.039	0.970	0.983	0.066	0.010	0.007
	2000	0.913	0.948	0.985	0.992	0.037	0.015	0.987	0.993	0.041	0.004	0.004

6 研究三：实证数据分析

本研究采用两个 TIMSS（Trends in International Mathematics and Science Study）数据，分别为 2011 年 8 年级和 2007 年 4 年级数学测试的数据。TIMSS 2011 年的数据由 Park, Lee 和 Johnson（2017）标定了 Q 矩阵，Ma 和 de la Torre(2019)将该数据用于多级计分 Q 矩阵修正的分析。该数据共包括 23 个题目、7 个属性，共 748 名学生的作答。其中第 11 题为多级计分的题目，其余题目为 0-1 计分的题目，Q 矩阵见表 6。

表 6 TIMSS 2011（8 年级）数据 Q 矩阵及修正结果

Item	Code	类别	A1	A2	A3	A4	A5	A6	A7
1	M042041	1	0	1	0	0	0	0	0
2	M042024	1	0	1	0	0	0	0	0
3	M042016	1	1	0	0	0	0	0	1*#
4	M042002	1	1	0	0	0	0	0	0
5	M042198A	1	0	0	1	0	0	0	0*#
6	M042198B	1	0	0	1	0	0	0	0
7	M042198C	1	0	0	1	0	0*	0	0
8	M042077	1	1	0	0	1	0	0	0
9	M042235	1	0	0	0	1	0*	0	0
10	M042150	1	0	0	0	0	1	0	0
11	M042300Z	1	0	0	0	0	0	1	1
11	M042300Z	2	0	0	0	0	1	0	0
12	M042169A	1	0*	0	0	0	0	0	1
13	M042169B	1	0	0	0	0	0	0	1
14	M042169C	1	0*	0	0	0	0	0	1
15	M032352	1	1	0	1*#	0	0	0	1*
16	M032725	1	0	1*	0	0	0	0*#	0
17	M032738	1	0	0	0	1	0	0	0
18	M032295	1	0	0	0	1	0	0	0
19	M032331	1	0	0	0	0	1	1	0
20	M032679	1	0	0	0	0	1	1*	0
21	M032047	1	1	0	0	1*#	0	0	0
22	M032398	1	0*	0	0	0	1	0	0
23	M032424	1	0	0*#	0	1	0	0	0

注：A1,整数和自然数；A2，分数、小数和比例；A3，模式；A4，表达式、方程式和函数；A5，线条、角度和形状；A6，位置和运动；A7，数据的组织、表达和解读。“*”为 BIC 方法调整的属性；“#”为 stepwise 方法调整的属性。

分析之前先比较该数据与各模型之间的拟合指标（偏差、AIC、BIC），结果显示各个指标对应的最优模型不同，因此为了避免模型选择错误，这里使用 seq-GDINA 模型来拟合数据，而对于 0-1 计分的题目则等价于用 GDINA 模型来拟合数据。两种方法对原始 Q 矩阵的调整结果见表 6，其中带“*”的属性为 BIC 方法建议调整的属性，带“#”的属性为 stepwise

方法建议调整的属性。BIC 方法共调整了 12 个题目 14 个属性，stepwise 方法调整了共 6 个题目 6 个属性，并且 stepwise 方法调整的 6 个属性全部包含于 BIC 方法调整的属性中。而对于第 11 题，两种方法均未对该题的两个类别 q 向量进行调整。

表 7 呈现了不同 Q 矩阵之间的一致率，BIC 方法和 stepwise 方法修正后的 Q 矩阵与原始 Q 矩阵之间的一致率分别为 0.92 和 0.96，而 BIC 方法和 stepwise 方法修正后的 Q 矩阵之间的一致率为 0.95，Q 矩阵之间具有较高的一致率。

表 7 TIMSS 2011（8 年级）数据不同方法 Q 矩阵修正一致率

	$Q_{original}$	Q_{BIC}	$Q_{stepwise}$
$Q_{original}$	1		
Q_{BIC}	0.92	1	
$Q_{stepwise}$	0.96	0.95	1

为了比较两种方法修正后的 Q 矩阵与原有的 Q 矩阵，分别计算修正前后 Q 矩阵的相对拟合指标（-2*LL、AIC、BIC）和绝对拟合指标（M_2 检验(Liu et al., 2016)、RMSEA(Liu et al., 2016)和 SRMSR），结果如表 8。从表 8 可以看出，两种方法修正后的 Q 矩阵在相对拟合指标上均优于原有 Q 矩阵。在绝对拟合上，修正前的 Q 矩阵 M_2 检验为 $p<0.01$，而修正后的 Q 矩阵检验结果为 $p=0.2$ 和 0.3，因此修正后的 Q 矩阵与数据更加拟合。在 RMSEA 和 SRMSR 指标上，两种方法修正后的 Q 矩阵也优于原有 Q 矩阵。而两种方法修正后 Q 矩阵的拟合指标相近，$Q_{stepwise}$ 的 M_2 检验和 RMSEA 优于与 Q_{BIC}，而 Q_{BIC} 的相对拟合指标和 SRMSR 优于 $Q_{stepwise}$。

表 8 TIMSS 2011（8 年级）数据原有 Q 矩阵和两种方法修正后 Q 矩阵的拟合指标

Q	相对拟合指标			绝对拟合指标				
				M_2 检验				
	-2*LL	AIC	BIC	M_2	df	P	RMSEA	SRMSR
$Q_{original}$	18888.23	19274.23	20165.39	123.51	83	0.003	0.026	0.059
Q_{BIC}	18624.73	19014.73	19915.13	89.02	81	0.254	0.012	0.044
$Q_{stepwise}$	18757.88	19139.88	20021.88	89.90	85	0.337	0.009	0.050

此外，我们又对 TIMSS 2007 年的数据进行了分析。该数据由 Lee，Park 和 Taylan（2011）标定了 Q 矩阵，Ma 和 de la Torre(2016)将该数据用于多级计分模型的分析。该数据共包括 11 个题目、8 个属性，共 823 名学生的作答，其中第 3、7、9 题为多级计分的题目，其余题目为 0-1 计分的题目。该数据原始 Q 矩阵如表 9 所示。

表 9 TIMSS 2007（4 年级）数据 Q 矩阵及修正结果

Item	Code	类别	A1	A2	A3	A4	A5	A6	A7	A8
1	M041052	1	1	1	0	0	0	0	0	0
2	M041281	1	0	1	1*	0	1*	0	0	0
3	M041275	1	1	0	0	0	0	1	0	1*
3	M041275	2	1*	0	0	0	0	1	0	1*
4	M031303	1	0	1	1	0	0	0	0	0
5	M031309	1	0	1	1	0	0	0	0	0
6	M031245	1	0	1	0	1	0	0	0	0
7	M031242A	1	0	1	1	0	1	0	0	0
7	M031242B	2	0	0	0	0	0	0	1	0
8	M031242C	1	0	1*	1*	0	1	0	1*	0
9	M031247	1	0	1*	1	1	0	0	0	0
9	M031247	2	0	1	1	1	0	0	0	0
10	M031173	1	0*	1*	1	0	0	0	0	0
11	M031172	1	1*	1*	0	0	0	1*	0	1

注：A1，表示、比较和排序整数以及说明排序位置的价值；A2，识别倍数，使用四步操作计算整数并估算；A3，解决问题，包括现实情境中的问题（如测量和资金问题）；A4，查找缺失数据，或对包含未知的句子和表达进行操作和建模；A5，描述模式及其扩展的关系，通过给定规则生成整数对，并为给定整数对的每个关系确定规则；A6，从表格，象形图，条形图和饼图中读取数据；A7，比较和理解如何使用数据中的信息； A8，了解不同的表达，用表格、象形图和条形图组织数据。 “*”为 BIC 方法调整的属性。

分析结果为 stepwise 方法和 BIC 方法分别调整了 17 个属性和 14 个属性，而 stepwise 方法调整后属性 5（A5）没有被任何题目测量，因此这里不详细展示该方法的具体结果，BIC 调整后的 Q 矩阵如表 9。同样计算 BIC 方法修正后的 Q 矩阵与原始 Q 矩阵的绝对拟合和相对拟合指标，由于该 Q 矩阵修正前后 M_2 检验的自由度过低，因此这里不能进行 M_2 检验。原有 Q 矩阵和 BIC 方法修正后的 SRMER 指标分别为 0.0312 和 0.0246。在相对拟合指标上，BIC 方法修正后的 Q 矩阵（AIC=11222.25; BIC=12677.42）也比原有 Q 矩阵（AIC=11513.79; BIC=13195.01）拟合更好。因此两个实证数据分析的结果均显示 BIC 方法修正后的 Q 矩阵与数据拟合更好。

7 结论与讨论

7.1 结论

本研究探讨了基于类别水平的多级计分认知诊断测验 Q 矩阵修正，并采用 Monte Carlo 模拟研究和实证研究验证和比较了 stepwise 方法和相对拟合指标用于 Q 矩阵修正的效果及特性，为实践中多级计分的测验 Q 矩阵修正提供了方法支持。研究发现：（1）BIC 方法对多级计分认知诊断模型的 Q 矩阵修正具有较高的模式判准率和属性判准率，其对 Q 矩阵的恢复率也高于 stepwise 方法，BIC 方法修正后的 Q 矩阵与数据更加拟合。（2）在复杂模型

中，相对拟合指标 BIC 比 AIC 和-2LL 表现更好，在实践中，使用者可以选择 BIC 法进行测验 Q 矩阵修正。（3）Q 矩阵修正效果受到被试人数的影响，增加被试人数可以提高 Q 矩阵修正的正确率。

7.2 讨论

（1）多级计分认知诊断 Q 矩阵修正方法

多级计分的题目是实际测验中常见的题型，并且多级计分题目比 0-1 计分题目能提供更多的信息，因此多级计分认知诊断模型的开发对认知诊断的发展具有重要作用。本研究将相对拟合统计量用于多级计分认知诊断模型 Q 矩阵修正中，并改进 Q 矩阵修正算法，研究发现 BIC 方法在多级计分模型中的 Q 矩阵修正具有很好的效果。并且该方法受到被试人数的影响较少，在不同 Q 矩阵错误类型下均有较好的修正效果。为了提高运算效率，本文中使用了顺序算法，模拟研究发现本文中使用的顺序算法与穷尽算法之间的属性判准率差异不超过 1%。而本文中的顺序算法为先增加属性后删除属性的算法，其表现与先删除属性再增加属性的算法一致。增加属性算法和删除属性的算法略差，这可能是由于后两种算法没有利用到专家给定的 q 向量信息。此外在模型不确定的情况下，可以使用饱和模型（seq-GDINA）来进行 Q 矩阵修正，模拟研究显示使用饱和模型进行 Q 矩阵修正并不会降低 Q 矩阵修正的效果。

（2）多级计分认知诊断下类别水平与项目水平 Q 矩阵修正

类别水平的 Q 矩阵需要在每个类别上分别标定 Q 矩阵，因此能更准确地探查出被试的解题过程，且分类准确性也更高。但是为每个类别标定 Q 矩阵不仅有难度且会增加 Q 矩阵标定的工作量。而项目水平的 Q 矩阵的标定相对简单，但产生的结果是忽略了每个步骤的信息，从而分类准确性有所降低(Ma & de la Torre，2016)。在 Q 矩阵修正上，对类别水平 Q 矩阵的修正也更难。本研究在多级计分模型下对类别水平的 Q 矩阵修正进行研究，并且发现 BIC 方法具有较好的效果，为多级计分模型下类别水平 Q 矩阵的标定与修正提供了方法支持。

（3）Q 矩阵修正结果应与专家意见相结合

从作答数据出发提出 Q 矩阵修正方法可以避免专家标定 Q 矩阵的主观性，也可以减轻专家的负担。但是客观方法标定的 Q 矩阵不能直接作为最终的 Q 矩阵，应该与专家的意见相结合。从作答数据出发进行的 Q 矩阵标定可以作为专家标定 Q 矩阵的参考和依据，但是不能忽视专家在测验设计和 Q 矩阵标定中的重要作用。而本文中 BIC 方法修正后的 Q 矩阵与数据更加拟合也并不代表修改的属性恰当，需要由专家最后决定是否对 Q 矩阵进行修改。

当客观方法得出的 Q 矩阵与专家意见不同时，可以由多个专家讨论决定，或者将有争议的题目删除。

（4）未来研究方向

本研究尝试提出在多级计分模型下的 Q 矩阵修正方法，并发现 BIC 方法可用于多级计分认知诊断模型的 Q 矩阵修正。但是本研究还存在一些需要进一步探究的地方，如不同题目参数质量对 Q 矩阵修正的影响、项目水平 Q 矩阵下不同方法的表现如何、属性间有层级关系时 Q 矩阵的修正效果等。此外对于多级计分模型下 Q 矩阵的标定还有更多的问题需要进行研究，如 Q 矩阵完备性和可识别性的推导证明、当属性个数有误时如何自动识别以及更多的真实数据研究等。总之，对多级计分模型下的 Q 矩阵修正方法还需要进一步的研究。

参考文献

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions on Automated Control, 19,* 716–723.

Chang, H.-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*, 1–20.

Chang, H.-H., & Wang, W. Y. (2016). ‘‘Internet plus’’ measurement and evaluation: a new way for adaptive learning. *Journal of Jiangxi Normal University (Natural Science), 40*(5), 441–455.

[张华华, 汪文义. (2016). “互联网+”测评:自适应学习之路. *江西师范大学学报(自然科学版), 40*(5), 441–455.]

Chen, J. S.(2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement, 41*, 277–293.

Chen, J. , Torre, J. & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50,* 123–140.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*, 598–618.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: development and applications. *Journal of Educational Measurement, 45*, 343–362.

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

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253–273.

Haertel, E. H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement, 8*, 333–346.

Hansen, M. (2013). *Hierarchical item response models for cognitive diagnosis.* unpublished doctoral dissertation. University of California at Los Angeles.

Hartz, S. M., & Roussos, L. A. (2008). The fusion model for skills diagnosis: blending theory with practice. *Educational Testing Service, Research Report, RR-08-71*. Princeton, NJ: Educational Testing Service.

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in massachusetts, minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11*, 144–177.

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. , Andersson, B. & Tian, W. (2019). Information matrix estimation procedures for cognitive diagnostic models. *British Journal of Mathematical and Statistical Psychology, 72,* 18–37

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

Ma, W. & Torre, J. (2019). An empirical Q-matrix validation method for the sequential generalized DINA model. *British Journal of Mathematical and Statistical Psychology.* https://doi.org/10.1111/bmsp.12156.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212.

Park, J. Y., Lee, Y.-S., & Johnson, M. S. (2017). An efficient standard error estimator of the DINA model parameters when analysing clustered data. *International Journal of Quantitative Research in Education, 4,* 159–190.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

Tu, D. B., Cai, Y., & Dai, H. Q. (2012). A new method of Q-Matrix validation based on DINA model. *ActaPsychologica Sinica, 44*(4), 558–568.

[涂冬波，蔡艳，戴海崎. (2012). 基于 DINA 模型的 Q 矩阵修正方法. *心理学报, 44*(4), 558–568.]

Tu, D. B., Cai, Y., Dai, H. Q., & Ding, S. L. (2010). A polytomous cognitive diagnosis model: P- DINA model. *ActaPsychologica Sinica, 42*(10), 1011–1020.

[涂冬波,蔡艳,戴海琦,丁树良.(2010). 一种多级评分的认知诊断模型:P-DINA 模型的开发.*心理学报, 42*(10), 1011–1020.]

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

Wang, D. X., Gao, X. L., Cai, Y., & Tu, D. B. (2018). A new Q-matrix estimation method: ICC based on ideal response. *Journal of Psychological Science, 41*(2), 466–474.

[汪大勋,高旭亮,蔡艳,涂冬波.（2018）.一种非参数化的 Q 矩阵估计方法:ICC-IR 方法开发. *心理科学, 41*(2), 466–474.]

Wang, D. X., Gao, X. L., Han, Y. T., & Tu, D. B. (2018). A simple and effective Q-matrix estimation method: from non-parametric perspective. *Journal of Psychological Science, 41*(1), 180–188.

[汪大勋，高旭亮，韩雨婷，涂冬波.(2018).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. *心理科学, 41*(1), 180–188.]

Xu, G. & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2017.1340889.

A method of Q-matrix validation for polytomous response cognitive diagnosis model based on relative fit statistics

WANG Daxun[1]; GAO Xuliang[2]; CAI Yan[1]; TU Dongbo[1]

([1] School of Psychology, Jiangxi Normal University, Nanchang, 330022) ([2] School of Psychology, Guizhou Normal University, Guiyang, 550000)

Abstract

Cognitive diagnostic assessments (CDAs) can provide fine-grained diagnostic information about students' knowledge states, so as to help to teach in accordance with the students' aptitude. The development of cognitive diagnosis model for polytomous response data expands the application scope of cognitive diagnostic assessment. As the basis of CDAs, Q-matrix has aroused more and more attention for the subjective tendency in Q-matrix construction that is typically performed by domain experts. Due to the subjective process of Q-matrix construction, there inevitably have some misspecifications in the Q-matrix, if left unchecked, can result in a serious negative impact on CDAs. To avoid the subjective tendency from experts and to improve the correctness of the Q-matrix, several objective Q-matrix validation methods have been proposed. Many Q-matrix validation methods have been proposed in dichotomous CDMs, however, the research of the Q-matrix validation method under polytomous CDMs is stalling lacking. To address this concern, several relative fit statistics (i.e., -2LL, AIC, BIC) were applied to the Q-matrix validation for polytomous cognitive diagnosis model in this research. The process of Q-matrix validation is as follows:

First, the reduced Q-matrix is represented by Q_r , which represents a set of potential q-vectors and contains $2^K - 1$ possible q-vectors when attributes are independent. When validating the q-vector of the first category of item *j*, all possible q-vectors in Q_r can be used as the q-vector of the first category of item *j*, and the Q-matrix of remaining items remains intact. From this, the item parameters and the attribute patterns of students can be estimated, and the -2LL, AIC, and BIC can be calculated accordingly. The q-vector with the largest likelihood (or smallest AIC/BIC) is regarded as the q-vector of the first category of item *j*. The q-vector of the next category of the item *j* can also be obtained in the same way. The algorithm stops when the validated Q-matrix is same as the previous Q-matrix, or every item has been reached. In order to improve the efficiency of the method, a sequential search algorithm was proposed.

Several simulation studies were conducted to evaluate the effectiveness and practicality of these methods, and the performance of the methods in this paper was compared with the stepwise method (Ma & de la Torre, 2019). Three experimental factors were considered in simulation studies, including sample size, Q-matrix error types and CDMs. The results show that (1) BIC method can be used for Q-matrix validation under polytomous response CDMs, and the performance of the BIC method is better than the stepwise method. (2) In general, the performance of the three methods from good to bad is the BIC method, AIC method, and -2LL method. (3) The performance of Q-matrix validation methods is affected by the sample size, and increasing the number of sample size can improve the accuracy of the Q-matrix validation.

In this study, Q-matrix validation methods for polytomous response CDMs were studied. It was found that the BIC method can be used for the Q-matrix validation under polytomous response CDMs. The method proposed in this paper can not only improve the accuracy of Q-matrix specification but also increase the model-data fit level. Besides, the data-based Q-matrix validation method can also reduce the workload of experts in Q-matrix construction and improve the classification accuracy of cognitive diagnosis.

Key words cognitive diagnostic assessment; Q-matrix; seq-GDINA; BIC