

# 一种非参数化的 Q 矩阵估计方法：ICC-IR 方法开发\*

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** 相对于参数化的方法, 本研究根据题目测量模式关系开发出 ICC 指标, 并提出基于理想得分的 ICC 指标法进行 Q 矩阵估计。Monte Carlo 模拟研究与实证研究发现 (1) 基于理想得分 ICC 指标法估计 Q 矩阵具有很好的效果, 当属性个数越少、基础题个数越多, 估计效果越好。(2) 相对于以往方法—— $D^2$  统计量的方法, ICC-IR 法效果更好, 并且是一种非参数化的方法, 计算简单快捷。(3) 实证数据分析表明, ICC-IR 法估计的 Q 矩阵在模型拟合度上也优于  $D^2$  统计量方法。

**关键词** 认知诊断 Q 矩阵 ICC 指标 DINA 模型

## 1 引言

认知诊断包括两大部分：“Q 矩阵界定”和“诊断分类” (Tatsuoka, 2009)。Q 矩阵表述了题目与属性之间的关系, 是认知诊断的基础 (Leighton, Gierl & Hunka, 2004)。在以往研究中, 我们通常是假设所界定的测验 Q 矩阵是正确的, 并由此对被试进行诊断分类, 然而 Q 矩阵的界定并不容易。在实际运用中, 通常采用的方法是让多位专家对题目的属性进行标定来确定测验的 Q 矩阵, 而专家们所定义的 Q 矩阵不尽相同。已有研究发现 Q 矩阵错误会增大参数估计误差和降低被试诊断正确率 (涂冬波, 蔡艳, 戴海崎, 2012; Rupp & Templi, 2008; de la Torre, 2009)。因此 Q 矩阵界定的困难一定程度上限制了认知诊断在实际中的应用 (DeCarlo, 2011)。此外, 研究者还尝试利用作答数据进行 Q 矩阵估计或修正。在 Q 矩阵估计上, 汪文义等人 (汪文义, 丁树良, 2010; 汪文义, 丁树良, 游晓锋, 2011) 及陈平和辛涛 (陈平, 辛涛, 2011) 分别对题目属性向量估计进行了研究。喻晓锋等人 (2015) 提出使用似然比  $D^2$  统计量来进行 Q 矩阵估计。在 Q 矩阵修正方面, 研究者提出了如  $\delta$  法 (de la Torre, 2008),  $\gamma$  法 (涂冬波等, 2012), RSS (residual sum of squares) 法 (Chiu, 2013) 等方法。已有的 Q 矩阵估计和修正方法中, 除 RSS 法以外, 均需要进行参数估计, 属于参数化的方法。然而这类参数化

的方法一般运算量大, 花费时间较长。同时 Q 矩阵估计的效果还有提升的空间, 如似然比  $D^2$  统计量在属性个数为 4 个或 5 个时, 在 8 个基础题的情况下, 平均估计正确率在 50% 左右。因此开发一种计算简单、效果较好的 Q 矩阵估计方法对认知诊断的简单化和推广具有重要意义。

本研究受 HCI 指标的启发, 开发出 ICC 指标 (详见文章第 2 部分), 提出基于理想作答的 ICC 指标法用于 Q 矩阵估计。并通过 Monte Carlo 模拟研究和实测数据研究, 探查该方法进行 Q 矩阵估计时的效果, 为认知诊断测验 Q 矩阵的估计提供新的方法支持。

## 2 基于理想得分的 ICC 指标 (ICC-IR) 法开发

### 2.1 HCI 指标

HCI (hierarchy consistency index) 指标是由 Cui, Leighton, Gierl 和 Hunka (2006) 开发的属性层级一致性指标, 用于属性层级关系合理性的界定。公式如下:

$$HCI_i = 1 - \frac{2 \sum_{j \in \text{Correct}_i} \sum_{g \in S_j} X_{ij}(1 - X_{ig})}{N_{ci}} \quad (1)$$

上式中,  $\text{Correct}_i$  是指所有被试  $i$  正确作答项目的集合。  $X_{ij}$  是被试  $i$  在第  $j$  题上的得分,  $S_j$  是指所测属性是项目  $j$  测量属性的子集的项目集,  $X_{ig}$  是指

\* 本研究得到国家自然科学基金 (31660278, 31760288, 31300876, 31100756)、江西省高等院校教学改革研究课题 (JXJG-15-2-26)、江西省高校人文社科项目 (XL1507, XL1508)、江西省社会科学规划项目 (17JY12)、江西省教育厅人文社科重点项目 (JD17077) 和武汉市卫计委支撑课题 (WG16C08) 的资助。

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

DOI:10.16719/j.cnki.1671-6981.20180233

被试在项目集  $S_j$  中题目上的得分,  $N_{ci}$  是所有答对题目的总共比较次数。当被试  $i$  在第  $j$  题上得 1 分, 而在第  $g$  题 ( $g \in S_j$ ) 上得 0 分, 则异常次数加 1 次。通过计算所有被试 HCI 指标的平均值可以界定属性层级关系的合理性。

## 2.2 ICC 指标开发

受 HCI 指标启发, 本文提出项目一致性指标 (item consistency criterion, ICC), 用于考核题目 Q 阵的合理性。在该指标上, 可以考虑以下几种项目 Q 阵异常情况: 如果被试在第  $j$  题上答对, 而在测量属性为第  $j$  题子集题目上答错; 被试在第  $j$  上答错, 而在测量属性为第  $j$  题父集题目上答对; 被试在第  $j$  上答对 / 错, 而在测量属性与第  $j$  题测量模式相同的题目上答错 / 对。ICC 指标计算公式为:

$$ICC_{qmj} = 1 - \frac{2 \sum_i \left[ \sum_{g \in S_m} X_{jg}(1-X_{gi}) + \sum_{j \in S_m^*} X_{jg}(1-X_{gi}) + \sum_{h \in S_m^{**}} [X_{jh}(1-X_{hi}) + X_{hi}(1-X_{jh})] \right]}{N_{cmj}} \quad (2)$$

上式中,  $ICC_{qmj}$  是第  $j$  题的第  $m$  种测量模式的指标,  $m$  是所有可能的测量模式的一种,  $X_{jg}$  是被试在第  $j$  题上的得分,  $S_m$  表示测量属性为题目  $j$  的子集项目集合,  $X_{g}$  为被试在测量属性为  $j$  题的子集题目上的得分。  $S_m^*$  表示项目  $j$  的父集项目集合,  $X_{g}$  为被试在  $j$  题的父集题目上的得分,  $S_m^{**}$  为测量模式和项目  $j$  相同的项目集合。  $X_{h}$  为被试在与  $j$  题有相同测量模式题目上的得分,  $N_{cmj}$  为总共的比较次数。由上式可见 ICC 指标越大代表异常越少。根据公式 (2), 当已知某些题目的测量模式时, 通过依次比较所有可能测量模式与已知题目测量模式的关系, 就可以计算“新题”(即未定义测量模式的题目)所有可能测量模式的 ICC 指标, 选择 ICC 指标较好的测量模式作为“新题”的测量模式, 由此进行 Q 矩阵估计。

## 2.3 基于原始得分的 ICC 指标与基于理想得分的 ICC 指标

在计算不同测量模式的 ICC 指标时, 可以使用两种方法, 分别是基于原始得分的 ICC 指标 (ICC based on observed response, ICC-OR) 和基于理想得分的 ICC 指标 (ICC based on ideal response, ICC-IR)。基于原始得分的 ICC 指标, 是在计算 ICC 指标时直接使用原始得分数据; 而 ICC-IR 则是使用理想得分数据。使用基于理想得分的 ICC 指标的原因及过程如下:

在认知诊断中, 被试在项目上的作答, 不仅受到被试掌握模式和题目考核模式之间的关系影响,

还受到猜测 ( $g$ ) 和失误 ( $s$ ) 的影响。即被试答对了项目  $j$  可能是被试掌握了该题所有的测量属性, 也可能是未全部掌握该题的测量属性, 但猜对了。同样被试答错了项目  $j$  可能是被试没有全部掌握该题的测量属性, 也有可能是全部掌握了该题的测量属性, 但失误答错了。在理想情况下, 如果被试作答没有受到猜测和失误的影响, Q 矩阵是正确的, 则新题中正确测量模式的 ICC 指标应该为 1。但是由于存在猜测和失误, 使每个新题的正确测量模式的 ICC 指标往往小于 1, 并且还会出现正确测量模式的 ICC 指标小于错误测量模式的 ICC 指标, 从而影响到使用 ICC 指标为题目选择合适的测量模式。因此尽量纠正被试作答矩阵中由于猜测和失误引起的异常可以提高 ICC 指标方法的效果。因此从理论上, ICC-IR 比 ICC-OR 更具优势。

那么, 在实际中 ICC-IR 法的理想得分如何获取呢? 具体过程为: 首先通过已经定义 Q 矩阵的题目估计被试的掌握模式, 然后将相同掌握模式的被试在每个题目上的作答情况进行比较。如果相同掌握模式的被试在项目  $j$  上答对 (得 1 分) 的人数多于答错 (得 0 分) 的人数, 则有理由相信该种掌握模式的被试在项目  $j$  上答错是由于失误造成的。相反, 如果相同掌握模式的被试在项目  $j$  上答对 (得 1 分) 的人数少于答错 (得 0 分) 的人数, 则认为该种掌握模式的被试在项目  $j$  上答对是由于猜测造成的。因此我们可以先通过海明距离 (Chiu & Douglas, 2013) 法估计被试的掌握模式 (为了计算简洁, 这里使用海明距离法估计被试掌握模式, 海明距离法进行诊断分类, 需要先构建被试的理想反应模式, 然后计算被试的得分向量与每个理想反应模式的海明距离。计算公式为:

$$d_h(\mathbf{y}, \boldsymbol{\eta}) = \sum_{j=1}^J |y_j - \eta_j|$$

被试在该题上作答的众数作为该类被试在该题上的理想得分。

## 2.4 基于理想得分的 ICC 指标法 (ICC-IR) 进行 Q 矩阵估计的步骤

若测验测量了  $k$  个属性, 如果属性之间没有层级关系, 则一共有  $2^k$  种属性掌握模式, 有  $2^k - 1$  种题目测量模式。在进行 Q 矩阵界定时, 已知的部分题目 Q 矩阵, 记为  $Q_{base}$ , 未知的 Q 矩阵的题目称为“新题”。具体步骤如下:

第一步: 根据  $Q_{base}$  定义所有新题的属性向量。

(1) 根据  $Q_{base}$ , 使用海明距离估计每个被试的掌握模式。

(2) 根据每个被试的掌握模式, 构造所有被试在所有题目上的理想得分矩阵(方法如前所述)。

(3) 根据  $Q_{base}$ , 用步骤(2)构造出的理想得分矩阵计算新题  $j$  的所有测量模式的 ICC 指标(有  $2^k-1$  个 ICC 指标)。

(4) 将 ICC 指标最大的测量模式作为新题  $j$  的测量模式。

(5) 将新题  $j$  加入到  $Q_{base}$  中, 作为下一个新题界定的基础题目。

重复步骤(1)-(4), 由此界定所有题目的属性向量矩阵  $\hat{Q}$ 。

第二步: 对  $Q$  矩阵进行循环修正。

记第一步得到的  $Q$  矩阵为  $\hat{Q}_0$ , 这个  $\hat{Q}_0$  可能包含一些错误。

(1) 从  $\hat{Q}_0$  中依次挑选一个题目  $j$ , 以其余题目的  $Q$  矩阵作为基础, 估计所有被试的掌握模式。

(2) 根据上一步估计的被试掌握模式, 构造所有被试在所有题目上的理想得分矩阵。

(3) 对于题目  $j$ , 寻找 ICC 指标最大的测量模式, 看其是否与当前题目  $j$  的测量模式相同, 如果不同, 则将 ICC 指标最大的测量模式赋予题目  $j$ 。

重复步骤(1)-(3), 当所有题目都被校正, 得到  $\hat{Q}$ , 则完成一次迭代。当迭代次数达到 20 次或两次迭代间  $Q$  矩阵相同, 则停止循环。

算法结束。

使用以上步骤进行  $Q$  矩阵界定时, 如果第一步中的  $Q_{base}$  题目数量比较少或者  $Q_{base}$  比较单一的时候, 可能会出现某个测量模式在基础题中没有子集、父集和测量模式相同的题目, 从而导致该种测量模式无法计算 ICC 指标。此时将该种测量模式的 ICC 指标指定为 0。依旧使用上述步骤进行  $Q$  矩阵界定。

### 3 ICC 指标计算及理想得分矩阵构建示例

#### 3.1 ICC 指标计算

为了便于理解, 这里给出一个示例展示 ICC 指标的计算方法。假设已知 3 个题目的测量模式(即  $Q_{base}$ ) 如下,

$$Q_{base} = \begin{bmatrix} 100 \\ 110 \\ 111 \end{bmatrix}$$

图 1 已知题目的  $Q$  矩阵

现根据这 3 题已知的测验  $Q$  矩阵来计算第 4 题第一种测量模式的 ICC 指标。这里只使用前 5 个被试的数据, 数据格式如表 1, 其中括号内为题目测量模式。计算如下:

表 2 中, 前 5 个被试在前 3 个题目上的异常作答已经用加粗表示出来。因此可以统计第 4 题为第 1 种测量模式的异常作答次数, 并除以总的比较次数。因此第 4 题的第 1 种测量模式的 ICC 指标为  $1-[2*(1+0+3+1+1)/15]=0.2$ 。其余测量模式 ICC 指标计

表 1  $N$  名被试在前 4 题的观察作答数据

被试	第 1 题 (100)	第 2 题 (110)	第 3 题 (111)	第 4 题 (?)
1	1	0	0	0
2	1	0	0	1
3	1	1	1	0
4	0	1	0	1
5	0	1	1	1
...	...	...	...	...
$N$	1	1	1	1

表 2 第 4 题为第 1 种测量模式的作答异常统计

被试	第 1 题 (100)	第 2 题 (110)	第 3 题 (111)	第 4 题 (100)?	异常作答次数
1	<b>1</b>	0	0	0	1
2	1	0	0	1	0
3	<b>1</b>	<b>1</b>	<b>1</b>	0	3
4	<b>0</b>	1	0	1	1
5	<b>0</b>	1	1	1	1
...	...	...	...	...	...
$N$	1	1	1	1	0

算方法相同。

### 3.2 ICC-IR 法理想得分矩阵构建

当使用理想得分矩阵来计算 ICC 指标时，需要先根据已知题目估计被试的掌握模式，然后构建理想得分矩阵，再计算 ICC 指标。这里假设已经根据已知题目估计出了被试的掌握模式，现根据被试掌握模式构建理想得分矩阵。这里只呈现第 1 种掌握

模式的被试（假设为 5 人）在前 5 个题目上的理想得分矩阵构建方法，其余掌握模式被试的理想得分矩阵类似。过程如下：

统计相同掌握模式的被试在每个题目上作答的众数，将所有相同掌握模式的被试在该题上的作答替换为其众数。在 0~1 计分的测验中，如果被试在题目  $j$  上答对的人数多于答错的人数，则其众数为 1，

表 3 第 1 种掌握模式被试在前 5 个题目上的作答矩阵

	第 1 题	第 2 题	第 3 题	第 4 题	第 5 题
被试 1 (000)	0	0	0	1	0
被试 2 (000)	0	1	0	0	1
被试 3 (000)	0	0	0	0	0
被试 4 (000)	0	0	1	1	0
被试 5 (000)	0	0	0	0	0
Sum	0	1	1	2	1

表 4 第 1 种掌握模式被试在前 5 个题目上的理想得分矩阵

	第 1 题	第 2 题	第 3 题	第 4 题	第 5 题
被试 1 (000)	0	0	0	0	0
被试 2 (000)	0	0	0	0	0
被试 3 (000)	0	0	0	0	0
被试 4 (000)	0	0	0	0	0
被试 5 (000)	0	0	0	0	0

否则为 0。理想得分如表 4。

## 4 研究 1：ICC-IR 法估计 Q 矩阵效果验证

限于文章篇幅，本文仅重点讨论理论上更具优势的 ICC-IR 法。参考国内外已有研究，实验设置如下：

### 4.1 研究设计

#### 4.1.1 Q 矩阵

实验采用 Liu, Xu 和 Ying (2012) 的 Q 矩阵，属性个数分别为 3、4、5 个，题目数量为 20。Q 矩阵如图 2 所示。

$$Q_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \quad Q_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \quad Q_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

图 2 真实 Q 矩阵（引自 Liu 等 (2012)）

#### 4.1.2 题目参数和被试参数模拟

参考 Liu 等人 (2012) 和喻晓锋等人 (2015) 的研究，同时为了研究间的可比性（研究 2），采用与这两个研究相同的认知诊断模型（即 DINA 模型）。模型参数和样本数量设定与喻晓锋等 (2015) 相同，即题目参数按照均匀分布产生，题目参数  $s$  和  $g$  的取值区间为  $[0.05, 0.25]$ 。被试掌握模式按照均匀分布产生，分别产生 400、500、800、1000 人。

#### 4.1.3 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（Bernoulli distribution）中产生被试  $i$  在题目  $j$  上的 0~1 作答反应得分  $response(i, j)$ 。即  $response(i, j) = Bernoulli(P_{ij})$ 。

#### 4.1.4 基础题个数

本研究设置基础题个数为 6、8、10、12 个。从真实 Q 矩阵中随机选取基础题，然后以上述方法进行估计，具体方法前文已有阐述。

#### 4.1.5 评价指标

采用成功估计次数 ( $N_{\text{successful}}$ ) 作为评价指标，即随机生成 100 批数据，使用上述方法完全正确估计 Q 矩阵的次数作为评价指标。计算每次估计的 Q

矩阵中所有题目的测量模式与真实 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), 并计算 100 次实验的平均值。根据每次实验估计的 Q 矩阵来估计被试的掌握模式, 计算所有被试的平均模式判准率 (pattern match ratio, PMR)。

$$N_{\text{successful}} = \sum_{r=1}^{100} (n_{r\_correct}) \quad (3)$$

$$PMR_{\text{item}} = \frac{\sum_{j=1}^J n_{j\_correct}}{J} \quad (4)$$

$$PMR_{\text{subject}} = \frac{\sum_{i=1}^N n_{i\_correct}}{N} \quad (5)$$

式 (3) 中,  $n_{r\_correct}$  表示第  $r$  次实验估计的 Q 矩阵与真实 Q 矩阵是否完全一致, 完全一致则为 1, 否则为 0。式 (4) 中,  $J$  为题目个数,  $n_{j\_correct}$  为估计的第  $j$  题 Q 矩阵是否与真实 Q 矩阵中第  $j$  题一致, 完全一致则为 1, 否则为 0。式 (5) 中,  $N$  为被试人数,  $n_{i\_correct}$  表示估计的被试  $i$  的掌握模式是否与被试  $i$  的掌握模式一致, 如果一致则为 1, 否则为 0。

#### 4.2 研究结果

表 5 呈现了 ICC-IR 法在不同实验条件下的成功估计次数。表 6 是 ICC-IR 法在不同实验条件下的平均题目模式判准率。表 7 是根据 ICC-IR 法估计的 Q 矩阵计算被试的模式判准率。

从表 5 可以看出, 在所有实验条件下 ICC-IR 法的成功估计次数均很高, 特别是当基础题在 8 个以上时, ICC-IR 法在三个 Q 矩阵下的成功估计次数均

表 5 ICC-IR 法进行 Q 矩阵估计的结果 (100 次)

真实 Q 矩阵	人数	基础题个数			
		6	8	10	12
Q1	400	100	100	100	100
	500	100	100	100	100
	800	100	100	100	100
	1000	99	100	100	100
	Mean	99.75	100	100	100
Q2	400	89	100	97	99
	500	90	99	100	99
	800	99	99	100	100
	1000	97	99	100	100
	Mean	93.75	99.25	99.25	99.5
Q3	400	98	100	100	100
	500	92	100	100	100
	800	96	100	100	100
	1000	100	100	99	100
	Mean	96.5	100	99.75	100

接近 100。基础题为 6 个时, 在 Q1 条件下, ICC-IR 法也接近 100 次, 而随着属性个数增多, ICC-IR 法的正确估计次数有下降趋势, 但也能保持在 90 次以上。在三个 Q 矩阵下, 成功估计次数并没有表现出随被试人数增加而增长的趋势。

具体在每个 Q 矩阵下来看, ICC-IR 法在 Q1 下除了 6 个基础题 1000 人的时候出现了一次错误以外, 其他实验条件都能 100% 估计正确。在 Q2 和 Q3 条件下 ICC-IR 法的平均估计正确率也在 90% 以上, 在基础题达到 8 题以上时平均估计正确率能达到 99% 以上。ICC-IR 法在 Q3 下的效果略优于 Q2, 而

Q1 下的效果也优于 Q2。按常理, 属性个数越多对 Q 矩阵的估计越难, 而这里 Q3 属性个数多于 Q2, 但效果却好于 Q2。通过分析 Q 矩阵的形式发现, ICC-IR 法之所以在 3 个 Q 矩阵上的表现反常是由于 3 个 Q 矩阵的情况不同, 调查发现 ICC 指标在测量属性为全属性 (即某个题目考察了所有属性) 与比全属性少一个的属性的测量模式之间差异很小, 如 (1111) 和 (1110)。这是因为计算这两个测量模式的 ICC 指标时所寻找的子集题目、父集题目、测量模式相同的题目范围很相似, 所以 ICC 指标值很接近, 容易出现误判的情况。虽然 Q1 中测量 2 个

属性和 3 个属性的题目较多，但 Q1 属性个数较少，而 Q2 中测量 3 个属性和 4 个属性的题目较多，属性个数也偏多。因此 ICC-IR 法在 Q2 上更容易犯错，

而 Q3 则没有测量属性较多的题目，所以 ICC-IR 法在 Q3 上的表现比较好。这与喻晓锋等人（2015）的研究结论相同。

表 6 ICC-IR 法进行 Q 矩阵估计的题目模式判准率 ( $PMR_{item}$ )

真实 Q 矩阵	人数	基础题个数			
		6	8	10	12
Q1	400	1	1	1	1
	500	1	1	1	1
	800	1	1	1	1
	1000	.994	1	1	1
	Mean	.999	1	1	1
Q2	400	.968	1	.999	1
	500	.952	.999	1	1
	800	.997	.996	1	1
	1000	.986	1	1	1
	Mean	.976	.999	1	1
Q3	400	.992	1	1	1
	500	.964	1	1	1
	800	.983	1	1	1
	1000	1	1	.998	1
	Mean	.985	1	.999	1

表 6 和表 5 相对应，在 Q1 下 ICC-IR 法的平均题目模式判准率非常接近于 1。在 Q2 条件下，ICC-IR 法的平均题目模式判准率同样很高，在基础题为 6 个时，题目模式判准率也在 97% 以上。在 Q3 情况下，ICC-IR 法的题目模式判准率也接近于 1。

表 7 是被试的模式判准率，这里使用的是海明距离判别法。对于估计出的 Q 矩阵，如果越接近于真

实的 Q 矩阵，模式判准率应该越高。在 Q1 下，模式判准率均保持在 .9 以上。在 Q2 条件下，模式判准率也接近 .8。而随着属性个数增多，模式判准率有所下降，这也反映出属性个数对模式判准率的影响。

## 5 研究 2：ICC-IR 法与 $D^2$ 统计量的比较

为了比较 ICC-IR 法和已有方法在估计 Q 矩阵上

表 7 ICC-IR 法进行 Q 矩阵估计的被试平均模式判准率 ( $PMR_{subject}$ )

真实 Q 矩阵	人数	基础题个数			
		6	8	10	12
Q1	400	.922	.92	.922	.92
	500	.921	.921	.922	.922
	800	.922	.922	.924	.92
	1000	.914	.922	.922	.921
	Mean	.919	.921	.922	.920
Q2	400	.774	.805	.803	.802
	500	.763	.801	.804	.798
	800	.803	.795	.8	.801
	1000	.790	.803	.804	.802
	Mean	.782	.801	.802	.800
Q3	400	.742	.754	.751	.757
	500	.729	.748	.749	.755
	800	.739	.759	.758	.75
	1000	.752	.751	.751	.751
	Mean	.740	.753	.752	.753

的效果, 将 ICC-IR 法与喻晓峰等人 (2015) 提出的  $D^2$  统计量方法进行比较。 $D^2$  统计量方法通过计算不同掌握模式被试在题目上的实际答对比例与期望答对概率之间的残差, 认为当题目的测量模式正确时, 残差最小。因此可以通过该方法来为题目选择合适的测量模式。 $D^2$  统计量需要先估计被试的掌握模式和项目参数, 然后才能计算。 $D^2$  统计量计算公式如下:

$$D_{q_j}^2 = 2 \sum_{i=1}^{2^k} \left[ r_i \log \frac{p_{ij}}{(1-s_j)^{r_i} g_j^{1-r_i}} + (N_i - r_i) \log \frac{(1-p_{ij})}{s_j^{r_i} (1-g_j)^{1-r_i}} \right] \quad (6)$$

实验设置与研究 1 相同。由于  $D^2$  统计量计算量大, 估计时间较长, 特别是当属性个数越多, 估计时间越长。因此仅对 3 属性和 4 属性的情况进行实验。

研究结果如下:

表 8 呈现了 2 种方法在不同实验下的成功估计次数, 表 9 呈现了不同实验条件下的平均题目模式判准率。表 8 结果显示, 总体上 ICC-IR 法效果优于  $D^2$  统计量效果。当基础题个数越少,  $D^2$  统计量的效果越差。而随着基础题个数增加,  $D^2$  统计量的方法估计准确率有所上升, 但在 Q2 条件下, 即使基础题增加到 12 个,  $D^2$  统计量方法的成功率还是只能达到 60% 左右。因此  $D^2$  统计量的估计效果在基础题个数较少时不太理想, 特别属性个数较多时。而 ICC-IR 法则在基础题个数较少和属性个数增多时均表现得比较稳健。

表 8 2 种方法进行 Q 矩阵估计 100 次实验完全正确的次数

真实 Q 矩阵	人数	基础题个数							
		6		8		10		12	
		$D^2$	ICC-IR	$D^2$	ICC-IR	$D^2$	ICC-IR	$D^2$	ICC-IR
Q1	400	23	100	53	100	80	100	98	100
	500	30	100	48	100	75	100	94	100
	800	21	100	54	100	81	100	95	100
	1000	21	99	57	100	89	100	98	100
	Mean	23.75	99.75	53	100	81.25	100	96.25	100
Q2	400	4	89	12	100	29	97	55	99
	500	6	90	17	99	39	100	56	99
	800	5	99	12	99	31	100	60	100
	1000	4	97	9	99	48	100	70	100
	Mean	4.75	93.75	12.5	99.25	36.75	99.25	60.25	99.5

表 9 2 种方法进行 Q 矩阵估计的题目模式判准率 ( $PMR_{item}$ )

真实 Q 矩阵	人数	基础题个数							
		6		8		10		12	
		$D^2$	ICC-IR	$D^2$	ICC-IR	$D^2$	ICC-IR	$D^2$	ICC-IR
Q1	400	.424	1	.662	1	.865	1	.990	1
	500	.461	1	.627	1	.842	1	.964	1
	800	.414	1	.677	1	.884	1	.975	1
	1000	.397	.994	.707	1	.931	1	.992	1
	Mean	.424	.999	.668	1	.880	1	.980	1
Q2	400	.270	.968	.376	1	.565	.999	.719	1
	500	.280	.952	.398	.999	.602	1	.734	1
	800	.263	.997	.379	.996	.539	1	.738	1
	1000	.284	.986	.336	1	.627	1	.816	1
	Mean	.274	.976	.372	.999	.583	1	.752	1

### 6 研究 3: ICC-IR 法和 $D^2$ 统计量方法在实测数据中的比较研究

为进一步研究 ICC-IR 法和  $D^2$  统计量方法在实测数据中的效果, 本研究采用 Tatsuoka 的分数减法数据进行实验, 该数据是由 Tatsuoka (1984) 收集的, 包

括 15 个题目和 536 个被试作答。共测量了 5 个属性。分数减法的 Q 矩阵详见 (de la Torre, 2008)。

为保证选取的基础题是相对有把握的, 我们选取  $s$  和  $g$  平均值最小的题目作为基础题, 因为  $s$  和  $g$  小说明题目既没有属性多余也没有属性缺失, 则可以

认为题目的测量模式是正确的。根据已有研究 (de la Torre, 2008) 估计的项目参数, 将  $s$  和  $g$  平均值从小到大进行排序。分别选取前 6、7、8、9、10 个题目

为基础题, 以剩余的题目作为待估计的题目。结果如下, 其中括号内为相同元素的个数占总元素个数的比例, 本例中总元素为  $15 \times 5 = 75$  个。

表 10 ICC-IR 法和  $D^2$  统计量方法对分数减法数据估计的 Q 阵与专家 Q 阵的一致率

Q 矩阵估计方法	基础题个数				
	6	7	8	9	10
$D^2$	32 (.426)	31 (.413)	35 (.466)	42 (.56)	40 (.533)
ICC-IR	54 (.72)	58 (.773)	60 (.8)	64 (.853)	69 (.92)

从表 10 可以看出, 在实测数据中,  $D^2$  统计量的方法的估计效果相对于 ICC-IR 法低。ICC-IR 法的估计效果随着基础题个数增加有所上升, 当基础题达到 10 个时, 估计的 Q 矩阵与真实 Q 矩阵之间一致率

能达到 92%。而  $D^2$  统计量的估计效果随基础题个数的增加变化不明显。

为了比较两种方法估计的 Q 矩阵的优劣, 我们计算了不同 Q 矩阵下的模型拟合指标 (负 2 倍的对数似然、AIC 和 BIC 指标)。结果如下:

表 11  $D^2$  统计量方法与 ICC-IR 法估计的 Q 矩阵的模型拟合度比较

Q 矩阵估计方法	模型拟合度	(估计的 Q 矩阵) 基础题个数				
		6	7	8	9	10
$D^2$	$-2 \cdot \log(L)$	7240.00	7118.09	7065.39	7270.38	7342.77
	AIC	7362.00	7240.09	7187.39	7392.38	7464.77
	BIC	7623.33	7501.42	7448.72	7653.71	7726.10
ICC-IR	$-2 \cdot \log(L)$	6889.93	6895.39	6928.53	6969.54	6911.51
	AIC	7011.93	7017.39	7050.53	7091.54	7033.51
	BIC	7273.26	7278.73	7311.86	7352.87	7294.85

从表 11 可以看出,  $D^2$  统计量估计的 Q 矩阵在模型拟合上均不如 ICC-IR 法。因此从实证数据的结果可以看出, ICC-IR 法也优于  $D^2$  统计量的方法。

## 7 结论与讨论

### 7.1 结论

(1) ICC-IR 法能在 Q 矩阵估计上具有很好的效果, 并且当 Q 矩阵中题目测量的属性个数越少、基础题个数越多, 估计效果越好。

(2) 相对于  $D^2$  统计量的方法, ICC-IR 法效果更好, 并且是一种非参数化的方法, 计算简单快捷。

(3) 实证数据分析表明, ICC-IR 法估计的 Q 矩阵在模型拟合度上也优于  $D^2$  统计量方法。

### 7.2 讨论

#### (1) ICC-IR 法的优势和局限

ICC-IR 法不需要进行参数估计, 相对于参数化的方法而言更简单。在运行时间上也具有优势, 如在人数为 1000 人, 属性个数为 3 个, 20 题中抽取 6 个基础题来估计剩余题目,  $D^2$  统计量 100 次实验平均每次实验需要 1857 秒, 而 ICC-IR 法不超过 30 秒, 运算时间约为  $D^2$  统计量的 1/60。但是该方法会受到基础 Q 矩阵形式以及需要界定的题目测量模式的影

响, 不容易区分全属性题目与测量属性个数为  $k-1$  个的题目; 该方法暂时只能用于 0-1 计分的题目, 对于多级计分数题目的 Q 矩阵估计方法还需要进一步探索。此外 ICC-IR 法估计 Q 矩阵的效果还应该在更多真实数据中进行检验。

#### (2) 关于基础题和有争议题目的处理

ICC 指标需要比较测量模式之间的关系, 因此应尽量保证基础题 Q 矩阵的丰富性。最好使每种测量模式在基础 Q 矩阵中有相关 (子集、父集、测量模式相同) 的题目。如果在此基础上能增加基础题的个数, 则对 Q 矩阵估计更加有益。此外不同的 Q 矩阵估计方法可以用来相互验证所估计的 Q 矩阵的合理性。将不同的方法估计出来的 Q 矩阵进行比较, 将有争议的题目提取出来进行专家讨论或者删除这些有争议的题目。这样既减轻了专家的负担、避免了专家的主观性, 也避免了客观方法的局限性。

#### 参考文献

- 陈平, 辛涛. (2011). 认知诊断计算机化自适应测验中在线标定方法的开发. *心理学报*, 43(6), 710-724.
- 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.
- 汪文义, 丁树良, 游晓峰. (2011). 计算机化自适应诊断测验中原题的属性标定. *心理学报*, 43(8), 964-976.

- 喻晓峰, 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.
- Chiu, C. Y. (2013). Statistical refinement of the q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, 37(8), 598-618.
- Chiu, C. Y., & Douglas, J. A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal item response patterns. *Journal of Classification*, 30(2), 225-250.
- Cui, Y., Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2006). A person-fit statistic for the attribute hierarchy method: The hierarchy consistency index. *Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.*
- DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, 35(1), 8-26.
- 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.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, 41(3), 205-237.
- Liu, J. C., Xu, G. J., & Ying, Z. L. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, 36(7), 548-564.
- 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.
- Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems*. Urbana, IL: Computer-Based Education Research Library, University of Illinois.
- Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Routledge.

## A New Q-matrix Estimation Method: ICC based on Ideal Response

Wang Daxun, Gao Xuliang, Cai Yan, Tu Dongbo

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

**Abstract** Nowadays, we are not satisfied with a total score from measurement, but hope to get an informative report. As the core of the new generation test theory, Cognitive Diagnosis (CD) attracts more and more people's attention. Since it can reveal the result and form a microscopic perspective, such as individuals' knowledge structures, processing skills and cognitive procedure etc, it would help us to take individualized teaching and promote students' development. Cognitive diagnosis assessments infer the attribute mastery pattern of respondents by item responses based on Q-matrix. The Q-matrix plays the role of a bridge between items and respondents. Many studies have shown that misspecification of the Q-matrix can affect the accuracy of model parameters and result in the misclassification of respondents. In practice, Q-matrix is established by experts. However, different experts may provide different Q-matrices. To avoid the subjectivity from experts in Q-matrix specification and ensure the correct of Q-matrix, researchers are trying to look for objective methods. Nevertheless, existing methods need information from parameter and a large amount of computation.

To simplify the method of Q-matrix estimation, this article introduces a new Q-matrix estimation method based on ICC (Item Consistency Criterion). The logic of the method is as follows: if the measurement pattern of the item A is a subset of the item B. The logic of the ICC method is that it is impossible for a person to get "0" score on item A, but to get "1" on item B. Of course, if item A and item B have the same measurement pattern, it is impossible that a person gets "1" score on item A, but get "0" on item B (or, the other way around). From this logic we come up with Item Consistency Criterion. In order to improve the effect of ICC method, we come up with ICC-IR (ICC based on ideal response) method.

In order to explore the effect of this method, we considered different numbers of participants, different numbers of base items and different Q-matrices whose attribute numbers are different. The item parameters and attribute mastery pattern of respondents are obeyed a uniform distribution. In addition, we compared with the Likelihood D2 Statistic.

The Monte Carlo simulation study and real data study showed that: generally, the ICC-IR method could recover the real Q-matrix with a high rate of success. Compared with the Likelihood D2 Statistic, the ICC-IR method was better. Furthermore, the ICC-IR method was easier to understand and needed less computation. The real data study also showed that the ICC-IR method could estimate the Q-matrix with a high success rate. Besides, without the needs of parameters estimation, the method was not affected by the deviation caused by the misfit between model and data. In a word, the method is simple and effective in Q-matrix Estimation, which is meaningful to the simplification of cognitive diagnosis.

**Key words** cognitive diagnosis, Q-matrix, Item Consistency Criterion, DINA model