心理学探新 2019,Vol.39,No.1,47-56
PSYCHOLOGICAL EXPLORATION

多维计算机化自适应测验中项目曝光控制选题策略的比较*

毛秀珍 王娅婷 杨 睿
(四川师范大学教育科学学院,成都 610066)

摘 要:在MCAT中考查四种项目选择指标在有无曝光控制条件下的选题表现。项目选择指标分别是:(1)贝叶斯的D优化方法(D-optimality)、后验期望Kullback-Leibler方法(KLP)、基于等权重复合分数的最小误差方差方法(the minimized error variance of the linear combination score with equal weight,V1)和基于最优权重复合分数的最小误差方差方法(the minimized error variance of the composite score with optimized weight,V2)。将针对认知诊断CAT项目曝光控制的的限制阈值方法(Restrictive Threshold,RT)和限制进度(Restrictive Progressive,RPG)方法、单维CAT中的最大优先指标方法(Maximum Priority Index,MPI)推广到MCAT。模拟研究表明:(1)KLP,D-优化和V1对领域分数估计准确,能力返真性比V2更好。(2)尽管V1和V2方法相比KLP和D-优化方法提高了题库利用率,但这四种选题指标都产生不均匀的项目曝光率分布。(2)三种曝光控制策略都极大地提高项目曝光均匀性,且不明显降低测量精度。(3)MPI与RPG方法在曝光控制方面表现类似,且比RT的方法表现更好。

关键词:多维项目反应理论;计算机化自适应测验;选题方法;测量精度;项目曝光率
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003-5184(2019)01-0047-10

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,打破了千人一卷的考试模式,真正实现了测验的"量体裁衣",极大地提高了测验效率。CAT与传统纸笔测验相比,因其效率更高、测验时间更短、测验参加者更少的压力等优势使其受到广大实践者和研究者的青睐。CAT的另一个特征是可以运用各种项目反应模型开展测验。例如,单维项目反应理论模型(Unidimensional Item Response Theory,UIRT)、多维项目反应理论(multidimensional IRT,MIRT)模型、认知诊断模型以及展开模型等等。

MCAT兼具MIRT和CAT的优点,在实践中突显了测验的高效、快捷和诊断功能。一方面,基于不同条件的大量研究都一致表明MCAT与单维CAT相比具有更高的测验精度和测量信度,换句话说在达到相似测量精度的条件下能大大减少测验长度、缩短测验时间。例如,Segall(1996)基于军队服务职业倾向题组测验(ASVAB)数据的9个自适应模拟研究发现MCAT与单维CAT在达到相似或更高测量精度的情况下减少了约三分之一的测验项目。又如,Luecht(1996)研究表明具有内容约束的情况下,MCAT可以减少25%~40%的测验项目。再如,Wang和Chen(2004)在不同潜特质相关、潜特质数量和评分水平的情况下表明MCAT的测验效率比单维CAT更高。另外,MCAT同时估计被试在多个维度上的能力水平,提供关于领域能力和总体能力的详细诊断信息。正是因为MCAT具有高效和多维性优点,使得它比单维CAT更适用于实际测验。因此,许多MCAT研究都基于实际测验如TerraNova(Yao,2010),美国大学入学考试(ACT)(Veldkamp & vanderLinden,2002)和ASVAB(Segall,1996;Yao,2012,2014a)等。

自Bloxom和Vale(1987)将UCAT推广到多维以来,MCAT越来越受到研究者的关注,其相关研究在最近几年取得了突破性进展。由于选题策略对测验质量和测量精度具有重要影响,从而成为当前研究热点。因此,大多数研究者关注提出新的项目选

* **基金项目**:国家自然科学基金青年项目(31400897)。
通讯作者:毛秀珍,E-mail:maomao_wanli@163.com。

择指标以减少能力估计的误差。Yao(2014a)研究表明大部分项目选择方法总是倾向于选择特定类型的项目,导致项目曝光不均匀。她还以 Kullback - Leibler 指标为例,指出该方法倾向于选择所有维度具有高区分度的项目或者不同维度之间区分度相差较大的项目。又如,D - 优化方法倾向于选择在某一维度具有高区分度的项目(Wang, Chang, & Boughton,2011)。目前,CAT 已广泛应用于多种测验。因此,控制项目曝光率在 MCAT 应用中极其重要,尤其是在高风险测验中的应用。此外,在 MACT 中关于控制项目曝光率的研究很少。因此,本文的目的是比较 MCAT 中多种曝光控制方法的表现。

至今,单维 CAT 中的许多项目曝光控制方法已经推广到 MCAT 情景。例如,Finkelman, Nering 和 Roussos(2009)将 Sympson - Hetter(S - H)(Sympson & Hetter,1985)和 Stocking - Lewis(S - L)(Stocking & Lewis,1998)方法推广到 MCAT。他们发现 SH 方法、推广的 SH 方法和推广的 SL 方法都能较好地控制最大项目曝光率,但还存在较多曝光不足的项目。另外,它们都需要较长的时间事先模拟来确定曝光率控制参数。另外,Yao(2014a)比较了 S - H 方法和固定曝光率程序。固定曝光率程序类似于 Cheng 和 Chang(2009)针对单维 CAT 中提出的最大优先指标方法(MPI)。她指出,S - H 方法的测量精度更高,固定比率方法的题库利用率更高,项目曝光更均匀。

Lee,Ip 和 Fuh(2008)借鉴 UCAT 中 a - 分层方法的思想,提出按区分度向量 $a = (a_{j1}, a_{j2})$ 的函数 $|a_{j1} - a_{j2}|$ 对题库分层的项目选择方法,结果表明该方法能提高大部分曝光过低项目的使用率,显著降低卡方值。但这种方法不能保证没有过度曝光的项目。因此,Huebner, Wang, Quinlan 和 Seubert(2015)将按该方法与项目合格方法(van der Linden & Veldkamp,2007)结合来增强项目曝光平衡性。这种组合方法提高大部分曝光率较低的项目的使用率,同时控制最大项目曝光率,但它只适用于二维能力空间。对更高维度的情况下建构合适的项目区分度参数的函数是今后的一个重要研究问题。

众所周知,项目曝光率均匀性受到过度曝光和曝光不足项目数量的影响。在上述曝光控制方法中,S - H 方法、推广的 S - H 方法、推广的 S - L 方法、固定曝光率和项目合格性方法在控制最大项目曝光率方面表现良好;按 $|a_{j1} - a_{j2}|$ 对题库分层的项目选择方法能有效提高曝光率较低项目的使用率。虽然 Huebner 等(2015)使用的组合方法在两个方面都表现良好,但它只适合于二维能力空间。

MCAT 在实践应用中,特别是应用于高风险测验时,项目曝光均匀性和测量精度是需要考虑的两个重要问题。因为二者总是相互抵消,实践者希望找到能保证测量精度且能平衡项目曝光均匀性的项目选择方法。然而,没有很好的方法能有效的平衡高维测验的项目曝光率。Wang, Chang 和 Huebner(2011)报告限制进度(Restrictive Progressive, RPG)方法和限制阈值(Restrictive Threshold, RT)方法在认知诊断 CAT 中能很好地平衡项目曝光率。另外,目前没有研究考察它们在 MCAT 中的表现。因此,本文的目的是考察它们在 MCAT 中能否控制最大项目曝光率且提高曝光不足项目的使用率,同时不显著损失测量精度,并进一步比较它们和 MPI 方法的表现。

第二部分将介绍采用的 MIRT 模型和能力估计方法,第三部分介绍项目选择指标和曝光率控制方法,接下来的三个部分分别是研究设计、结果、结论和讨论。

2 选用的多维项目反应理论模型和能力估计方法

2.1 多维两参数逻辑斯蒂克模型(Multidimensional two parameters Logistic Model, M - 2PL)

MIRT 模型按完成任务时某一能力维度上的不足是否可以被其它优势能力所补偿分为补偿模型和非补偿模型。Bolt 和 Lall(2003)指出二者能很好拟合非补偿模型产生的数据,但是非补偿模型不能很好地拟合补偿模型产生的数据。目前,大部分研究选用补偿的二级评分模型((van der Linden,1999; Veldkamp & van der Linden,2002; Mulder & van der Linden,2010)。鉴于补偿模型的优势,M - 2PL 模型将用于模拟被试的作答。

M - 2PL 模型(McKinley & Reckase,1982)中项目 j 包括斜率(截距)参数 b_j 和区分度向量 $a_j = (a_{j1}, a_{j2}, \cdots, a_{jD})^T$,其中 T 表示转置,D 表示测验的维度。那么,能力为 $\theta = (\theta_1, \theta_2, \cdots, \theta_D)^T$ 的被试正确作答项目 j 的概率可以根据(1)式计算而得。

$$P_j(\theta) = P(x_j = 1 \mid \theta, a_j, b_j) = \frac{1}{1 + \exp\left[-(a_j^T \cdot \theta - b_j)\right]}. \quad (1)$$

上式中 $a_j^T \cdot \theta - b_j = \sum_{l=1}^{D} a_{jl} \cdot \theta_l - b_j$ 表示二维空间中的直线。它表明所有使 $a_j^T \cdot \theta$ 相等的能力向

量 θ 具有相同的正确作答概率,体现了模型的补偿特征。

2.2　最大后验能力估计方法(maximize a posterior estimation,MAP)

Yao(2014b)研究比较了 MAP、期望后验估计方法(expected a posteriori,EAP)和极大似然估计方法(maximum likelihood estimation,MLE),并指出(1)MLE 方法产生更小的偏差和更大的均方误差根,然而 MAP 和 EAP 运用强先验或标准先验都得到更高精度的能力估计值;(2)MAP 和 EAP 表现类似,但是 EAP 的运行时间比 MAP 更长。最近,Huebner 等(2015)在 MCAT 中比较了 EAP 和 MLE 方法,并证明 EAP 在能力估计方面总是产生更稳定的结论和更低的均方误。基于上述研究和估计精度以及计算简便性考虑,本文采用 MAP 方法估计能力。

令 θ 的先验密度函数 $f(\theta)$ 服从均值为 μ_0、协方差矩阵 Σ_0 的多变量正态分布。记项目 j 的作答反应为 x_j,已做答 k-1 个项目的反应为 X_{k-1}。基于贝叶斯定理,有 $f(\theta \mid X_{k-1}) \propto L(X_{k-1} \mid \theta) \cdot f(\theta)$,$L(X_{k-1} \mid \theta)$ 代表似然函数。MAP 的目标就是找到后验密度函数 $f(\theta \mid X_{k-1})$ 的众数。也就是说,能力估计值 $\hat{\theta}^{MAP}$ 等价于求解

$$\frac{\partial \log f(\theta \mid X_{k-1})}{\partial \theta_l} = 0(l = 1,2,\cdots,D).$$

。Newton-Raphson 迭代方法可以用于求解这个方程组,详细方法请参见 Yao(2014b)。

3　项目选择指标和曝光控制策略

下文统一用 N 表示被试人数,L 代表测验长度,M 表示题库容量。$S_{k-1} = \{i_1, i_2, \cdots, i_{k-1}\}$ 表示已施测项目的集合,$R_k = \{1,2,\cdots,M\}/S_{k-1}$ 表示选择第 k-1 个项目之后的剩余题库。

3.1　项目选择指标

基于计算复杂性和运行时间的考虑,本文选用贝叶斯 D-优化方法(Mulder & van der Linden,2009)、后验期望 KullBack-Leibler 信息方法(VelderKamp & van der Linden,2002)、基于等权重复合能力最小误差方差方法和基于最优权重复合能力最小误差方差方法(Yao,2012)四种项目选择指标。

3.1.1　贝叶斯 D-优化方法

MIRT 中项目 Fisher 信息量不再是一个实数,而是一个矩阵。特别地,M-2PL 中的项目 j 的 Fisher 信息量为:

$$I_j(\theta) = P_j(\theta) \cdot (1 - P_j(\theta)) \cdot (a_j^T a_j). \quad (2)$$

施测 k-1 个项目后,能力估计值构成一个椭圆(球)V_{k-1}。于是,为使施测第 k 个项目后,V_k 的面积(体积)下降最快,Segall(1996)提出选择使后验测验 Fisher 信息矩阵行列式值最大的方法,又称为 D-优化方法(Mulder & van der Linden,2009)。具体而言,该方法的选择标准可以表示为:

$$D_k = \max \{| I_{k-1}(\hat{\theta}^{k-1}) + I_j(\hat{\theta}^{k-1}) + \Sigma_0^{-1} |, j \in R_{k-1}\}. \quad (3)$$

其中,$I_{k-1}(\hat{\theta}^{k-1})$ 表示已施测项目集在能力估计值处的测验信息量,$I_k(\hat{\theta}^{k-1})$ 表示候选项目在能力估计值处项目信息量。

3.1.2　后验期望 KL 信息量方法(Posterior Expected KL information,KLP)

KLP 方法是通过对根据能力的后验分布信息对项目 KL 信息进行加权而获得。也就是,第 k 个项目根据下式来选择

$$KLP_k = \max \{\int_\theta KL_j(\hat{\theta}^{k-1}, \theta) \cdot f(\theta \mid X_{k-1}) d\theta, j \in R_{k-1}\} \quad (4)$$

其中,

$$KL_j(\hat{\theta}^{k-1}, \theta) = E_\theta \log \left[\frac{P_j(x_j \mid \theta, a_j, b_j)}{P_j(x_j \mid \hat{\theta}^{k-1}, a_j, b_j)}\right] = P_j(\theta) \log \frac{P_j(\theta)}{P_j(\hat{\theta}^{k-1})} + (1 - P_j(\theta)) \log \frac{(1 - P_j(\theta))}{(1 - P_j(\hat{\theta}^{k-1}))}. \quad (5)$$

为简化计算,通常将缩小积分区间,得到(11)式。

$$KLP_k = \max \{\int_{\theta_1^{k-1}-\gamma_j}^{\theta_1^{k-1}+\gamma_j} \cdots \int_{\theta_D^{k-1}-\gamma_j}^{\theta_D^{k-1}+\gamma_j} KL_j(\hat{\theta}^{k-1}, \theta) \cdot f(\theta \mid X_{k-1}) d\theta_1 \cdots d\theta_D, j \in R_{k-1}\}, \quad (6)$$

其中 4 γ_j 等于 $3/\sqrt{j}$。因此,KLP 方法中,第 k 个项目选自使(10)式取最大值的项目。

3.1.3　基于相等权重复合能力最小误差方差方法(V_1)

van der Linden(1999)给出二维能力空间中计算复合能力估计方差的方法,然后提出第 k 个项目应选择使复合分数具有最小误差方差的项目。Yao(2012)进一步指出对 D 维线性复合能力 $\theta_\alpha = \sum_{l=1}^{D} \theta_l w_l$ 而言,在施测 k-1 个项目后,θ_α 的测量标准误为 $SEM(\theta_\alpha) = (V(\theta_\alpha))^{1/2} = (wV(\theta)w^T)^{1/2}$,其中 $V(\theta)$ 的值通常由 $I(\theta)^{-1}$ 来逼近。于是,V_1 方法设置所有权重 $w = (1/D, 1/D, \cdots, 1/D)$,那么

第 k 个项目将在剩余题库中选择使 $SEM(\theta_\alpha)$ 取值最小的项目。

3.1.4 基于最优权重复合能力最小误差方差方法（V_2）

V_2 与 V_1 方法不同的是，V2 中 θ_α 不是领域能力相等权重的线性组合，而是基于最优权重的线性组合。根据已施测项目信息量计算使复合能力估计误差最小的权重，称为最优权重。具体而言，Yao（2012）通过数理证明了使 $SEM(\theta_\alpha) = (wV(\theta)w^T)^{1/2}$ 取最小值的权重存在，而且该权重为

$$w = \frac{1}{\sum_{o=1}^{D}\sum_{l=1}^{D} b_{ol}} \cdot [1,1,\cdots,1]_{1\times D} \cdot I_{k-1}(\theta) \tag{7}$$

其中，b_{ol} 表示 $I_{k-1}(\theta)$ 的第 o 行 l 列的元素。因此，V_2 方法在选择每个项目之前根据已施测项目计算在当前能力估计值处的 Fisher 信息量矩阵并计算最优权重；然后在剩余题库中选择使 $SEM(\theta_\alpha)$ 值最小的项目。

3.2 项目曝光率控制策略

Wang 等（2011）提出的限制阈值方法（RT）和限制进度指标方法（RPG）是在认知诊断 CAT 中表现较好的两种项目曝光控制方法，下面将它们推广到 MCAT 情景。

3.2.1 RT 方法

该方法在每个被试参加测验之前将曝光率大于预先设定的最大值的那些项目从题库中去掉后形成一个影子题库，然后第 $k(k=1,2,\cdots,L)$ 个项目将从由剩余题库中项目选择指标值落在最大（最小）值的一个较小区间内的项目所构成候选项目集中随机选择。例如，当按 D－优化方法和 KLP 方法选题时，候选项目集由信息量落在区间 $[\max(Index)-\delta, \max(Index)]$ 的项目构成；当按 V_1 和 V_2 方法选题时，则将选题指标值落在区间 $[\min(Index), \min(Index)+\delta]$ 的项目放在一起构成候选项目集。其中 $\delta = [\max(Index) - \min(Index)] * (1 - k/L)^{\beta}$，L 为测验长度。$\beta$ 的值越大，δ 越小，测量越准确，项目曝光均匀性越差。因此，β 是平衡项目曝光分布和测量精度的权重，其值可根据测验要求灵活设置，本文令 β 等于 0.5。

3.2.2 RPG 方法

MCAT 中当采用 D－优化指标和 KLP 指标选题时，RPG 方法将根据式（8）选择第 k 题（Wang et al.，2011）：

$$i_k = \max\{(1-er_j/r^{\max})\cdot[(1-k/L)u_j + Index_j \times \beta k/L], j \in S_{k-1}\} \tag{8}$$

对 V_1 和 V_2 方法，则按式（9）选题：

$$i_k = \max\{(1-er_j/r^{\max})\cdot[(1-k/L)u_j + (C - Index_j) \times \beta k/L] j \in S_{k-1}\} \tag{9}$$

er_j 与 $r^{\max}$ 分别表示项目 j 的曝光率和期望项目曝光率，为了统一方向，常数 C 必须大于所有项目在复合能力处的估计误差，本文令 C 等于 10000。实验发现 SEM 总是在前几个项目很大，但是很快就下降到 1000 以下。因此，最好将 C 的值设置为大于 1000。令 H^* 等于剩余题库中所有项目信息量的最大值，那么 u_j 均匀取值于区间 $(0, H^*)$，β 是平衡项目曝光控制和测量准确性的权重，本文取 $\beta = 0.5$。

3.2.3 最大优先指标方法

根据 Yao（2014b），容易得到项目 j 基 D－优化方法和 KLP 方法选题时量的优先指标（Priority Index，PI）为：

$$PI_j = \frac{r^{\max} - n_j/N}{r^{\max}} \cdot Index_j, \tag{10}$$

n_j 为第 j 个项目被调用的次数，$Index_j$ 表示 D－优化或 KLP 指标，MPI 方法的任务就是找到使 PI 值最大的项目。对 V_1 和 V_2 方法，PI 指标相应地变为：

$$PI_j = \frac{r^{\max} - n_j/N}{r^{\max}} \cdot (C - Index_j) \tag{11}$$

C 的含义和值与 RPG 方法的相同，$Index_j$ 表示 V_1 或 V_2 指标。

4 方法

本文采用 MATLAB（R2010a）为工具编写 MCAT 代码，进行模拟实验。

4.1 模拟研究的设计方法

4.1.1 题库的模拟

尽管 Stocking（1994）建议题库应包含测验长度 12 倍以上的项目，MCAT 的大部分研究都采用了较为严格的题库。例如，van der Linden（1999）的实验中针对测验长度为 50 的 MCAT 使用包含 500 个项目的题库；Lee 等（2008）的研究中题库包含 480 个项目，测验长度为 30 和 60 两种情况；在 Veldkamp 和 van der Linden（2002）的研究中对测验长度为 30 的 MCAT 测验使用仅包含 200 个项目的题库。鉴于此，本文固定测验长度为 30，模拟产生包含 450 个项目的题库。

为简化实验条件，大部分研究都假设测验考察 2 或 3 个维度利用 M－2PL 或 M－3PL 产生项目参数和

做答反应(van der Linden,1999;Veldkamp & van der Linden,2002;Lee et al.,2008;Mulder & van der Linden,2009;Finkelman et al.,2009;Wang, Chang, & Boughton,2013;Wang & Chang,2011)。因此,不失一般性,本研究假设测验考察三个维度,利用 M-2PL 产生数据,并借鉴 Yao 和 Schwarz(2006),Wang 和 Chang(2011)等人的方法确定项目参数。对项目 j,每个维度的区分度从对数正态分布中产生,即 $(a_{j1},a_{j2},a_{j3})\,j(j=1,2,\cdots 450)$ 独立产生于 $\log N(0,0.5)$,项目难度 $b_j(j=1,2,\cdots 450)$ 从标准正态分布中随机产生,项目猜测参数均设置为 0。

4.1.2　模拟被试的真实能力水平和作答反应

借鉴 Wang 和 Chang(2011),Yao,Pommerich 和 Segall(2014),Wang 等(2013)研究,研究从多变量正态分布中模拟产生 5000 名被试。其中能力均值为 [0,0,0],考虑三种相关水平,并假设方差协方差矩阵为:$\begin{pmatrix}1&\rho&\rho\\ \rho&1&\rho\\ \rho&\rho&1\end{pmatrix}(\rho=0.3,0.6,0.8)$。运用 M2PL 模型计算被试 i 正确作答项目 j 的概率 P_{ij},并产生(0,1)区间的随机数 p_{ij}。如果 P_{ij} 大于 p_{ij},则被试 i 在项目 j 上的反应为 1,否则为 0。

4.1.3　能力估计方法

初始能力估计值从多变量标准正态分布中随机产生。假设能力先验分布为多变量标准正态分布,利用 MAP 方法估计能力值。

4.1.4　项目选择方法和测验终止规则

本文考查了四种项目选择指标:D-优化指标、KLP、V_1 和 V_2 在与项目曝光控制方法相结合前后的选题结果。项目曝光控制方法是 MPI、RT 和 RPG。测验长度固定为 30。

4.1.5　评价指标

每个能力维度的平均偏差与均方差用于表示能力估计精度,它们通过下面的式子计算。

$$Bias_l=\frac{1}{N}\cdot\sum_{i=1}^{N}(\hat{\theta}_l-\theta_l)(l=1,2,3)\quad(12)$$

$$MSE_l=\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_l-\theta_l)^2(l=1,2,3)\quad(13)$$

项目曝光率即项目的使用频率。本文选用未使用的项目个数、过度曝光的项目个数(即曝光率大于 0.20 的项目个数)、χ^2 统计量和测验重叠率评价各项目曝光率的结果。其中,$\chi^2=\sum_{i=1}^{N}[(er_i-\overline{er_i})^2/\overline{er_i}]$ 表示项目观察曝光率和期望曝光率之间的差异(Chang & Ying,1999)。项目 i 的期望曝光率 $\overline{er_i}$ 等于测验长度 L 除以题库容量 M。χ^2 越小,总体上项目观察曝光率与期望曝光率之间的差异越小。测验重叠率定义为随机选择的两个被试之间期望重叠的项目个数与测验长度之比。假设有 N 个被试参加长度为 L 的测验,它可以通过公式(21)(Chen,Ankenmann,& Spray,2003)计算,

$$\hat{T}=\frac{N}{L}S_{er}^2+\frac{L}{N}.\quad(14)$$

其中,S_{er}^2 表示项目曝光率的方差。测验重叠率越小,项目曝光控制越好。

5　模拟数据的结果

5.1　能力估计结果

由于每种方法在任意两个维度估计值的偏差相差极小,图 1 展示了三个维度的平均偏差。图 2 展示了各种相关水平下每个维度的 MSEs。根据图 1 和图 2 很容易得知:(1)D-优化、V1 方法和 V2 方法得到相似的估计偏差,且比 KLP 方法的偏差更大;(b)对每个维度的 MSE,KLP 方法的值最小,接下来是 D-优化、V1 方法和 V2 方法。总体上,KLP 方法的测量精度明显高于其它三种方法的结果,D-优化方法次之,V2 方法表现最差。

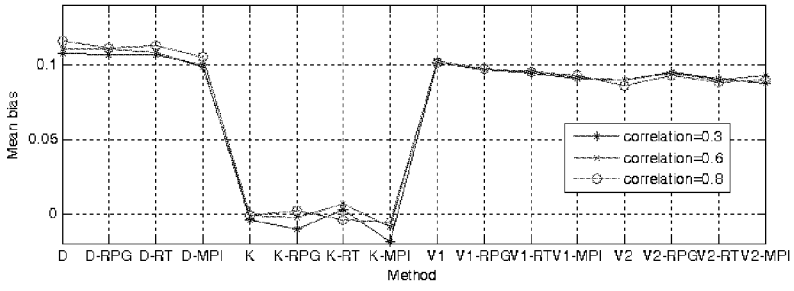

图 1　各种实验条件下每种方法在三个维度的平均偏差值

项目曝光控制方法对测量精度的效应通过以下几个方面进行检测。首先,从图 1 可知,固定项目选择方法,当与不同曝光控制方法结合选题前、后的偏差几乎没有差异。因此,项目曝光控制方法不会显著影响测量偏差。其次,根据图 2,比较各指标与曝光控制方法相结合前、后选题的测量 MSE,可发现除 V2 方法外,所有项目曝光控制策略都增加了 MSE 的值。V2 方法的 MSE 比 V2 - RT 方法的 MSE 更大。从下面的结果可知 V2 方法本身能提高题库利用率和项目曝光均匀性,这也使其在一定程度上降低了测量精度。总体上讲,结合曝光控制策略选题总会降低测量精度。

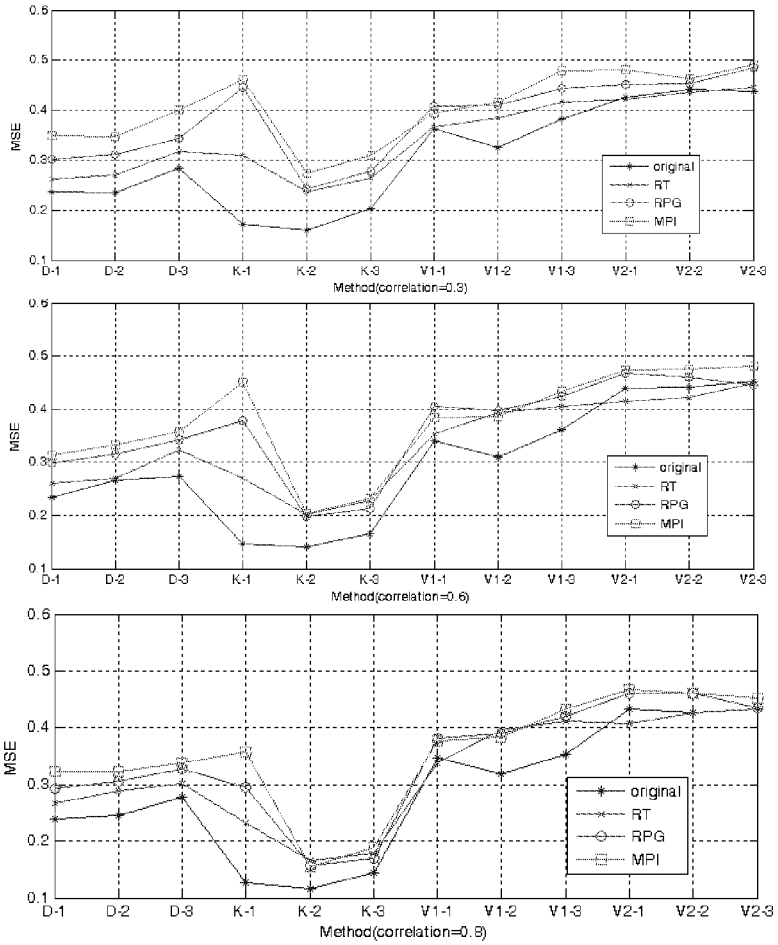

图 2 各种方法在每种实验条件下各个维度的 MSE 值

注:Original = 没有结合项目曝光控制方法的选题指标;D = D - optimality;K = KLP;'-1','-2',and '-3'代表第一、第二和第三维度。

第三,当曝光控制方法和 D - 优化方法,KLP 方法或者 V2 方法相结合时,他们在测量精度方面具有所差异。然而,当与 V1 方法相结合时,所有曝光控制方法都产生类似的测量精度。除此之外,能力相关越高,同一选题指标与项目曝光控制方法结合前后在测量精度的差异越低。

最后,比较不同项目曝光控制方法的结果可知 RT 方法总是产生最低的 MSE。因此,它比 RPG 和 MPI 方法的测量精度更高。尽管 RPG 和 MPI 方法在不同项目选择指标下表现的优劣次序有所波动,总体上二者的表现类似。RT 方法和 RPG 方法的表现和 Wang 等(2011)在认知诊断测验下的检验结果一致。一般地,不同项目曝光控制方法按测量精度从高到低排序为 RT、RPG 和 MPI 方法。

5.2 项目曝光率分布结果

每种项目选择指标在与曝光控制方法结合前后的结果呈现在表 1、图 3 和图 4。

表 1　不同条件下各选题方法的测验重叠率和卡方值

选题方法	重叠率	χ^2	选题方法	重叠率	χ^2
D	0.408/0.23/0.23	152.6/75.14/75.14	V1	0.253/0.241/0.237	83.5/78.78/76.29
D – RPG	0.067/0.065/0.068	3.78/2.53/3.97	V1 – RPG	0.124/0.124/0.124	25.90/25.95/25.83
D – RT	0.123/0.122/0.123	25.63/24.89/24.86	V1 – RT	0.099/0.101/0.098	14.76/14.72/14.84
D – MPI	0.075/0.073/0.069	0.97/0.974/0.96	V1 – MPI	0.072/0.073/0.072	2.52/2.59/2.55
KLP	0.145/0.238/0.325	42.02/78.54/96.15	V2	0.114/0.113/0.113	21.37/20.83/20.81
KLP – RPG	0.078/0.074/0.074	7.23/3.40/3.45	V2 – RPG	0.124/0.125/0.124	15.89/25.92/15.90
KLP – RT	0.121/0.119/0.118	24.45/23.47/23.10	V2 – RT	0.092/0.086/0.093	11.64/8.61/11.88
KLP – MPI	0.087/0.098/0.098	10.35/14.29/14.19	V2 – MPI	0.074/0.077/0.074	3.29/4.44/3.29

注：每个单元格中代表相关为 0.3/0.6/0.8 的结果。

首先，根据卡方值、测验重叠率、题库利用率和过度曝光项目比例很容易推知四种项目选择指标的项目曝光率分布极不均匀。其中 D – 优化和 KLP 方法的题库利用率不足 50%；D – 优化、KLP 方法和 V1 方法过度曝光的项目比例达到 10% 及以上。尽管 V2 方法中从未曝光的项目比例接近 0，测验重叠率和 χ^2 值也比其它三种方法更小，它也不能得到比较满意的项目曝光率分布。图形 4(a) 以项目曝光率升序的方式描述了四种项目选择指标项目曝光率分布的曲线图。从图形 4(a) 可知四种项目选择指标的项目曝光分布都不均匀。

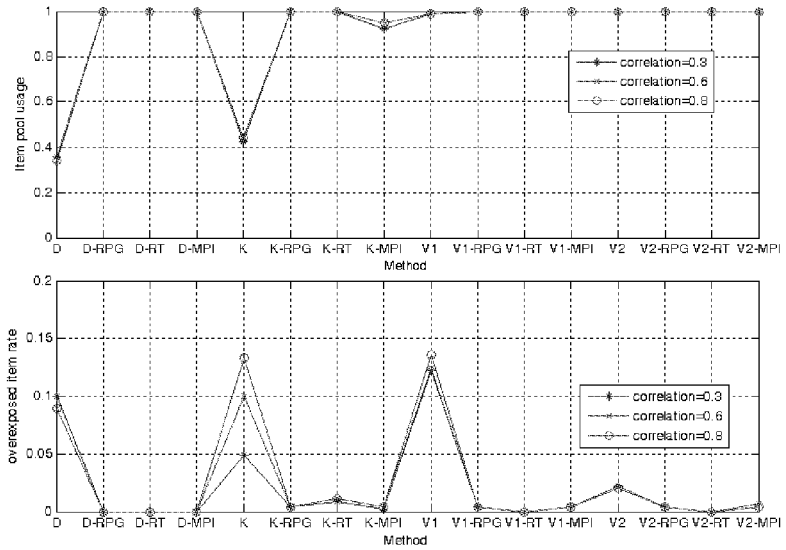

图 3　各种实验条件下不同方法的题库利用率和过度曝光项目比例

第二，所有项目曝光控制方法都增加了题库利用率，降低了过度曝光项目的比例、测验重叠率和卡方值，从而提高了项目曝光均匀性。根据表 1，尽管 RPG 方法与 MPI 方法表现类似，在大部分条件下都比其它方法表现更好。表 1 还可推知，相同项目曝光控制方法在与不同项目选择指标相结合的项目曝光率分布具有相似特征。因此，图 4(b) 以 KLP 选题指标为例，描绘了在能力相关为 0.6 的情况下，KLP 与不同曝光控制方法相结合前后的曝光率分布曲线。

另外，从图 4(b) 可以得知不同项目曝光控制方法的项目曝光率分布具有不同特征。结合图形 3，可知除了 KLP – MPI 方法外，其它方法的题库利用率达到 100%。换句话说，所有项目曝光控制方法都显著提高了题库利用率。检查过度曝光项目的比例，RPG 方法和 MPI 方法产生中过度曝光项目的数量在大部分条件下比 RT 方法的更多。一般地，RT 能将项目曝光率控制在允许的最大项目曝光率之下，而 RPG 和 MPI 方法都有少量过度曝光的项目。

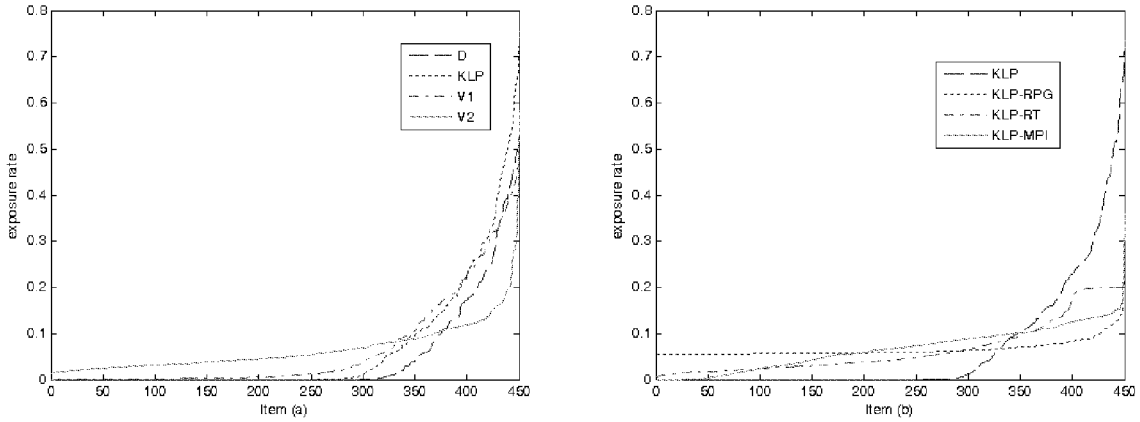

图 4　不同方法在能力相关为 0.6 时的曝光率分布图

注：图 a 为项目选择指标没有与项目曝光控制方法结合时的图形；图 b 为以 KLP 指标为例，与各种方法结合时的项目曝光率分布图。

根据表 1 和图 3，还发现一些特殊的情况。首先，当 MPI 与 D－优化、V1 和 V2 方法结合时，项目曝光率分布比它与 KLP 方法相结合时的曝光率分布更均匀。其次，当 RPG 方法与 V1 或 V2 方法相结合时，总会有 1 个或 2 个项目施测给所有被试。检查 V1－RPG 和 V2－RPG 的内部过程，发现在选择第一或第二个项目的时候，误差方差值总是显示“NaN”。换句话讲，V1－RPG 和 V2－RPG 方法中的过度曝光项目主要是由于项目信息矩阵非正定造成的。此外，V1－RPG 和 V2－RPG 的测验重叠率和卡方值显然也相应地受到第一个或前两个项目的影响。

总体上讲，尽管项目曝光控制策略的项目曝光率分布特征不尽相同，它们都能极大地提高项目曝光率分布的均匀性。这个结论可以通过直接比较图 4(a)和图 4(b)获得。研究结果也体现了测量精度和项目曝光率分布之间在一定程度上相互抵消的情况。

6　结论与讨论

许多研究已经表明 CAT 具有优于 P&P 测试和计算机测试的优点，例如它在减少测验长度，提高测量精度以及模型拟合方面表现得很好。对具有众多优势的 MCAT 而言，项目选择方法是 MCAT 成功应用于实践的关键之一(Wang & Chang，2011)。尽管已有项目选择指标能提高测量精度，但他们在处理过度曝光项目和曝光过低的项目时都显得脆弱无力。解决这个问题的有效方法是在项目选择过程中融合项目曝光控制策略。因此，本文基于模拟数据，考察了四种项目选择指标在与项目曝光控制策略结合前后的选题表现。

研究表明，V2 相对于 D－优化方法，KLP 和 V1 具有更高的项目库使用率，更少的过度曝光项目和更低的测试重叠率。通常，项目选择指标在不使用项目曝光控制策略时在项目曝光统计方面不令人满意，并且按照心理测量精度的的高低可以排序为 KLP，D－优化方法，V1 和 V2。此外，当使用项目曝光控制方法时，在所有项目选择指标的测量精度趋于降低。

在比较不同项目曝光控制方法产生的项目曝光率分布时，尽管 RPG 方法和 MPI 方法表现类似，RPG 方法在大多数情况下的表现优于其他方法，RT 方法表现最差。此外，每个项目曝光控制方法在不同的项目选择指标下产生相同的曝光率模式。当比较测量精度时，不同曝光控制方法可以排序为 RT，RPG 和 MPI。Chang 和 Twu(1998)曾指出在许多研究中观察到测量精度和项目曝光率的均匀性之间总是相互抵消。换句话言，为保证项目曝光率达到期望值，必将在一定程度上牺牲测量精度。

在本研究和 Wang 等(2011)的研究一致表明，在相同的测试条件下 RT 方法的测量精度优于 RPG 方法，在项目曝光分布的均匀性方面，RT 方法略差于 RPG 方法。总之，RT 与 RPG 方法能平衡测量精度和项目曝光均匀性，然而 MPI 方法在项目曝光分布方面表现较好。但在测量精度方面表现较差。

关于 MCAT 的项目选择方法的几个问题值得进一步研究。首先，虽然 D－优化性，V1 和 V2 比 KLP 快得多，但运行时间通常会随测试维度的增加而增加。因此，耗时的缺点可能影响 MCAT 在处理复杂测试条件时的应用。事实上，MCAT 优于单维 CAT 的特点主要在于能从多维度获得的详细的认

知信息。因此,需要更多关于减少项目选择方法的计算时间的有效算法,在已有选题方法上进行简化,提出有效简单的选题策略。例如 Wang 等(2011)提出的两个简化的 KL 指标。其次,MCAT 项目选择方法虽然可以保证每个维度的测试的测量精度,但在实际测试中遇到许多其他约束。因此,研究如何处理 MCAT 的非统计约束非常重要。

参考文献

Bloxom, B. M. ,& Vale, C. D. (1987). *Multidimensional adaptive testing: A procedure for sequential estimation of the posterior centriod and dispersion of theta.* Paper presented at the meeting of the Psychometric society, Montreal, Canada.

Bolt, D. M. , & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, *27*, 395 – 414.

Chang, S. W. ,& Twu, B. Y. (1998). A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing. *ACT Research Report Series*, 98 – 113.

Chang, H. H. , & Ying, Z. L. (1999). a – Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*, 211 – 222.

Chen, S. Y. , Ankenmann, R. D. ,& Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, *40*, 129 – 145.

Cheng, Y. ,& Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*, 369 – 383.

Finkelman, M. , Nering, M. L. ,& Roussos, L. A. (2009). A conditional exposure control method for multidimensional adaptive testing. *Journal of Educational Measurement*, *46*, 84 – 103.

Huebner, A. R. , Wang, C. , Quinlan, K. ,& Seubert, L. (2015). Item exposure control for multidimensional computer adaptive testing under maximum likelihood and expected a posterior estimation. *Behavior Research Methods*, DOI 10. 3758/s13428 – 015 – 0659 – z.

Lee, Y. H. , Ip, E. H. ,& Fuh, C. D. (2008). A strategy for controlling item exposure in multidimensional computerized adaptive testing. *Educational and Psychological Measurement*, *68*, 215 – 232.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, *20*, 389 – 404.

McKinley, R. L. ,& Reckase, M. D. (1982). *The use of the general Rasch model with multidimensional item response data* (Research Report ONR 82 – 1). American College Testing, Iowa City, IA.

Mulder, J. ,& van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria. *Psychometrika*, *74*, 273 – 296.

Mulder, J. ,& van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback – Leibler information item selection. In W. J. van der Linden & C. A. W. Glas(Eds.), *Elements of Adaptive Testing, Statistics for Social and Behaviroal Sciences*. Springer Science + Businesws Media.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331 – 354.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Report No. 94 – 5). Princeton, NJ: Educational Testing Service.

Stocking, M. L. ,& Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *23*, 57 – 65.

Sympson, J. B. ,& Hetter, R. D. (1985). Controlling item – exposure rates in computerized adaptive testing. *In Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973 – 977). San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error – variance criterion. *Journal of Educational and Behavioral Statistics*, *24*, 398 – 412.

van der Linden, W. J. , & Veldkamp, B. P. (2007). Conditional item exposure control in adaptive testing using item – ineligibility probabilities. *Journal of Educational and Behavioral Statistics*, *32*, 398 – 418.

Veldkamp, B. P. ,& van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*, 575 – 588.

Wang, C. ,& Chang, H. H. (2011). Item selection in multidimensional computerized adaptive testing – gaining information from different angles. *Psychometrika*, *76*, 363 – 384.

Wang, C. , Chang, H. H. , & Boughton, K. A. (2011). Kullback – Leibler information and its applications in multidimensional adaptive testing. *Psychometrika*, *76*, 13 – 39.

Wang, C. , Chang, H. H. , & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *37*(2), 99 – 122.

Wang, C. , Chang, H. H. , & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, *48*, 255 – 273.

Wang, W. C. , & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychologica Measurement*, *28*, 295 – 316.

Yao, L. (2010). Reporting valid and reliability overall score and domain scores. *Journal of Educational Measurement*, *47*, 339 – 360.

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and applications. *Psychometrika*, *77*, 495 – 523.

Yao, L. (2014a). Multidimensional CAT item selection methods for domain scores and composite scores with item exposure control and content constrains. *Journal of Educational Measurement*, *51*, 18 – 38.

Yao, L. (2014b). Multidimensional item response theory for score reporting. In Y. Cheng & H. – H. Chang (Eds.), *Advances in modern international testing: Transition from summative to formative assessment.* Charlotte, NC: Information Age.

Yao, L., Pommerich, M., & Segall, D. O. (2014). Using Multidimensional CAT to Administer a Short, Yet Precise, Screening Test. *Applied Psychological Measurement*, *38*, 614 – 631.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed – format tests. *Applied Psychological Measurement*, *37*, 3 – 23.

A Comparison Study of Item Selection Strategies with Item Exposure Controlling in MCAT

Mao Xiuzhen Wang Yating Yang Rui

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

Abstract: Four item selection indexes with and without exposure control are evaluated and compared in multidimensional computerized adaptive testing (CAT). The four item selection indices are D – optimality, Posterior expectation Kullback – Leibler information (KLP), the minimized error variance of the linear combination score with equal weight (V1), and the minimized error variance of the composite score with optimized weight (V2). The maximum priority index (MPI) method for unidimensional CAT and two item exposure control methods (the restrictive threshold (RT) method and restrictive progressive (RPG) method, originally proposed for cognitive diagnostic CAT) are extended to the miltidimentional CAT. The results show that: (1) KLP, D – optimality, and V1 perform well in recovering domain scores, and all outperform V2 in psychometric precision; (2) KLP, D – optimality, V1, and V2 produce an unbalanced distribution of item exposure rates, although V1 and V2 offer improved item pool usage rates; (3) all the exposure control strategies improve the exposure uniformity greatly and with very little loss in psychometric precision; (4) RPG and MPI perform similarly in exposure control, and are both better than RT.

Key words: multidimensional item response theory; computerized adaptive testing; item selection methods; exposure control strategy; psychometric precision