

双因子模型 MCAT 中多级评分项目选题策略的比较*

毛秀珍** 刘欢 唐倩

(四川师范大学教育科学学院, 成都, 610066)

摘要 针对多级评分项目计算机化自适应测验, 推导双因子等级反应模型 Fisher 信息量的计算, 然后将“维度缩减”方法运用到后验加权 D 优化 (PDO)、后验加权 Kullback-Leibler (PKL)、连续熵 (CEM) 和互信息 (MI) 方法, 并在低、中、高双因子模式下比较它们的表现。模拟研究表明: (1) 双因子模式越强, 全局因子估计精度降低, 局部因子估计精度提高; (2) 相同实验条件下, CEM 方法的测量精度最高, PKL 方法的估计精度最低, 其它方法没有显著差异。

关键词 多维项目反应理论 双因子等级反应模型 计算机化自适应测验 选题方法 测量精度

1 引言

计算机化自适应测验 (computerized adaptive testing, CAT) 根据被试潜在特质水平自适应地选择测验项目, 极大地提高了测验效率。CAT 的另一个特征是它可以灵活应用各种测量模型, 如项目反应理论模型、多维项目反应理论模型和认知诊断理论模型。

双因子模型假设测验考察一个全局因子 (general factor, G) 和多个特定领域因子或称为局部因子 (specific domain factor or group factor, S)。它是一类特殊的多维模型, 曾用于大量有关智力测验 (Watkin & Beaujean, 2014)、自我领导测验 (Furthner, Rauthmann, & Schse, 2015)、学校适应不良行为 (Wiesner & Schanding, 2013)、教育成就测验 (Cai, Yang & Hansen, 2011; DeMars, 2006) 以及精神病诊断筛选测验 (psychiatric diagnostic screening questionnaire, PDSQ) 的结构分析 (Gibbons, Rush, & Immekus, 2009)。这些研究表明双因子模型符合认知能力、心理特质、精神病理等多类测验的结构特征, 很多情况下比其它竞争模型 (如单维、高阶和相关特质模型) 能更准确地反映量表维度。

针对双因子模型多维 CAT (MCAT), Weiss 和 Gibbons (2007) 在具有双因子结构的题库中将 CAT 分为多个阶段, 每个阶段测量一种能力。通过多个单维 CAT 依次测量全局因子和局部因子。与纸笔测验相比, 双因子 CAT 的测验长度平均减少了 80%, 实际节省约 82% 的测验时间。Zheng, Chang 和 Chang (2013) 在双因子题库中考察了内容约束条件对测量精度和项目使用率的影响。由此可见, 双因子模型 MCAT 实际可行, 兼具双因子模型和 CAT 的优势, 具有广泛的应用前景。

Seo (2011) 将 MCAT 项目选择和能力估计方法应用到双因子项目反应理论模型 (Cai et al., 2011), 在 MCAT 中同时估计全局因子和局部因子。Seo 和 Weiss (2015) 进一步在低双因子结构、高双因子结构和似双因子结构题库中比较了 D- 优化、DS- 优化, A 优化和 E 优化四种项目选择方法。

上述研究都针对二级评分项目探讨双因子 MCAT 项目选择方法。多级评分项目与相同项目二分化后相比, 能提供更多信息量 (Samejima, 1976)。van Rijn, Eggen, Hemker 和 Sanders (2002) 发现, 与二级评分项目库相比, 多级评分项目库对极端能力被试的估计误差更小。特别地, CAT 背景

* 本研究得到国家自然科学基金青年项目 (31400897) 的资助。

** 通讯作者: 毛秀珍。E-mail: maomao_wanli@163.com

DOI:10.16719/j.cnki.1671-6981.20190128

下使用多级评分项目还能降低测验长度、调整项目曝光率 (Ayala, Dodd, & Koch, 1992)。Lin (2012) 指出自从在 K-12 系统和其它高风险测验中使用多级评分项目, 测量多种能力和使用多级评分项目的测验已变得越来越重要。鉴于此, 本文以多级评分项目为研究对象。

MCAT 中随着测验考察维度的增加, 多维能力积分的计算越复杂。这常常阻碍后验加权 Kullback-Leibler 信息 (PKL)、连续熵 (continuous entropy method, CEM) 和互信息 (mutual information, MI) 的应用。而双因子模型通常假设全局因子与局部因子, 局部因子之间相互独立。在此基础上, 双因子模型“维度缩减”方法在能力估计中能将多个维度能力的积分运算化简为多个二维迭代积分 (Cai et al., 2011)。于是, “维度缩减”方法能否应用于 PKL、CEM 和 MI 方法以简化计算成为本文的核心问题。研究通过推导双因子等级反应模型下 Fisher 信息量, 运用模拟实验比较上述选题方法的表现。

2 双因子模型

2.1 双因子等级反应模型

Holzinger 和 Swineford (1937) 在 Spearman 两因素模式基础上推广提出双因子模型。项目水平的双因子模型要求每个观察变量测量全局因子和一个局部因子, 多个观察变量共同测量一个局部因子。

双因子模型中假设测验考察全局因子 θ_0 和 G 个局部因子 $\theta_1, \dots, \theta_G$ 即 $\theta = [\theta_0, \theta_1, \dots, \theta_G]^T$ 。项目 j 只考察全局因子和一个局部因子 θ_g , 因而其区分度向量为 $a_j = (a_{j0}, 0, \dots, a_{jg}, \dots, 0)$ 。令项目 j 有 K 个反应类别, $x_{ij} (x_{ij} = 0, 1, 2, \dots, K-1)$ 表示被试 i 在项目 j 的反应, $b_t (t = 1, 2, \dots, K-1)$ 表示项目难度。则双因子等级反应模型作为两步模型, 首先需要计算作答反应大于等于类别 $t (t = 1, 2, \dots, K-1)$ 的概率:

$$P_{ijt}^* = P(x_{ij} \geq t | \theta_{i0}, \theta_{ig}) = \frac{1}{1 + \exp[-1.702 \cdot a_j(\theta_i - b_j \cdot t)]} \quad (1)$$

特别地, $P(x_{ij} \geq 0) = 1, P(x_{ij} \geq K) = 0$ 。于是, 第 t 个类别上的作答概率 P_{ijt} 等于相邻两个类别累积作答概率的差, 即

$$P_{ijt} = P(x_{ij} = t | \theta_{i0}, \theta_{ig}) = P_{ijt}^* - P_{ijt+1}^* \quad (2)$$

2.2 双因子等级反应模型的 Fisher 信息矩阵

记反应 $x_{ij} = t (t = 0, 1, \dots, K-1)$ 时, 指示函数 $I(x_{ij} = t) = 1$, 否则 $I(x_{ij} = t) = 0$ 。假设测验考察 p 个维度, 项目 j 的 Fisher 信息量为 $p \times p$ 维矩阵, 其 $(m,$

$n)$ 元素定义为:

$$I_{j,(m,n)}(\theta) = -E\left[\frac{\partial^2 \ln P(x_{ij} | \theta)}{\partial \theta_m \partial \theta_n}\right] \quad (3)$$

其中 $\ln P(x_{ij} | \theta) = \ln \prod_{t=0}^{K-1} P_{ijt}^{I(x_{ij}=t)} = \sum_{t=0}^{K-1} I(x_{ij}=t) \cdot \ln P_{ijt}$, 并且

$$\frac{\partial \ln P(x_{ij} | \theta)}{\partial \theta_m} = \sum_{t=0}^{K-1} I(x_{ij}=t) \frac{1}{P_{ijt}} \cdot \frac{\partial P_{ijt}}{\partial \theta_m} \quad (4)$$

$$\frac{\partial^2 \ln P(x_{ij} | \theta)}{\partial \theta_m \partial \theta_n} = \sum_{t=0}^{K-1} \left[\left(\frac{-I(x_{ij}=t)}{P_{ijt}^2} \right) \cdot \frac{\partial P_{ijt}}{\partial \theta_m} \cdot \frac{\partial P_{ijt}}{\partial \theta_n} + \frac{I(x_{ij}=t)}{P_{ijt}} \cdot \frac{\partial^2 P_{ijt}}{\partial \theta_m \cdot \partial \theta_n} \right] \quad (5)$$

经化简、求导和求期望后, 可得项目 j 的 Fisher 信息矩阵为:

$$I_j(\theta) = -E\left(\frac{\partial^2 \ln P(x_{ij} | \theta)}{\partial \theta^2}\right) = a_j^T \cdot a_j \cdot \sum_{t=0}^{K-1} P_{ijt}(1 - P_{ijt}^* - P_{ijt+1}^*)^2 \quad (6)$$

注意到, 二级评分两参数 logistic 模型的结果是 (6) 式的特例。

3 项目选择方法

令 $S_{k-1} = \{i_1, i_2, \dots, i_{k-1}\}$ 、 $X_{k-1} = \{x_{i1}, x_{i2}, \dots, x_{ik-1}\}$ 和 R_k 分别表示被试 i 已施测项目集合、对应项目的反应和剩余题库。首先, 记 S_{k-1}^g 包含集合 S_{k-1} 中考察第 g 个特殊因子的项目, 并记 $\theta_g^* = [\theta_0, \theta_g]^T$ 表示影响项目作答的能力因子。当项目 i_j 考察第 g 个特殊因子时, 指示函数 $u_{ij,g} = 1$, 反之 $u_{ij,g} = 0$ 。其次, 函数 $P(\mathbf{X}_{k-1} | \theta) \cdot f(\theta) = \prod_{i_j \in S_{k-1}} \prod_{t=0}^{K-1} P(x_{ij} = t | \theta)^{I(x_{ij}=t)} \cdot f(\theta)$ 中能力满足独立性假设, 即联合分布为 $f(\theta) = f(\theta_0, \theta_1, \dots, \theta_G) = f(\theta_0) \cdot f(\theta_1) \cdot \dots \cdot f(\theta_G)$ 。另外, 项目反应满足局部独立, 并且考察不同特殊因子的项目集合之间不相交, 于是

$$P(\mathbf{X}_{k-1} | \theta) \cdot f(\theta) = f(\theta_0) \cdot \prod_{g=1}^G \left[\prod_{i_j \in S_{k-1}^g} \prod_{t=0}^{K-1} P(x_{ij} = t | \theta_g^*)^{I(x_{ij}=t)} \cdot f(\theta_g) \right], \quad (7)$$

$$\text{令 } \prod_{i_j \in S_{k-1}^g} \prod_{t=0}^{K-1} P(x_{ij} = t | \theta_g^*)^{I(x_{ij}=t)} = L_{S_{k-1}^g}(t), \text{ 则} \\ A_{k-1} = \int P(\mathbf{X}_{k-1} | \theta) \cdot f(\theta) d\theta = \int_{\theta_0} \{f(\theta_0) \cdot \prod_{g=1}^G [\int_{\theta_g} L_{S_{k-1}^g}(t) \cdot f(\theta_g) d\theta_g]\} d\theta_0. \quad (8)$$

3.1 贝叶斯 D- 优化方法 (DO)

贝叶斯 D- 优化方法是 MCAT 中最基本和常用的选题方法 (Segall, 1996), 其项目选择指标表示为: $D_{i_k} = \arg \max \det(I_{S_k}(\hat{\theta}^{k-1}) + \Sigma_0^{-1})$ 。 (9)

其中, $I_{S_k}(\hat{\theta}^{k-1}) = I_{S_{k-1}}(\hat{\theta}^{k-1}) + I_{i_k}(\hat{\theta}^{k-1})$, $I_{S_{k-1}}$ 与 I_{i_k} 分别表示已经施测项目集合和候选项目的 Fisher 信息矩阵, Σ_0^{-1} 代表能力先验分布方差协方差矩阵的逆矩

阵。

3.2 后验加权 Fisher 信息 D- 优化方法 (Posterior weighted Fisher D-optimality, PDO)

由于测验初期能力估计不准确, PDO 方法将选择使后验加权 Fisher 信息矩阵行列式值最大的项目, 即:

$$PDO_{i_k} = \arg \max_{i_k \in R_k} \det \left[\int_{\theta \in N(\hat{\theta}^{k-1}, \delta_k)} I_{S_k}(\theta) \cdot P(\theta | \mathbf{X}_{k-1}) d\theta \right] \quad (10)$$

其中 $N(\hat{\theta}^{k-1}, \delta_k)$ 表示以 $\hat{\theta}^{k-1}$ 为中心, δ_k 为半径的领域, δ_k 通常取值为 $3/\sqrt{k}$ (Wang & Chang, 2011)。根据矩阵积分计算法则, (10) 式中 m 行 n 列元素等于对 $I_{S_k}(\theta)$ 中 m 行 n 列元素积分, 即 $I_{S_k, (m, n)} = \int_{\theta \in N(\hat{\theta}^{k-1}, \delta_k)} (\sum_{i_j \in S_k} I_{i_j, (m, n)}) \cdot P(\theta | \mathbf{X}_{k-1}) d\theta$

由于 $P(\theta | \mathbf{X}_{k-1}) \propto P(\mathbf{X}_{k-1} | \theta) \cdot f(\theta)$, 根据 (8) 式, 有

$$I_{S_k, (m, n)} = \sum_{i_j \in S_k} \int_{\theta_0 - \delta_k}^{\theta_0 + \delta_k} \left\{ \prod_{g=1}^G \left[\int_{\theta_g - \delta_k}^{\theta_g + \delta_k} (I_{i_j, (m, n)})^{u_{ij, g}} \cdot L_{S_{k-1}}^g(t) \cdot f(\theta_g) d\theta_g \right] \right\} f(\theta_0) d\theta_0. \quad (11)$$

由此可见, 利用“维度缩减”可将 PDO 的计算简化为 G 个二维迭代积分计算。特别地, 若记每个因子有 Q 个积分节点, 全局因子维度的积分节点记为 o_{q_0} , 第 g 个局部因子的积分节点记为 o_{q_g} , 则 (11) 式运用数值积分近似等于:

$$I_{S_k, (m, n)} \approx \sum_{i_j \in S_k} \sum_{q_0}^Q \left\{ \prod_{g=1}^G \left[\sum_{q_g}^Q (I_{i_j, (m, n)}(o_{q_0}, o_{q_g}))^{u_{ij, g}} \cdot L_{S_{k-1}}^g(o_{q_0}, o_{q_g}) \cdot w(o_{q_g}) \right] \right\} \cdot w(o_{q_0}). \quad (12)$$

3.3 后验期望 Kullback-Leibler 方法 (Posterior Expected Kullback-Leibler Information, PKL)

PKL 依据 (13) 式来选择第 i_k 个项目 (Wang & Chang, 2011),

$$PKL_{i_k} = \arg \max_{i_k \in R_k} \int_{\theta \in N(\hat{\theta}^{k-1}, \delta_k)} KL_{i_k}(\hat{\theta}^{k-1} \| \theta) \cdot P(\theta | \mathbf{X}_{k-1}) d\theta \quad (13)$$

将项目 KL 信息量代入 (13), 依据 (8) 式对 (13) 式化简, 可得,

$$PKL_{i_k} \propto \sum_{t=0}^{K-1} P(x_{i_k} = t | \hat{\theta}^{k-1}) \int_{\theta_0 - \delta_k}^{\theta_0 + \delta_k} \left\{ \prod_{g=1}^G \int_{\theta_g - \delta_k}^{\theta_g + \delta_k} \left[\log \left(\frac{P(x_{i_k} = t | \hat{\theta}^{k-1})}{P(x_{i_k} = t | \theta_g^t)} \right) \right]^{u_{i_k, g}} \cdot L_{S_{k-1}}^g(t) \cdot f(\theta_g) d\theta_g \right\} \cdot f(\theta_0) d\theta_0$$

3.4 连续熵方法 (Continuous Entropy method, CEM)

CEM 方法依据被试 i 在施测第 i_k 个项目后能力后验分布 $P(\theta | \mathbf{X}_{k-1}, x_{i_k})$ 的期望后验连续熵选择项目, 其标准为:

$$CEM_{i_k} = \arg \min_{i_k \in R_k} \sum_{t=0}^{K-1} H(P(\theta | \mathbf{X}_{k-1}, x_{i_k})) \cdot P(x_{i_k} = t | \mathbf{X}_{k-1}). \quad (14)$$

其中, $H(P(\theta | \mathbf{X}_{k-1}, x_{i_k})) = \int P(\theta | \mathbf{X}_{k-1}, x_{i_k} = t) \cdot \log[1/P(\theta | \mathbf{X}_{k-1}, x_{i_k} = t)] d\theta$ 。经化简, 项目 i_k 的连续熵与

$\sum_{t=0}^{K-1} (C(t) - D(t))$ 成正比, 其中

$$C(t) = \left[\log \int P(\mathbf{X}_{k-1}, x_{i_k} = t | \theta) \cdot f(\theta) d\theta \right] \cdot \left[\int P(\mathbf{X}_{k-1}, x_{i_k} = t | \theta) \cdot f(\theta) d\theta \right] \quad (15)$$

$$D(t) = \int P(\mathbf{X}_{k-1}, x_{i_k} = t | \theta) \cdot f(\theta) \cdot \log(P(\mathbf{X}_{k-1}, x_{i_k} = t | \theta) \cdot f(\theta)) d\theta \quad (16)$$

在双因子模型假设下, $D(t)$ 可以继续分解为 $E+F$, 且,

$$E = \int_{\theta_0} \left[\prod_{g=1}^G \int_{\theta_g} L_{S_k}^g(t) \cdot f(\theta_g) d\theta_g \right] \cdot f(\theta_0) \cdot \log f(\theta_0) d\theta_0 \quad (17)$$

$$F = \sum_{g'=1}^G \int_{\theta_0} \left[\prod_{\substack{g=1 \\ g \neq g'}}^G \int_{\theta_g} L_{S_k}^g(t) \cdot f(\theta_g) d\theta_g \right] \cdot \left[\int_{\theta_{g'}} L_{S_k}^{g'}(t) \cdot f(\theta_{g'}) \cdot \log(L_{S_k}^{g'}(t) \cdot f(\theta_{g'})) d\theta_{g'} \right] \cdot f(\theta_0) d\theta_0 \quad (18)$$

3.5 互信息项目选择方法 (mutual information method, MI)

假设 X 代表 $P(x_{i_k} = t | \mathbf{X}_{k-1})$, Y 代表 $P(\theta | \mathbf{X}_{k-1})$, 互信息定义为 X 与 Y 的联合的分布与它们边际分布乘积的 KL 距离, 即 MI 选题标准为:

$$MI_{i_k} = \arg \max_{i_k \in R_k} \left\{ \sum_{t=0}^{K-1} \int P(\theta, x_{i_k} = t | \mathbf{X}_{k-1}) \log \frac{P(\theta, x_{i_k} = t | \mathbf{X}_{k-1})}{P(\theta | \mathbf{X}_{k-1}) P(x_{i_k} = t | \mathbf{X}_{k-1})} d\theta \right\} \quad (19)$$

项目 i_k 的互信息可化简为:

$$I_{MI_{i_k}} = (1/A_{k-1}) \sum_{t=0}^{K-1} \left[\int P(\mathbf{X}_{k-1}, x_{i_k} = t | \theta) \cdot f(\theta) \log P(x_{i_k} = t | \theta) d\theta + B_k \cdot \log(A_{k-1}/A_k) \right] \quad (20)$$

其中, $\int P(\mathbf{X}_{k-1}, x_{i_k} = t | \theta) \cdot f(\theta) \cdot \log P(x_{i_k} = t | \theta) d\theta$ 也可化为二维迭代积分

$$\int_{\theta_0} f(\theta_0) \cdot \prod_{g=1}^G \left[\int_{\theta_g} L_{S_k}^g(t) \cdot f(\theta_g) \cdot (\log P(x_{i_k} = t | \theta))^{u_{i_k, g}} d\theta_g \right] d\theta_0$$

由此可见, “维度缩减”方法可以运用到上述方法的计算。

4 研究设计

研究采用 MATLAB (R2010a) 为工具编写 CAT 代码, 在低、中和高双因子模式题库中比较贝叶斯 D-优化方法、PDO、PKL、CEM 和 MI 方法的选题表现。

4.1 题库的模拟

借鉴 Seo 和 Weiss (2015) 的研究, 本文模拟生成低、中和高双因子模式题库, 每个题库包含 400 个项目。假设每个题库中前 200 个项目考察第一个局部因子, 区分度向量为 $(a_{j_0}, a_{j_1}, 0)$ ($j = 1, 2, \dots, 200$)

；后 200 个项目考察第二个局部因子，区分度向量为 $(a_{j_0}, 0, a_{j_2}) (j = 201, \dots, 400)$ 。特别地， a_{j_0} 服从对数正态分布，即 $a_{j_0} \sim \log N(0, 0.2)$ 。 a_{j_1} 和 a_{j_2} 服从均匀分布，即 $a_{j_1} \sim U(0, a_{j_0}p)$ 和 $a_{j_2} \sim U(0, a_{j_0}p)$ 。其中， $p = \cos(\pi/3), \cos(\pi/6), 1$ 分别对应低、中和高双因子模式的题库。 p 值越大，全局因子和局部因子在项目上的区分度的差异越小，题库的双因子结构越强，因而称为高双因子模式。 p 值越小，项目在这两个维度上的区分度差异越大，则项目主要考察全局因子，从而称为低双因子模式。

另外，van Rijin 等（2002）指出多级评分单维 CAT 背景下，包含三个反应类别的项目表现最好。因此，研究假设多级评分项目具有 3 个评分类别。为保证项目步骤难度递增且其分布能覆盖能力范围，假设项目 j 的步骤难度参数服从均匀分布，即 $b_{j_1} \sim U(-4, 2)$ 和 $b_{j_2} \sim U(b_{j_1}, 4)$ 。

4.2 模拟被试及作答反应

从标准多变量正态分布中随机产生 2000 名被试。运用双因子多维等级反应模型计算被试 i 在项目 j 上反应类别 k 的累积概率 $P_{ijk}^* (k = 1, 2)$ ，并产生 $(0, 1)$ 区间的随机数 p_{ij} 。当 $p_{ij} \leq 1 - P_{ij2}^*$ 时，反应为 0；当 $P_{ij1}^* < p_{ij} < 1 - P_{ij2}^*$ 时，反应为 1，否则反应为 2。

4.3 能力估计方法

依据 Yao（2014）和 Huebner, Wang, Quinlan 和 Seubert（2016）关于能力估计方法研究的结果，基于能力估计精度和计算简便性考虑，本文采用最大后验估计（maximum a posterior, MAP）估计能力。能力先验分布服从标准多变量正态分布，迭代结束标准为所有能力维度上、后两次估计值差的绝对值小于 0.001，最大迭代次数为 25。

4.4 评价指标

测验长度为 40，每施测 5 个项目时依据 $EDU = (1/N) \cdot \sum_{i=1}^N \sqrt{\sum_{l=0}^2 (\theta_{il} - \hat{\theta}_{il})^2}$ 计算欧氏距离。根据能力估计值与真值的相关、能力估计值的平均绝对值误差（average abosutle bias, ABS）、平均均方根误差（root mean square error, RMSE）和欧氏距

离来评估能力估计精度。其中 $l(l=0, 1, 2)$ 维度上的 ABS 和 RMSE 分别为： $ABS_l = (1/N) \cdot \sum_{i=1}^N |\hat{\theta}_{il} - \theta_{il}|$ 和 $RMSE_l = (1/N) \cdot \sum_{i=1}^N (\hat{\theta}_{il} - \theta_{il})^2$ 。一般地，ABS、RMSE 和 EDU 的值越小，表明能力估计越准确。

5 结果

表 1、2 和 3 分别统计了低、中和高双因子模式题库中各能力维度估计值与真值的相关、均方根误差和绝对值偏差。

固定任一选题方法时，双因子模式越强，全局因子 G 的相关系数越低、RMSE 和 ABS 的值越大；而局部因子 S_1 和 S_2 的相关越高、RMSE 和 ABS 的值越低。换句话说，双因子模式越强，全局因子的估计精度越低，局部因子的估计精度越高。这与 Seo 和 Weiss（2015）的结论一致。因为模拟实验中，全局因子在项目上的区分度分布保持不变，而局部因子在项目上的区分度的值随双因子模式的增强而增大。于是，双因子模式越强，项目对局部因子提供的测量信息越多，局部因子的估计精度越高。双因子模式越弱，项目则主要考察全局因子。因此全局因子的估计精度随双因子模式的增强而有所降低。

固定双因子模式时，无论是全局因子还是局部因子都有（1）PKL 方法中相关系数的值明显低于其它方法的结果，其它方法的相关系数没有显著差异；（2）PKL 方法的 ABS 和 RMSE 显著大于其它方法的结果；（3）CEM 方法的 ABS 和 RMSE 总是小于其它方法的结果；（4）贝叶斯 D-优化、PDO 和 MI 方法的 ABS 和 RMSE 没有明显差异。总体上，依据相关、RMSE 和 ABS 推知 CEM 的测量精度最高，PKL 方法的能力估计精度最低。PDO 方法，DO 方法和 MUI 方法的结果相似，稍次于 CEM 方法的结果。

图 1 为不同因子模式下，各种项目选择方法在测验长度为 5 的倍数时欧氏距离的折线图。40 个项目共记录 8 个欧氏距离。由图 1 可知，（1）双因子模式越强，相同方法在同一测验长度处的欧氏距离越小。换句话说，双因子模式越强，题库项目在局

表 1 各种方法在低、中和高双因子模式题库中能力估计值与真值之间的相关

方法	低-双因子模式			中-双因子模式			高-双因子模式		
	θ_0	θ_1	θ_2	θ_0	θ_1	θ_2	θ_0	θ_1	θ_2
DO	.975	.742	.700	.967	.858	.873	.963	.880	.882
PDO	.974	.740	.740	.969	.864	.877	.969	.884	.878
PKL	.973	.742	.663	.961	.866	.837	.961	.875	.872
MUI	.973	.755	.705	.967	.856	.872	.963	.885	.881
CEM	.974	.707	.751	.971	.878	.877	.964	.879	.882

表 2 各种方法在低、中和高双因子模式下 RMSE 的结果

方法	低-双因子模式			中-双因子模式			高-双因子模式		
	θ_0	θ_1	θ_2	θ_0	θ_1	θ_2	θ_0	θ_1	θ_2
DO	.225	.683	.719	.254	.496	.493	.274	.470	.470
PDO	.229	.684	.677	.249	.486	.487	.262	.463	.465
PKL	.232	.684	.754	.278	.511	.552	.280	.480	.489
MUI	.232	.668	.714	.254	.500	.494	.271	.461	.472
CEM	.228	.706	.665	.240	.463	.486	.260	.459	.452

表 3 各种方法在低、中和高等双因子模型下的平均绝对值偏差条件统计表

方法	低-双因子模式			中-双因子模式			高-双因子模式		
	θ_0	θ_1	θ_2	θ_0	θ_1	θ_2	θ_0	θ_1	θ_2
DO	.179	.55	.573	.204	.396	.391	.212	.392	.359
PDO	.183	.552	.544	.201	.389	.386	.207	.366	.362
PKL	.184	.543	.599	.223	.411	.438	.223	.38	.385
MUI	.184	.541	.574	.204	.397	.389	.214	.352	.39
CEM	.18	.556	.533	.19	.367	.384	.207	.327	.37

部因子维度上的区分度提高,项目质量更优。因而,整体能力水平的估计精度提高。(2)固定双因子模式时,PKL方法的欧氏距离最大,CEM方法的欧氏距离最小,MUI、PDO和D-优方法的欧氏距离没有明显差异。模拟实验中测验长度为40,测验初期能力估计不准确,各种方法之间的差异不明显。随着测验的进行,不同方法的特征慢慢突显,当测验长度大于20时,不同方法的欧氏距离差异更加明显。

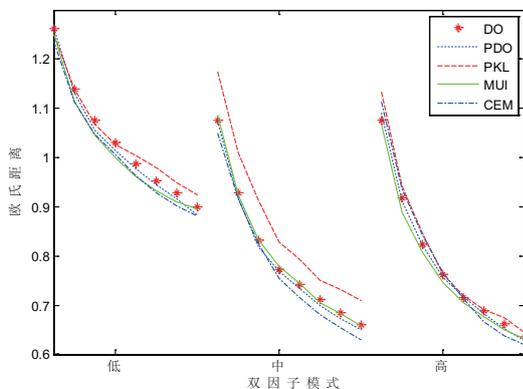


图 1 低、中和高双因子模式下欧氏距离图

6 结论和讨论

双因子模型作为一类特殊的多维模型,已广泛应用于分析心理量表、病人自我报告、教育调查以及教育评估测验获得的数据。双因子项目反应理论模型的提出和CAT技术的成熟与实践,进一步推动了双因子模型的应用。目前,双因子模型MCAT的研究还很少,主要针对二级评分项目背景下考察项目选择问题。多级评分项目作为一种重要的项目类型,其项目选择计算往往比二级评分项目更复杂。

特别是MCAT中,具有前途的项目选择方法(如:Kullback-Leibler信息量、互信息、连续熵)往往因测验考察维度数量的增加而更加复杂。

本文针对多级评分项目,推导了双因子等级反应模型Fisher信息量的计算,然后考察双因子模型“维度缩减”方法在项目选择方法中的应用。结果发现,对PKL、CEM和MI方法而言,“维度缩减”方法对能力的积分运算进行恰当重组,将高维积分简化为多个二维迭代积分,从而降低计算复杂度,缩短计算时间。研究中2000名被试作答40个项目,运用D-优化方法选题仅需1小时,其它方法也只需要两小时。平均来讲,每个被试完成40题需要4秒时间,这完全满足CAT对运行时间的要求。其次,CEM方法的估计精度最高,PKL方法的测量精度最低。PDO、DO方法和MUI方法的精度相似,稍次于CEM方法的结果。这与MCAT中研究结论有许多一致之处(Lin, 2012; Wang & Chang, 2011)。

值得注意的是,本文采用模拟研究方式,项目参数和被试特征不能完全反映实际情况。今后还有待在其它实验条件开展CAT研究,建构双因子模型题库开展实证研究。另外,MAP,MLE和EAP能力估计方法具有不同特点,考察他们是否影响不同方法的选题结果也是今后有意义的研究问题。此外,双因子模型CAT作为一个重要的应用领域还有很多问题值得进一步研究。例如,Zheng等(2013)只考察了修正的内容约束方法在实际题库中的表现,研究受到实际题库容量小的限制。今后还有待在容量更大的题库中考察多种内容约束方法的表现。又如,针对包含二级和多级评分项目的混合测验而言,

探索最优测验设计如多级评分项目的施测比例和呈现方式等也都是有意义的研究问题。

参考文献

- Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*(3), 221–248.
- De Ayala, R. J., Dodd, B. G., & Koch, W. R. (1992). A comparison of the partial credit and graded response models in computerized adaptive testing. *Applied Measurement in Education, 5*(1), 17–34.
- DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145–168.
- Furthner, M. R., Rauthmann, J. F., & Schse, P. (2015). Unique self-leadership: A bifactor model approach. *Leadership, 11*(1), 105–125.
- Gibbons, R. D., Rush, A. J., & Immekus, J. C. (2009). On the psychometric validity of the domains of the PDSQ: An illustration of the bi-factor item response theory model. *Journal of Psychiatric Research, 43*(4), 401–410.
- Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*(1), 41–54.
- Huebner, A. R., Wang, C., Quinlan, K., & Seubert, L. (2016). Item exposure control for multidimensional computer adaptive testing under maximum likelihood and expected a posteriori estimation. *Behavior Research Methods, 48*(4), 1443–1453.
- Lin, H. Y. (2012). *Item selection methods in multidimensional computerized adaptive testing adopting polytomously-scored items under multidimensional generalized partial credit model*. Unpublished doctoral dissertation of University of Illinois at Urbana-Champaign.
- Samejima, F. (1976). Graded response model of the latent trait theory and tailored testing. In C. K. Clark (Ed.), *Proceedings of the first Conference on Computerized Adaptive Testing* (pp. 5–17). Washington, DC: U. S. Government Printing Office.
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*(2), 331–354.
- Seo, D. G. (2011). *Application of the bifactor model to computerized adaptive testing*. Unpublished doctoral dissertation of The University of Minnesota.
- Seo, D. G., & Weiss, D. J. (2015). Best design for multidimensional computerized adaptive testing with the bifactor model. *Educational and Psychological Measurement, 75*(6), 954–978.
- van Rijin, P. W., Eggen, T. J. H. M., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26*(4), 393–411.
- Wang, C., & Chang, H. H. (2011). Item selection in multidimensional computerized adaptive testing—gaining information from different angles. *Psychometrika, 76*(3), 363–384.
- Watkins, M. W., & Beaujean, A. A. (2014). Bifactor structure of the wechsler preschool and primary scale of intelligence—fourth edition. *School Psychology Quarterly, 29*(1), 52–63.
- Weiss, D. J., & Gibbons, R. D. (2007). *Computerized adaptive testing with the bifactor model*. Paper presented at the New CAT Models session at the 2007 GMAC Conference on Computerized Adaptive Testing.
- Wiesner, M., & Schanding, G. T. (2013). Exploratory structural equation modeling, bifactor models, and standard confirmatory factor analysis models: Application to the BASC-2 behavioral and emotional screening system teacher form. *Journal of School Psychology, 51*(6), 751–763.
- Yao, L. H. (2014). Multidimensional item response theory for score reporting. In Cheng, Y., & Chang, H. H. (Eds.), *Advances in modern international testing: Transition from summative to formative assessment*. Charlotte, NC: Information Age.
- Zheng, Y., Chang, C. H., & Chang, H. H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research, 22*(3), 491–499.

Comparison of Bifactor MCAT Item Selection Criteria for Polytomous Items

Mao Xiuzhen, Liu Huan, Tang Qian

(Institute of Educational Science, Sichuan Normal University, Chengdu, 610066)

Abstract Bifactor model assumes that the test involves a general factor and multiple group factors. Numerous analyses on the structures of psychological trait measurement, school education survey, medical survey, and diagnostic testing have shown that the bifactor model could well represent the construct structures of the tests, surveys, or scales, and it has shown better model-data fit than other competing models (e.g. unidimensional, higher-order, and correlation models). The bifactor CAT has proved to be a practical approach that could substantially reduce the burden of respondents while increasing testing efficiency (Gibbons, et al., 2007). However, the number of dimensions in multidimensional CAT usually becomes an obstacle to the application of many famous item selection method, especially for the polytomous items.

This study focused on the formula of information matrix for polytomous items and how to simplify the computation of item selection method using the dimension reduction method. First, the Fisher information for bifactor grade response model was derived; Then, the dimension reduction method was applied to the computation of item selection methods including the posterior weighted Fisher D-optimality method (PDO), the posterior weighted Kull-Leilber information method (PKL), the continuous entropy method (CEM), and the mutual information method (MI); Last, these methods were then compared with simulated data under three different bifactor pattern designs, using the original D-optimality method as the baseline. We conducted Monte Carlo simulation using a MATLAB program (R2010a) to write the CAT code and evaluate different item selection methods in terms of the correlation between real and estimated abilities, root mean squared error, absolute deviation, and Euclidean distance.

The results showed that: (1) The information of the bifactor graded response model is easily obtained and it is the generation of the information of the 2-parameter logistic model; (2) Simulation results showed that for each item selection method, the correlation in high bifactor pattern was the highest, the root mean square and the absolute was lowest; (3) Under the same simulation condition, the CEM item selection method produced the highest correlation of real ability and estimated ability, lowest root mean square, absolute bias and Euclidean distance among all the item selection methods while the PKL method performed the worst according to these indices; (4) The PDO, MI and DO methods produced very similar results when fixing the test condition; (5) The Euclidean distance of all the methods showed that their difference became significant when the test length was larger than 20 items.

In conclusion, the dimension reduction method can be easily used to simplify the computation of item selection methods including PDO, PKL, CEM and MI. This method can simplify the multidimensional integration contained in each method to multiple 2-dimensional integrations. The simulation results further show that when the difference between the discrimination parameters of the group factors and those of the general factor are smaller, estimates of the group factors become more accurate and vice versa for the estimates of the general factor. Some problems like controlling the exposure rate, meeting the content constraint and item selection for mix-form test are valued to be explored further.

Key words multidimensional item response theory, bifactor graded response model, computerized adaptive testing, item selection criteria, measurement precision