·统计、测量与方法·

一种广义的认知诊断 Q 矩阵修正新方法*

汪大勋 高旭亮 蔡 艳 涂冬波**
（江西师范大学心理健康教育研究中心，江西师范大学心理学院，南昌，330022）

摘 要 本文提出了一种新的 Q 矩阵修正方法——两阶段法（two–stage method），该方法不仅适用于简化的认知诊断模型，也适合于饱和的认知诊断模型，在实践应用中更具灵活性。模拟研究和实证研究表明：第一，两阶段方法整体上优于国际上知名的 ς^2 法 (de la Torre & Chiu, 2016)；第二，两阶段方法受被试人数和 Q 矩阵的错误率影响较小，尤其在小样本时仍有相对理想的正确率；第三，实证数据研究表明，两阶段法修正后的 Q 矩阵与数据拟合更好。

关键词 认知诊断 Q 矩阵 G–DINA 似然比 两阶段法

1 引言

与经典测量理论（CTT）和项目反应理论（IRT）相比，认知诊断评估（cognitive diagnosis assessment，CDA）可以提供更详细的诊断信息，因而受到越来越多研究者及实践者的关注 (Leighton & Gierl, 2007; Rupp, Templin, & Henson, 2010; Tatsuoka, 2009)。 在教育测评中，认知诊断可以诊断出学生对各个知识点的学习状态，从而为教师进行补救教学和因材施教提供重要依据 (Chang, 2015; Chen, 2017)。认知诊断主要包括“Q”矩阵和“诊断分类”两大部分（Tatsuoka，2009），认知诊断是在 Q 矩阵的基础上通过被试在题目上的作答信息来对被试进行诊断分类，因此 Q 矩阵是整个认知诊断的基础；同时，大量研究 (涂冬波，蔡艳，戴海崎，2012; de la Torre, 2009; Rupp & Templin, 2008) 表明，Q 矩阵错误界定会导致题目参数估计误差增大和被试诊断正确率降低，这进一步说明 Q 矩阵在认知诊断中具有重要的基础性作用，它的好坏直接决定认知诊断的效果。在实践中，Q 矩阵一般是由领域专家来进行标定，但专家标定 Q 矩阵会受到专家主观因素的影响，并且不同专家标定的 Q 矩阵也往往不尽相同。因此，为了克服专家标定 Q 矩阵的主观性以及提高 Q 矩阵标定的正确性，学者们提出了不同的方法进行 Q 矩阵估计或修正。在国内，研究者提出了 γ 法（涂冬波等，2012），D^2 统计量方法（喻晓锋等，2015），海明距离的方法（汪大勋，高旭亮，韩雨婷，涂冬波，2018）等；在国外，研究者们提出了 δ 法（de la Torre，2008）及在此基础上拓展的 ς^2 法 (de la Torre & Chiu, 2016)，RSS（residual sum of squares）法（Chiu，2013），基于残差的方法 (Chen, 2017)、数据驱动法（Liu, Xu, & Ying, 2012）、基于 EM 算法的方法（Wang et al., 2018）等。这些方法修正 Q 矩阵具有不同的特点，例如 D^2 统计量、δ 法、ς^2 法、基于 EM 算法的方法中的 MLE 和 MMLE 方法都是在参数估计以后使用特定的指标来评价题目的 Q 矩阵正确与否；γ 法和基于 EM 算法中的 ID 法则是在参数估计以后从经验的角度来分析题目的 Q 矩阵；而数据驱动法和基于残差的方法则是通过比较数据的观察分布和不同 Q 矩阵产生的期望分布之间的差异来进行 Q 矩阵修正。

但是这些国内外开发的方法大部分在实际使用中存在一些限制：如 γ 法、海明距离法、δ 法和 RSS 法等均有认知诊断模型（CDMs）限制，它们

** 本研究得到国家自然科学基金（31660278，31300876，31100756）、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。

** 通讯作者：涂冬波。E–mail: tudongbo@aliyun.com

DOI:10.16719/j.cnki.1671–6981.20190431

一般仅适用于完全非补偿的 DINA 模型或完全补偿的 DINO 模型中，而难于或无法应用于一些限制条件相对宽松的饱和认知诊断模型（saturated CDMs，如 G-DINA 模型等），从而限制了这些方法在实际中的应用范围。当这些方法中，仅有 ς^2 法和基于残差的方法可用于饱和的认知诊断模型，它们在样本容量较大时具有较理想的效果，但是当样本量减少时其效果如何，国内外还未见相关研究报道。我们通过先前的预研究发现，当样本量较少时（如 N=500），ς^2 法的 Q 矩阵修正的正确率会大大降低，也即它在样本量较少的情况下的表现很不理想。而开发出适用于小样本的 Q 矩阵修正方法不仅可以解决数据量有限时的 Q 矩阵修正，还有助于推动认知诊断在小规模测评中的应用。

鉴此，本研究拟在国内外以往研究的基础上，尝试开发出一种全新的既适用于简化模型又适用于饱和模型的 Q 矩阵修正方法，这类方法最大优点是不受所使用的认知诊断模型的限制，因而在实践中的使用范围更广，同时也避免了传统 Q 矩阵修正方法受模型限制的不足。具体来讲本研究开发新的 Q 矩阵修正方法需实现：（1）适用于各类认知诊断模型。即新开发的方法既适用于简化的认知诊断模型（reduced CDMs），又适用于饱和的认知诊断模型（saturated CDMs）；（2）在小样本容量下仍具有相对较理想的 Q 矩阵修正的正确率，以克服当前方法在小样本容量下准确率低的不足。文章拟采用 Monte Carlo 模拟研究和实证数据研究相结合的范式，探查新开发方法的效果，并将该方法的效果与国际知名的 ς^2 法进行比较研究，以探讨该新方法的科学性、有效性及其优势。

2 认知诊断模型

在认知诊断的发展中，目前国际上已经开发出多种认知诊断模型。如 DINA (Haertel, 1984;)、NIDA (Maris, 1999)、DINO (Templin & Henson, 2006)、R-RUM (Hartz & Roussos, 2008)、A-CDM 和 G-DINA (de la Torre, 2011) 等。其中前 5 种模型是简化的认知诊断模型（reduced CDMs），具有一定的约束条件。而 G-DINA 模型则是一种饱和模型，它在一定的条件下可以转换为其他简化模型（详见下面介绍）。这里对本文中涉及的认知诊断模型进行介绍如下：

对于题目 j，被试在该题上的答对概率受到题目 j 测量属性的影响。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，$L=2^{K^*_j}$ 为所有简化后的掌握模式。α^*_{lj}表示第 l 种掌握模式。对于掌握模式$\boldsymbol{\alpha}^*_{lj}=(\alpha_{l1},\cdots,\alpha_{lk},\cdots,\alpha_{lK^*_j})$，其在 G-DINA 模型下的答对概率表示为：

$$P(X_j=1|\alpha^*_{lj})=\delta_{j0}+\sum_{k=1}^{k^*_j}\delta_{jk}\alpha_{lk}+\sum_{k'=k+1}^{k^*_j}\sum_{k=1}^{k^*_j-1}\delta_{jkk'}\alpha_{lk}\alpha_{lk'}+\cdots+\delta_{j12\ldots k^*_j}\prod_{k=1}^{k^*_j}\alpha_{lk} \quad (1)$$

其中$P(X_j=1|\boldsymbol{\alpha}^*_{lj})$是掌握模式为$\boldsymbol{\alpha}^*_{lj}$ 的被试答对题目 j 的概率。δ_{j0} 是题目 j 的截距参数，δ_{jk} 是属性 k 的主效应，$\delta_{jkk'}$ 是属性 k 和 k' 的交互效应，$\delta_{j12\ldots k^*_j}$ 是所有属性的交互效应。

当假设属性之间没有交互效应时，G-DINA 模型可以简化为 A-CDM 模型。A-CDM 模型的公式如下：

$$P(X_j=1|\alpha^*_{lj})=\delta_{j0}+\sum_{k=1}^{k^*_j}\delta_{jk}\alpha_{lk} \quad (2)$$

R-RUM 模型与 A-CDM 模型相似，同样没有交互效应。R-RUM 模型取了对数以后则与 A-CDM 模型相同，R-RUM 模型表达式如下：

$$\log\left[P(X_j=1|\alpha^*_{lj})\right]=\delta_{j0}+\sum_{k=1}^{k^*_j}\delta_{jk}\alpha_{lk} \quad (3)$$

将除了 δ_{j0} 和$\delta_{j12\ldots k^*_j}$参数以外的所有参数设置为 0，则 G-DINA 模型可以简化为 DINA 模型。DINA 模型的公式如下：

$$P(X_j=1|\alpha^*_{lj})=\delta_{j0}+\delta_{j12\ldots k^*_j}\prod_{k=1}^{k^*_j}\alpha_{lk} \quad (4)$$

设置 $\delta_{jk}=-\delta_{jkk'}=\cdots=(-1)^{K^*_j+1}\delta_{j12\ldots K^*_j}$，则 G-DINA 模型可以简化为 DINO 模型，DINO 模型的公式如下：

$$P(X_j=1|\alpha^*_{lj})=\delta_{j0}+\delta_{jk}\alpha_{lk}. \quad (5)$$

3 Q 矩阵修正新方法：两阶段法

本研究提出的 Q 矩阵修正方法分为两个阶段，因此命名为两阶段方法 (two-stage method，TS)。两阶段法在进行 Q 矩阵修正时，是对题目依次循环进行修正，即每次只考察 / 修正一道题目的 q 向量。在修正题目 j 的 q 向量时，其余 J-1 个题目的 Q 矩阵保持不变，具体分为两个阶段：

（1）Stage 1：q_j^{Stage1} 属性的确定

分别确定每个属性是否单独对被试答对项目 j 的概率有显著影响，也即分别对每个属性的主效应进行显著性检验（H0：$\delta_{jk}=0$），我们把所有主效应显著的属性暂定为项目 j 的测量属性，记为 q_j^{Stage1}。若 q_j^{Stage1}=(10100)，则说明属性 1 和属性 3 在项目 j 上的主效应显著，其余属性的主效应不显著。q_j^{Stage1} 的确定过程为：

分别将单个属性的测量模式（如 [10000]）作为题目 j 的测量模式，与其他题目 Q 阵一起估计题目的参数。对于题目 j，则可以估计出属性 k 在题目 j 上的主效应参数 $\hat{\delta}_{jk}$ 及其误差 $SE_{\delta_{jk}}$。理论上，如果题目 j 测量了属性 k，则属性 k 的主效应 $\hat{\delta}_{jk}$ 应该不等于 0，因此可对这一零假设 H0：$\delta_{jk}=0$ 进行 t 检验，即：

$$t=\frac{\hat{\delta}_{jk}-0}{SE_{\delta_{jk}}} \tag{6}$$

如果 t 检验显著（p<.01 或 .05），则说明属性 k 在题目 j 上的主效应显著，即题目 j 可能测量了属性 k。对于题目 j，需要分别对 K 个单属性的测量模式进行估计，并进行 t 检验。将所有主效应显著的属性的集合定义为 C_j，则题目 j 测量的属性应该包含于集合 C_j。

（2）Stage 2：q_j^{Stage1} 属性的进一步验证

在 Stage 1 中，由于每个属性的显著性检验是单独进行的，因此无法考察多个属性间的相互影响。也即属性 k 单独检验时，其主效应显著，但当增加另一个属性 s，则属性 k 的主效应可能不显著。因此在 Stage 2 中需要进一步检验 q_j^{Stage1} 中的属性。

在集合 C_j 中，题目 j 应该更倾向于测量了 t 检验中 p 值更小的属性。因此，在 Stage 2 中，在 p 值最小的属性基础上，根据 p 值从小到大从 C_j 中依次增加属性，并检验新增加属性后的模型拟合度是否显著优于未增加属性的模型拟合度。本研究中采用似然比（Likelihood ratio test，LRT）检验，具体如下：

由于题目 j 至少需要考察一个属性，因此将 p 值最小的属性作为题目 j 的基础属性，并在此基础上根据 p 值大小从集合 C_j 中依次增加一个属性，将增加属性之前的 q 向量表示为 q_{before}，增加属性之后的 q 向量表示为 q_{after}。分别将增加属性前后的测量模式作为题目 j 的测量模式，并将两种测量模式下得到的两个模型似然进行似然比检验。两种测量模式下的似然比（表示为 D）检验表示为：

$$D=-2\ln\left(\frac{L\left(X\middle|\hat{\beta},Q^T_{j_{q_{before}}}\right)}{L\left(X\middle|\hat{\beta},Q^T_{j_{q_{after}}}\right)}\right) \tag{7}$$

其中 $L\left(X\middle|\hat{\beta},Q^T_{j_{q_{before}}}\right)$ 和 $L\left(X\middle|\hat{\beta},Q^T_{j_{q_{after}}}\right)$ 分别是增加属性前后模型的似然。$Q^T_{j_{q_{before}}}$ 是当题目 j 的测量模式为 q_{before} 时整个测验的 Q 矩阵，$Q^T_{j_{q_{after}}}$ 是当题目 j 的测量模式为 q_{after} 时整个测验的 Q 矩阵。

似然比统计量（D 值）服从自由度为模型参数个数之差的卡方分布，

$$D\sim\chi^2, df=\left(df_{Q^T_{j_{q_{after}}}}-df_{Q^T_{j_{q_{before}}}}\right) \tag{8}$$

如果属性增加前后模型之间没有显著差异，则根据简单性原则，不增加该属性，则题目 j 的测量模式仍为 q_{before}。如果属性增加前后模型之间有显著差异，则增加该属性，并将增加属性后的测量模式 q_{after} 作为基础测量模式。接着再从集 C_j 中增加下一个属性并进行似然比检验，直到增加属性前后模型之间没有显著差异或者增加完集合 C_j 中的全部属性为止，由此得到题目 j 的测量模式。

如果两阶段法得到题目 j 的测量模式与原有 Q 矩阵中的测量模式不同，则更新题目 j，然后修正下一题。当所有题目均被修正，记为一次迭代。当修正前后的 Q 矩阵相同或迭代次数达到事先界定的最大次数，则算法结束。

为了探讨本文提出的两阶段法的科学性和有效性，本研究采用 Mont Carlo 模拟研究和实证研究相结合的研究范式。研究 1 比较了两阶段法和 ς^2 法（de la Torre & Chiu, 2016）在四个简化模型中的效果；研究 2 比较了两阶段法和 ς^2 法在饱和模型中的效果；研究 3 对两种方法在实证数据中的表现进行了验证及比较。

4 研究 1 两阶段法及 ς^2 法在四个简化模型中的比较研究

4.1 研究 1 设计

4.1.1 Q 矩阵

本研究采用 de la Torre 和 Chiu（2016）的 Q 矩阵（见表 1），包含 30 个题目，5 个属性。

4.1.2 认知诊断模型、被试参数和题目参数模拟

表 1 测验 Q 矩阵

Item	A1	A2	A3	A4	A5	Item	A1	A2	A3	A4	A5
1	1	0	0	0	0	16	0	1	0	1	0
2	0	1	0	0	0	17	0	1	0	0	1
3	0	0	1	0	0	18	0	0	1	1	0
4	0	0	0	1	0	19	0	0	1	0	1
5	0	0	0	0	1	20	0	0	0	1	1
6	1	0	0	0	0	21	1	1	1	0	0
7	0	1	0	0	0	22	1	1	0	1	0
8	0	0	1	0	0	23	1	1	0	0	1
9	0	0	0	1	0	24	1	0	1	1	0
10	0	0	0	0	1	25	1	0	1	0	1
11	1	1	0	0	0	26	1	0	0	1	1
12	1	0	1	0	0	27	0	1	1	1	0
13	1	0	0	1	0	28	0	1	1	0	1
14	1	0	0	0	1	29	0	1	0	1	1
15	0	1	1	0	0	30	0	0	1	1	1

研究 1 的认知诊断模型为四个简化的认知诊断模型，即 DINA、DINO、A-CDM 和 R-RUM。被试掌握模式按照均匀分布从 2^5=32 种模式中随机产生，分别产生 500、1000、2000 人。所有模型题目参数的模拟借鉴 de la Torre 和 Chiu（2016）的模拟方法：掌握项目 j 全部属性的被试和没有掌握任何项目 j 属性的被试（记为 P_{j1} 和 P_{j0}）的答对概率分别固定为 .8 和 .2。对 DINA 和 DINO 模型来说，猜测参数和失误参数均为 .2。对于 R-RUM 和 A-CDM 模型，其他掌握模式的答对概率从 $[P_{j0}, P_{j1}]$ 中随机产生且服从单调性约束。

4.1.3 Q 矩阵错误模拟

参考 de la Torre 和 Chiu（2016）的设计，Q 矩阵错误率设置为 5%。为了考察两阶段法在 Q 矩阵错误率高的情况下的表现，增加 10% 的错误率。即 Q 矩阵分别随机产生 5% 和 10% 的错误。

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 上的答对概率 P_{ij}，以 P_{ij} 为概率在贝努力分布中产生被试 i 在题目 j 上的 0~1 作答反应得分 *response (i, j)*，即 *response(i, j)= Bernoulli(P_{ij})*。

4.1.5 评价指标

计算每次修正后的 Q 矩阵与真实 Q 矩阵题目测量模式的一致性作为题目模式判准率（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正确修改的比例和正确界定的属性未被修改的比例。所有实验均重复 100 次，然后再计算 100 次实验的平均 PMR、AMR、FPR 以及 TPR。

$$PMR = \frac{\sum_{t=1}^{100}\sum_{j=1}^{J} n_{jt\_correct}}{100 \times J} \quad (9)$$

$$AMR = \frac{\sum_{t=1}^{100}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{jkt\_correct}}{100 \times J \times K} \quad (10)$$

公式 9 和 10 中，t 为第 t 次实验（t=1,2,⋯,100），J 为题目个数，$n_{jt\_correct}$ 为第 t 次实验中修正后的第 j 题 q 向量是否与真实 Q 矩阵中第 j 题一致，完全一致则为 1，否则为 0。公式（10）中，K 为属性个数，$n_{jt\_correct}$ 表示第 t 次实验中修正后的第 j 题的第 k 个属性（为 0 或者 1）是否与真实 Q 矩阵中第 j 题第 K 个属性一致，如果一致则为 1，否则为 0。

4.2 研究 1 结果

表 2 呈现了两种方法在 DINA 和 DINO 模型中的修正结果。表 3 呈现了两种方法在 A-CDM 和 R-RUM 模型中的修正结果。

由表 2 的结果可知，在 DINA 和 DINO 模型中，TS 法对 Q 矩阵修正的正确率总体上要高于 ς^2 法。相对于 ς^2 法，TS 法表现比较稳健。即使在错误率为 10%，人数在 500 时，TS 法的模式判准率和属性判准率也达到 95% 和 99% 左右。而 ς^2 法在人数少且错误率高的情况下的模式判准率不到 30%，非常

不理想。

对于样本容量影响，相比之下 ς^2 法更受被试人数的影响，当人数从 2000 降到 500 人时，ς^2 法的模式判准率下降了 60%~70%，属性判准率下降了 20% 左右。TS 法的模式判准率下降 5% 左右，属性判准率下降 1% 左右，远低于 ς^2 法的下降幅度。在不同的错误率上，Q 矩阵的错误率越高则正确恢复 Q 矩阵的难度越大。两种方法受 Q 矩阵错误率的影响不同，TS 法在 5% 和 10% 的错误率时的表现几乎一致，因此该方法受错误率的影响较小。当人数达到 2000 时，TS 法在两种错误率下对 Q 矩阵的修正正确率均能达到 100%。而相对于 TS 法，ς^2 法受 Q 矩阵错误率的影响更大。

在 FPR 和 TPR 指标上，TS 法均优于 ς^2 法。但两种方法在 FPR 指标上的差异更大，说明两种方法的区别在于错误属性的识别上。对于错误属性的识别，TS 法比 ς^2 法更加精确。

根据表 3 的结果，在 A-CDM 和 R-RUM 模型中，TS 法的表现总体上依旧好于 ς^2 法。特别是在小样本时，TS 法和 ς^2 法的差异更大，但当人数增加到 2000 时，两种方法的 Q 矩阵修正正确率相当。通过比较 FPR 指标和 TPR 指标，两种方法在 FPR 指标上的差异更大，同样也说明在错误属性的识别和修正上，ς^2 法的表现不如 TS 法。此外，两种方法在 A-CDM 模型和 R-RUM 模型中的表现不如 DINA 和 DINO 模型，原因是前面两个模型明显比后面两个模型更为复杂。

表 2 两种方法在 DINA 和 DINO 模型中 100 次实验的平均结果

模型	错误率	N	PMR		AMR		FPR		TPR	
			ς^2	TS	ς^2	TS	ς^2	TS	ς^2	TS
DINA	5%	500	.272	.948	.764	.988	.511	.989	.778	.988
		1000	.766	.999	.943	1	.765	.999	.953	1
		2000	.941	1	.986	1	.914	1	.990	1
	10%	500	.234	.942	.741	.987	.491	.991	.769	.986
		1000	.683	.998	.919	1	.709	.999	.942	1
		2000	.895	1	.975	1	.887	1	.984	1
DINO	5%	500	.268	.951	.762	.990	.486	.990	.778	.990
		1000	.772	.999	.944	1	.766	1	.955	1
		2000	.944	1	.987	1	.895	1	.993	1
	10%	500	.248	.945	.746	.987	.482	.986	.775	.988
		1000	.683	.999	.916	1	.709	1	.939	1
		2000	.885	1	.971	1	.877	1	.981	1

表 3 两种方法在 A-CDM 和 R-RUM 模型中 100 次实验的平均结果

模型	错误率	*N*	PMR		AMR		FPR		TPR	
			ς^2	TS	ς^2	TS	ς^2	TS	ς^2	TS
A-CDM	5%	500	.273	.530	.755	.876	.475	.883	.771	.875
		1000	.589	.636	.895	.909	.631	.896	.910	.909
		2000	.692	.699	.924	.927	.766	.911	.933	.928
	10%	500	.253	.511	.737	.867	.447	.858	.770	.868
		1000	.561	.614	.882	.900	.626	.903	.910	.899
		2000	.662	.672	.915	.921	.753	.908	.933	.922
R-RUM	5%	500	.276	.551	.754	.884	.503	.863	.768	.885
		1000	.609	.662	.901	.916	.648	.905	.915	.917
		2000	.712	.717	.930	.933	.799	.930	.938	.934
	10%	500	.240	.528	.736	.876	.449	.875	.768	.876
		1000	.557	.627	.882	.906	.623	.906	.911	.906
		2000	.697	.719	.924	.932	.807	.928	.937	.932

5 研究 2 两阶段法及 ς^2 法在饱和模型（G-DINA）中的比较研究

研究 1 中的认知诊断模型可以由饱和的认知诊断模型在一定的约束条件下转换而来，因此饱和模型具有更广的适用性。研究 2 则是对两种方法在饱和模型（G-DINA）中的效果进行验证及比较。

5.1 研究 2 设计

研究 2 的实验设计与研究 1 的实验设计相似，不同的是研究 2 模拟数据时使用的是饱和的认知诊断模型（G-DINA）。

5.2 研究 2 结果

表 4 呈现了两种方法在 G-DINA 模型中的修正结果。从表 4 可以看出，在 G-DINA 模型中，TS 法的表现依旧优于 ς^2 法。当被试人数减少，两种方法的差异更大，AMR 指标差异达到 20% 左右，PMR 指标差异达到 40% 左右。当被试人数增加到 2000 人时，两种方法的模式判准率和属性判准率相当。在错误率的影响上，ς^2 法受错误率的影响略微大于 TS 法。如当 N=2000 时，错误率从 5% 增加到 10%，ς^2 法的模式判准率降低幅度为 4%，TS 法降低幅度为 1%。此外，TS 法的 FPR 指标在所有实验条件下均高于 ς^2 法，并且随着人数减少，两种方法的 FPR 指标差异越大。这说明对错误属性的识别上，TS 法要比 ς^2 法更精确。TS 法的 TPR 指标总体上优于 ς^2 法，当被试人数达到 2000 时，ς^2 法的 TPR 指标略优于 TS 法。说明只有当被试人数达到一定的程度时，ς^2 法才能有效识别正确标定的属性。

表 4 两种方法在 G-DINA 模型中 100 次的平均结果

模型	错误率	N	PMR		AMR		FPR		TPR	
			ς^2	TS	ς^2	TS	ς^2	TS	ς^2	TS
G-DINA	5%	500	.278	.664	.758	.922	.531	.921	.770	.922
		1000	.694	.802	.926	.958	.695	.951	.939	.958
		2000	.884	.896	.976	.979	.854	.980	.983	.978
	10%	500	.241	.635	.739	.912	.478	.905	.768	.913
		1000	.617	.796	.901	.955	.627	.953	.932	.956
		2000	.844	.885	.965	.976	.822	.979	.981	.976

6 研究 3：实证数据研究

为了验证两阶段方法在实证数据中的效果及与 ς^2 法进行实证比较，本研究采用了 TIMSS（Trends in International Mathematics and Science Study，2003）的数据。该项目是由国际教育成就评价协会实施，Su 等人（Su, Choi, Lee, Choi, & McAninch, 2013）将 TIMSS 数据用于认知诊断分析。本研究中 TIMSS 数据的 Q 矩阵如表 5。

分析之前先计算 TIMSS 数据与各模型之间的拟合指标（偏差、AIC、BIC），结果显示各指标对应

表 5 TIMSS 数据 Q 矩阵

Item	Code	A1	A2	A3	A4	A5
1	M012002	0	0	0	1	0
2	M012016	0	1	0	0	1
3	M012042	0	0	1	0	1
4	M022050	0	0	0	1	0
5	M022185	0	0	1	0	0
6	M022191	0	0	0	1	0
7	M022196	0	0	1	0	0
8	M022198	0	1	0	0	0
9	M022232	1	0	0	0	0
10	M022251	0	0	1	0	0
11	M032570	1	0	0	0	0
12	M032643	0	1	0	0	1

注：A1，使用比率和速率推理来解决实际或数学问题；A2，将原来对数学的理解应用到有理数系统中；A3，将原来对算术的理解应用到代数表达式中并加以扩展；A4，单变量方程和不等式的推理与解答；A5，将原来对分数运算的理解拓展并应用到有理数的加、减、乘、除中。

的最优模型并不相同，为了避免模型选择错误，且与 ς^2 法保持一致，因此这里使用 GDINA 模型来进行分析。分别使用两阶段法和 ς^2 法对 TIMSS 数据进行分析，同时保证每个属性均被测量一次以上。两种方法对原始 Q 矩阵的修改结果如下：两阶段法调整了 4 个题目，共 7 个属性 (Item3[A1:0-1;A3:1-0]、Item4[A1:0-1;A4:1-0]、Item5[A2:0-1;A3:1-0]、Item12[A5:1-0])；ς^2 法没有修改任何属性，因此与原始 Q 矩阵相同。两阶段法对各个题目修正前后模型偏差的变化如表 6。从表 6 中可以看出，所有题目在建议的测量模式上增加属性均会导致模型偏差的降低。而对于题目 3，原始 Q 矩阵中的测量模式定义为 [00101]，两阶段法结果显示题目 3 的 α_3 的主效应也显著，并且根据 p 值的排序靠前，而似然比检验结果未将其作为题目 3 的属性。题目 4 也发生了相同的情况，其原始 Q 矩阵的测量模式为 [00010]，根据两阶段法的结果，题目 4 的 α_4 主效应也显著，且排序靠前，而似然比检验结果将其排除。这两个题目均将 α_1 作为了测量属性，但从题目 3（已知 x，求 -3x 的值）与题目 4（求 x/3>8 的 x 值）来看，均与 α_1 关系不明显。而同样测量了 α_1 的题目 9 和题目 11 却没有被修改，这不排除是随机因素的影响，此外这里需要学科专家来决定题目 3 是否需要删除 α_3、题目 4 是否需要删除 α_4。题目 12 删除了 α_5 只保留了 α_2，且具有很高的模型偏差，这可能是两个属性之间的交互作用导致了 α_5 的作用减小。从题目上来看，第 12 题（n 是否为负整数）与同样测量了 α_5 的第 2 题（包含 2.25 的数字区间）相比，题目测量的内容并不完全相同，因此更深层的原因需要由学科专家进行解释。同样对于题目 5 调整的两个属性，也需要专家判断其与题目测量内容是否匹配。

为了比较原有方法修正后的 Q 矩阵（与原始 Q 矩阵一致）与两阶段法修正后的 Q 矩阵的差异，分别计算不同 Q 矩阵下的模型拟合指标：偏差、AIC 和 BIC。结果如表 7。从表 7 可以看出，经过两阶段法调整后的 Q 矩阵在模型拟合上要优于 ς^2 法修正后的 Q 矩阵（与原始 Q 矩阵一致）。因此实证数据分析的结果显示，两阶段法修正后的 Q 矩阵与数据更拟合。

7 结论与讨论

7.1 结论

（1）本研究开发的两阶段方法整体上优于国际上知名的 ς^2 法。与 ς^2 法相比，两阶段法修正 Q 矩阵的效果更好，且更加稳健；在识别 Q 矩阵错误上，两阶段法比 ς^2 法更敏感；更为重要的是，在所有模拟实验条件下，两阶段法对 Q 矩阵修正的准确率（PMR）比 ς^2 法平均整体高出约 22%，说明本研究开发的新方法具有非常明显的优势。

（2）本研究开发的两阶段方法不仅具有较理想的 Q 矩阵修正正确率，并且该方法受被试人数和 Q 矩阵的错误率影响较小，尤其在小样本时仍有相对理想的正确率。

（3）两阶段法既适用于简化认知诊断模型，也适

表 6 两阶段法修改题目前后模型偏差变化

题目	测量模式	偏差（-2*LL）	D 值	自由度（df）
item3	10000	10736.52		
	10001Δ	10766.56	30.04	2
	10101	10735.22	-31.34	4
item4	10000Δ	10724.11		
	10010	10719.46	-4.65	2
item5	01000Δ	10718.29		
	01001	10711.94	-6.35	2
item12	01000Δ	10741.18		
	01010	10738.26	-2.92	2

备注：“Δ”表示该测量模式为建议的测量模式。

表 7 不同 Q 矩阵的模型拟合指标

Q-matrix	模型拟合		
	偏差	AIC	BIC
Q_original / Q_ς^2	10732.62	10853.85	11136.24
Q_TS	10741.18	10859.19	11132.32

合于饱和的认知诊断模型，在实践应用中更具灵活性。

7.2 讨论

（1）两阶段法的优点和缺点

与已有方法 ς^2 法的比较研究显示，两阶段法比 ς^2 法具有更高的正确率。此外，两阶段法受被试人数和 Q 矩阵错误率的影响较小，在小样本和 Q 矩阵错误率高的情况下，两阶段法表现依旧很稳健。但该方法修正 Q 矩阵时分为两个阶段，并且相对来说计算复杂。与 ς^2 法相比，该方法对 Q 矩阵的修正一般需要迭代 2~3 次，为了防止循环迭代，还需要设置一个迭代终止规则。在运算时间上，ς^2 法的平均运行时间为 3.47 秒，两阶段法的平均运行时间为 1233.44 秒，因此两阶段法要比 ς^2 法更耗时。

（2）两阶段法可以进一步研究的地方。

两阶段法在第一步是将单个属性的测量模式作为题目 j 的测量模式，并估计主效应 $\hat{\delta}_{jk}$，因此属性间主效应不能直接比较大小。可替代的方法是将考察了所有属性的测量模式（如 q = (11111) ）作为题目 j 的测量模式，由此估计各属性的主效应，并根据主效应大小排序进行第二步检验。但这样的做法还需要考虑由于属性间交互作用对属性主效应大小的影响。

此外，在两阶段法的第二步也可以用其他方法（ 如 Wald 检验（Ma, Iaconangelo, & de la Torre, 2016））来检验题目的属性是否多余。与似然比检验不同，Wald 检验是从题目的层面对各个属性进行检验，并且在确定下一个属性时，需要对所有剩余的属性都进行检验。而使用主效应大小排序以及使用其他检验方法来修正 Q 矩阵的表现如何还有待进一步深入研究。

（3）关于有争议题目的处理

对于Q 矩阵的修正，研究者提出了许多种方法。使用不同方法修正的 Q 矩阵可能不相同，不同的 Q 矩阵修正方法可以用来相互验证 Q 矩阵的修正结果。对于有争议的题目，需要交由专家进行讨论后再决定。这样既避免了由于专家的主观性导致的 Q 矩阵错误，也可以避免客观方法的局限性。通过客观方法对 Q 矩阵进行修正，在一定程度上也减轻了专家标定 Q 矩阵的负担。但是客观方法修正 Q 矩阵不能取代专家在 Q 矩阵标定和测验设计中的作用，客观方法只是为专家标定 Q 矩阵提供参考和依据。

参考文献

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.

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.

喻晓锋 , 罗照盛 , 高椿雷 , 李喻骏 , 王睿 , 王钰彤 . (2015). 使用似然比 D2 统计量的题目属性定义方法 . *心理学报 , 47(3),* 417–426.

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

Chen, J. S. (2017). A residual–based approach to validate Q–matrix specifications. *Applied Psychological Measurement, 41(4),* 277–293.

Chiu, C. Y. (2013). Statistical refinement of the Q–matrix in cognitive diagnosis. *Applied Psychological Measurement, 37(8),* 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(4),* 343–362.

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

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

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q–matrix validation. *Psychometrika, 81(2),* 253–273.

Haertel, E. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement, 8(3)*, 333–346.

Hartz, S., & Roussos, L. (2008). *The Fusion Model for skills diagnosis: Blending theory with practice*. Princeton, NJ: Educational Testing Service.

Leighton, J. P., & Gierl, M. J. (Eds). (2007). *Cognitive diagnostic assessment for education: Theory and applications.* New York, NY: Cambridge University Press.

Liu, J. C., Xu, G. J., & Ying, Z. L. (2012). Data–driven learning of Q–matrix. *Applied Psychological Measurement, 36(7)*, 548–564.

Ma, W. C., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40(3),* 200–217.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64(2),* 187–212.

Rupp, A. A., & Templin, J. (2008). The effects of Q–matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68(1),* 78–96.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* New York, NY: Guilford Press.

Su, Y. L., Choi, K. M., Lee, W. C., Choi, T., & McAninch, M. (2013). *Hierarchical cognitive diagnostic analysis for TIMSS 2003 mathematics.* CASMA Research Report 35. Iowa: Center for Advanced Studies in Measurement and Assessment (CASMA), University of Iowa.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method.* New York, NY: Taylor & Francis.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11(3),* 287–305.

Wang, W. Y., Song, L. H., Ding, S. L., Meng, Y. R., Cao, C. X., & Jie, Y. J. (2018). An EM–based method for Q–matrix validation. *Applied Psychological Measurement, 42(6),* 446–459.

A New General Method for Q-Matrix Validation in Cognitive Diagnosis Assessments

Wang Daxun, Gao Xuliang, Cai Yan, Tu Dongbo

(Research Center of Psychological Health Education, School of Psychology, Jiangxi Normal University, Nanchang, 330022)

Abstract As the basis of CDAs, Q-matrix has aroused more and more attentions for the subjective tendency of Q-matrix construction that is typically performed by domain experts. Due to the subjective process of Q-matrix construction, there inevitably have more or less misspecifications in the Q-matrix, if left unchecked, can result in serious negative impact on CDAs. To avoid the subjective tendency from experts and to improve the correctness of Q-matrix specification, a number of objective methods have been proposed. However, most of the existing methods have some limitations. For example, some methods can only be used for reduced cognitive diagnostic models, and the methods that can be used for saturated models require large sample sizes(N ⩾ 2000). Therefore, it is necessary to propose a Q-matrix validation method which can be used in a wide class of cognitive diagnosis models and has acceptable performance even when the simple size is small. To address this concern, this paper proposed a new method to validate Q-matrix-Two-stage method.

The method proposed in this paper has two stages in the process of Q-matrix validation. The purpose of the first stage is to estimate the parameters of item *j* by taking all the single attribute measurement patterns as the q-vector of the item *j*. Then T test was used to check whether there was a significant difference between the $\hat{\delta}_{jk}$ (the parameter of item *j*) and 0. All attributes with significant difference from 0 are selected as candidate attributes for item *j*. The purpose of the second stage is to further verify candidate attributes. For item *j*, the number of attributes increased from 1 to K, and only one attribute is added at a time. The likelihood ratio is used to test whether there is significant difference between the two models before and after adding attribute to item *j*. Finally, the q-vector of item *j* can be determined according to the result of likelihood ratio test.

Both simulation and real-life studies were conducted to examine the feasibility and effectiveness of the proposed method. Results indicate that the proposed method outperforms the existing methods introduced in this paper (ς^2 method) whatever the reduced or saturated CDMs are used. Moreover, the proposed method has acceptable performance even when the simple size is small but the ς^2 method has not. Besides, for the identification of misspecification in Q-matrix, the two-stage method is more sensitive than the ς 2method. A real-life example also shows that the Q-matrix adjusted by two-stage method has better goodness of fit and reliability, and is more reasonable than the original Q-matrix and the Q-matrix suggested by ς^2 method.

In this paper, a new general method for Q-matrix validation was developed. This method can be used to construct the Q-matrix in the cognitive diagnosis assessments and to adjust misspecification in Q-matrix. In addition, the proposed method is helpful to improve the accuracy of classification in cognitive diagnosis and promote the development of cognitive diagnosis theory.

Key words cognitive diagnostic assessment, Q-matrix, G-DINA, Likelihood ratio test, two-stage method