文章编号: 1000-5862(2012)05-0446-06

等级评分模型下的最大信息量分层选题策略

程小扬[1], 丁树良[2], 朱隆尹[1], 巫华芳[1]

(1.赣南师范学院数学与计算机科学学院, 江西 赣州 341000; 2.江西师范大学计算机信息工程学院, 江西 南昌 330022)

摘要: 对于 0-1 评分模型, R. B.Juan 等提出了最大信息量分层选题策略, 将此选题策略应用到等级反应评分模型(GRM)中, 即以项目 j 的最大信息量 $I_{\max}(j)$ 作为分层的依据, 以取得该项目的最大信息量时能力点值 $\theta_{\max}(j)$ 作为项目的综合难度, 分别用 $I_{\max}(j)$ 与 $\theta_{\max}(j)$ 替代张华华等提出的按 a 分层和按 b 分块按 a 分层方法中的 a、b 参数, 形成最大信息量按 a 分层选题策略(MI-AS)和最大信息量按 b 分块按 a 分层方法(MI-BS). 模拟实验结果表明: MI-AS 和 MI-BS 方法较传统的按 a 分层方法要好.

关键词: 计算机化自适应测验; 等级反应模型; 最大信息量分层法; 选题策略
中图分类号: O 626.4　　文献标志码: A

0 引言

随着科学技术的发展, 计算机化自适应测验(computerized adaptive testing, CAT)已成为一种非常重要的新型考试模式, 正逐步替代传统的纸笔考试模式. 虽然 CAT 能有效地缩短考试时间, 提高测量的精度, 但 E.S.Quellmalz 和 J.W.Pellegrino 指出, 在高风险测试中, CAT 的使用应受到限制[1]. 在 CAT 的测试中, 为提高测量的效率, 会倾向于选择区分度大的项目, 致使这些高区分度的项目频繁调用, 影响考试安全. 因此, 选题策略一直是 CAT 实施质量的关键[2]. 如何提高测量的精度, 同时又兼顾降低项目的曝光率, 这一直是选题策略要研究的重点问题.

目前, 比较典型的选题策略有 2 类: (1)F.M. Lord[3]提出的最大 Fisher 信息量方法(maximum fisher information, MFI), (2)Chang Huahua 等[4-5]提出的分层化的选题方法. 其中, MFI 方法测量的精度较高, 缺点是高区分度的项目会被频繁调用, 考试安全问题突出. 为此, 学者们提出了很多控制曝光率的方法, 其中著名的有 Sympson 和 Hetter 提出的 SH 方法; van der Linden 和 Veldkamp 提出的项目合格方法, Barrada, Veldkamp 和 Olea 提出的多重最大曝光率法等[6]. 这些方法都能在一定程度上抑制高曝光率项目的过度使用, 但都无法提高低曝光率项目的抽取几率. Chang Huahua 等在分析最大信息量的基础上, 提出了 a-stratification(简称 AS)方法[4], 该方法将题库按区分度 a 的非降序排列, 然后将题库分成大小相等的几层, 选题时从低 a 层开始选题, 随着测试深入, 逐步进入高 a 层. 这样, 低区分度的项目开始时使用, 高区分度项目在最后使用. 张华华等[2]认为, 该方法具有 2 个好处: (i)在开始阶段使用区分度低的项目, 使得考生不致于开始阶段答错几道题, 而严重低估考生的能力; (ii)在对测量的精准度降低不太大的基础上, 有助于降低项目的曝光率. 随后, 在 Chang Huahua 等考虑到实际使用的题库中, 难度参数 b 与区分度参数 a 存在正相关性, 因此, 在分层时将考虑到难度 b 在各层次的平衡, 提出了按 b 分块 a 分层的方法(简称 BS). R. B. Juan 等[7]提出, 对于 3PLM 模型来说, 在进行分层化选题时, 除了要考虑参数 a 和 b 外, 分层时还应考虑猜测度 c 对信息量的影响, 同时, 在每层选题时, 由于 c 的存在, 最大信息量并不在能力与难度相等处, 为此, 他们提出 MFI 应与 AS 或 BS 相结合, 形成了新的选题方法, 叫最大信息量按 a 分层选题策略(MI-AS)和最

收稿日期: 2012-06-16
基金项目: 国家自然科学基金(30860084, 31160203, 31100756), 江西省教育厅青年科学基金(GJJ10238)和江西省教学改革研究课题(JXJG101113)资助项目.
作者简介: 程小扬(1978-), 男, 江西九江人, 讲师, 硕士, 主要从事人工智能和计算机辅助教育的研究.

大信息量按 b 分块按 a 分层方法(MI-BS). 根据 R. B. Juan 等的结果, MI-AS、MI-BS 比相应的 AS、BS 的测验效率更高, 但曝光率稍差一些. 当然, R. B. Juan 等的方法对于 0-1 单维评分模型来说是合适的, 如何将该方法移植到多级评分模型以及移植以后其效果如何评估, 本文将就此展开讨论.

众所周知, 国外的考试过去一段时间过分偏好于客观题, 造成了国外的 CAT 的应用与研究多建立在 0-1 评分模型基础上. 但近年来, 许多学者对此进行反思, 认为多级评分模型将是今后一个研究的热点方向. 通过实践表明, 多级评分项目比 0-1 评分模型可以获得更多的被试信息[8]. 对多级评分模型的 CAT, 我国学者近年来对此进行大量研究, 戴海崎等[9]比较了等级评分模型下的中位数与平均数选题方法的优劣, 陈平等[8]对等级评分模型下去两端平均法、难度匹配法进行比较, 刘珍等[10]对 GPCM 模型下的 4 种选题方法进行了比较, 罗芬等[11]提出了在等级评分模型下的动态综合选题策略. 由于多级评分模型中难度(阈值)参数有多个, 而认为被试特质能力值是单维的, 因此, 为了与该单维的能力值相匹配, 一般需要对难度参数向量降维. 以上研究采用的降维办法是简单的求所有等级难度参数的平均值或部分 (甚至某一个) 参数的平均值, 通过罗照盛等[12]对等级评分模型下项目信息量的研究, 取得项目信息最大, 对应的能力点在“邻近难度等级占优”处. 但如何具体求出该占优点, 罗照盛等并未给出答案. 由此, 如何综合这些难度参数, 是一个值得研究的课题. 从信息量公式可以看出, 信息量能较好地将各难度参数综合起来, 于是, 本文从信息量公式出发, 求出试题所处的最大信息量的能力点. 虽然在多级评分模型中, 一般不考虑猜测度参数 c, 但项目所能提供的信息量大小并不完全由区分度参数来决定. 如假设有 2 个项目的难度等级分别为(–2, –1, 0, 1, 2), (–1, –0.5, 0, 0.5, 1), 当它们的区分度 a 都为 1 时, 虽然它们的平均值等指标都相等, 但它们在能力值为 0 处的项目信息量大小并不相同. 所以, 在分层时需要考虑到它们之间的这种差异, 而不仅仅只是考虑 a 参数. 因此, 本文试图以题库中该项目提供的最大信息量作为分层的依据, 探索这种选题策略的表现.

1　等级反应模型下的最大信息量分层选题策略

1969 年, Samejima 给出了有序多值评分项目的等级反应模型(GRM), 它把每个项目分成若干个等级, 各个等级难度要求严格递增, 记 $P^*_{\alpha j,t}$ 为被试在第 j 个项目得 t 分或 t 分之上的概率, 记 $P_{\alpha jt}$ 被试 α 在第 j 个项目恰得 t 分的概率, 则

$$P_{\alpha jt} = P^*_{\alpha j,t} - P^*_{\alpha j,t+1}, \tag{1}$$

$$P^*_{\alpha j,t} = \frac{1}{1+\exp[-Da_j(\theta_\alpha - b_{jt})]}, \tag{2}$$

其中 a_j 为第 j 个项目的区分度, b_{jt} 为第 j 个项目等级 t 的难度, 显然如该项目满分为 f 分, 则 $P^*_{\alpha j,0}$ =1, $P^*_{\alpha j,f+1}$=0.

GRM 模型项目的 Fisher 信息量公式为

$$I_j(\theta_\alpha) = \sum_{t=0}^{f_j} D^2 a_j^2 P_{\alpha jt}(1 - P^*_{\alpha j,t} - P^*_{\alpha j,t+1})^2, \tag{3}$$

然而(3)式是关于能力的非线性函数, 要解出使(3)式获取最大值时 θ_α 的值, 非常复杂, 求解有一定困难, 但可借助计算机来求近似值. 求解过程为: 将能力区间[−3, 3]等距地截取 601 个能力点 $\delta_1, \delta_2, \cdots, \delta_{601}$, 步长为 0.01, 将这些 $\delta_1, \delta_2, \cdots, \delta_{601}$ 分别替代 θ_α, 代入(3)式, 计算出使(3)式取得最大的值, 记为 $I_{\max}(j)$, 即项目 j 的最大信息量值; 同时记下使信息量最大时的能力点 δ_k, 记为 $\theta_{\max}(j)$. 在实际题库中一般中等难度的项目较多, 可能有很多试题计算出来的 $\theta_{\max}(j)$ 值相等, 为避免出现该情况, 在上面的计算过程中, 可将各能力点改为 $\delta_1+\varepsilon_j$, $\delta_2+\varepsilon_j, \cdots, \delta_{601}+\varepsilon_j$, 其中 ε_j 为(0, 0.01)之间的随机数.

若将以上计算出来的 $I_{\max}(j)$ 来替代项目 j 的区分度 a_j, 用 $\theta_{\max}(j)$ 表示项目难度的综合值, 以此将 Chang Huahua 等的按 a 分层法作如下改变, 形成最大信息量与按 a 分层方法结合的方法, 简称为 MI-AS. 具体步骤如下: (i)将题库中的项目按 $I_{\max}(j)$ 的升序排列, 然后将题库分成 K 层, 每层项目的数量大致相等, 第 1 层 $I_{\max}(j)$ 值最小, 第 K 层 $I_{\max}(j)$ 最大; (ii)在开始测试时, 被试首先进入第 1 层选题, 假设当前能力估计值为 $\hat{\theta}$, 则选择该层中使得下式最小的项目: $|\hat{\theta} - \theta_{\max}(j)|$; (iii)当满足该层的结束条件, 则进入下一层作答. 一直到结束第 K 层, 整个测试结束.

同理, 下面将 Chang Huahua 等[5]的按 b 分块按 a 分层法也作同样的改变, 记为 MI-BS, 具体步骤如

下: (i)假设题库中的项目数为 N, 将题库分成 K 层; (ii)将题库中项目按 $\theta_{\max}(j)$ 的升序排列, 并将题库分为 N/K 个区块, 每个区块有 K 个项目; (iii)将每个区块内的项目按 $I_{\max}(j)$ 从小到大分成 K 个层, 每层一个项目, 即第 1 层 $I_{\max}(j)$ 最小, 第 K 层 $I_{\max}(j)$ 最大; (iv)将每区块的第 i 层合并成一层, 并放在第 i 层, 其中 i=1, 2, …, K; (v)对应 K 层, 将测验也分成 K 个阶段, 第 i 个阶段就在第 i 层选题, 假设当前估计能力值为 $\hat{\theta}$, 选题时也是选择该层中使得下式最小的项目: $|\hat{\theta}-\theta_{\max}(j)|$; (vi)当满足该层的结束条件, 则进入下一层作答. 一直到结束第 K 层, 则整个测试结束.

2 CAT 设计过程

2.1 题库及被试的产生

基于GRM模型, 模拟产生400个项目, 每个项目的等级数都为5, 等级难度参数服从标准正态分布, 项目的区分度参数服从对数正态分布, 且取值范围为[0.2, 2.5].

被试的产生分 2 类: (i)群体 1: 模拟产生 5 000 名被试特质水平(能力), 被试的能力 θ 服从标准正态分布; (ii) 群体 2: 在[−2.6, 2.6]中, 每隔 0.2 取一个能力点(共 27 个能力点)上分别产生 500 名被试, 共 13 500 名.

群体 1 的被试分布为通常情况; 群体 2 主要考察不同水平的被试在各种选题策略上的测验精度或效率.

2.2 拟比较的选题策略

为查看本文中的 MI-AS、MI-BS 方法在等级评分模型下的效果, 在此选择 MFI、难度匹配法、平均数法、中位数法、去两端平均法等选题策略作为比较, 其中难度匹配法、平均数法、中位数法、去两端平均方法都是在按 a 分层下进行的. 因为基于 GRM 的每个项目有若干个难度参数, 故无法直接使用按 b 分块按 a 分层法(BS), 因此 BS 方法不能列入比较.

2.3 CAT 测试过程

初始阶段采用从当前层中随机选择 3 道试题作答, 然后将得分与失分之比的自然对数作为能力初始估计可以取为; 随后进入能力的精确估计阶段, 依据本文所介绍的选题策略进行选题. 再依据被试在项目上的得分, 用 EAP 估计被试的能力. 重复该步骤, 直至满足测验结束条件.

为从不同的条件下比较选题方法的优劣, 终止规则分别采用定长和不定长的2种方式进行. 在定长测验中, 设置测验的长度为16, 分成4层; 对于不定长测验, 采用累积信息量的大小作为结束条件, 信息量的大小设为16, 且最大测验长度不超过30, 也分成4层, 每层的结束方法采用戴海琦等的办法.

2.4 评价指标

从测验的精度和安全性2个方面来评价CAT的质量, 其中精度指标本文用*Bias*、*RMSE*、测验效率来评价, 安全性指标用 χ^2 和测试重叠率*Rt*来评价, 具体计算公式为

$$Bias=\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i-\theta_i),$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i-\theta_i)^2},$$

$$ABS=\frac{1}{N}\sum_{i=1}^{N}|\hat{\theta}_i-\theta_i|,$$

$$\text{测验效率}=\sum_{i=1}^{N}inf_i\Big/\sum_{i=1}^{N}L_i,$$

$$\chi^2=\sum_{j=1}^{M}\{[A_j-(\sum_{j=1}^{M}A_j/M)]^2/(\sum_{j=1}^{M}A_j/M)\},$$

$$Rt=\frac{2TO_{总}}{(N-1)\sum_{i=1}^{N}L_i},$$

其中 N 为被试总数, θ_i 为被试 i 的能力真值, $\hat{\theta}_i$ 为被试 i 的估计能力, inf_i 为被试 i 测验的信息总量, L_i 为被试 i 的测试长度; M 为试题数量, A_j 是第 j 题曝光率. A_j 的计算公式为 A_j= 第j题被使用的次数 / N, $TO_{总}$是考生的试题重叠总数, 计算方法为: $TO_{总}=\sum_{j=1}^{M}C_{M_j}^2$, M_j 是试题库中第 j 题使用次数, $C_{M_j}^2$ 为从 M_j 个元素中取出 2 个元素的组合数.

对于不定长测验来说, 如下定义的人均用题数(Nf)是一个非常重要的指标

$$Nf=\left(\sum_{i=1}^{N}r_i\Big/N\right),$$

其中 r_i 为第i个被试在模拟中作答的项目数.

3 实验结果及分析

依据以上介绍的评价标准, 将MI-AS、MI-BS方

法、MFI方法、4种按a分层下的选题方法应用在GRM模型下比较, 通过蒙特卡罗模拟实验得如下数据. 在实验中分层数都设为4,

不定长测验的结束规则设为测验信息量16, 定长测试的长度设置为16. 中位数、平均数等方法表示的都是按a分层下取得, 但为表述简洁, 下面的表述中都省略了按a分层.

3.1　不定长 CAT 的实验情况

被试群体 1 参加模拟的不定长 CAT, 各种选题策略在上述各评价指标上的总体表现如表 1 所示.

表1　不定长*CAT*下各种选题策略比较

选题策略	*Bias*	*RMSE*	*ABS*	测验效率	*Nf*	χ^2	*Rt*
MI-AS	–0.001	0.239	0.190	0.77	22.9	13.7	0.091
MI-BS	0.002	0.236	0.188	1.01	17.5	6.0	0.058
MFI	0.003	0.240	0.190	2.69	6.8	66.0	0.182
中位数	0.000	0.238	0.189	0.75	23.2	15.2	0.096
平均数法	–0.004	0.241	0.192	0.74	23.6	23.0	0.116
去两端平均数法	0.001	0.235	0.188	0.80	21.9	18.5	0.101
难度匹配法	0.001	0.244	0.194	0.74	23.8	7.1	0.077

从表1中可知, 各种选题方法中, *Bias*都较小, 基本接近于0, 说明各种方法都接近于无偏. 由于不定长测验下的各种选题策略采用相同的信息量作为结束条件, 所以反映测量精度的1个主要指标RMSE和ABS上差异都较小, 其中以去两端平均数法和MI-BS稍好一些. 而从另一个*Nf*指标上来看, MI-BS方法的平均测验长度是除MFI外, 平均测量长度最短的. 这说明MI-BS方法可用较少的试题(比传统的按a分层法大致要少5道试题), 达到其它选题策略相同的测量精度. 而MI-AS方法平均测验长度与其它几种选题策略方法基本接近.

从安全方面来考虑, MFI 方法在χ^2 和 *Rt* 指标上表现最差, MI-BS 在这 2 个指标上表现最好, 比难度匹配法都要优异. 而 MI-AS 法也要比平均数法、中位数、去两端平均数法都要好. 考试安全性是 CAT 一个需要重点考虑的方面, 为更详细的了解在不定长 CAT 下各个项目曝光率情况,图 1 记录下 400 个项目的曝光率分布情况.

从图1可知, MI-BS方法使项目曝光率接近于0.044 (MI-BS平均曝光率为17.5/400=0.044)的密度最大, 说明大多项目的曝光率集中在平均曝光率周围; 高曝光率项目的密度比其它几种方法要小, 低曝光率项目的密度比平均数法和MI-AS要小, 但比难度匹配法稍大.

为更加详细的了解各选题策略对不同能力水平的被试在测量精度的影响, 将被试群体2加入模拟的CAT进行比较研究. 在不定长测验中, 由于采用统一的测验信息量作为结束测验的依据, 则每个被试的标准误差基本相同, 即表现为*ABS*和*RMSE*基本一致. 这时, 每个被试的测验长度(*Nf*)可反映测验效率. *Nf* 在27个能力特征点上的表现如图2所示.

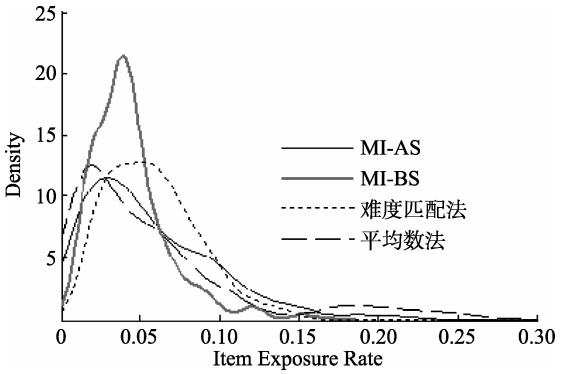

图1　不定长CAT下项目曝光率密度分布图

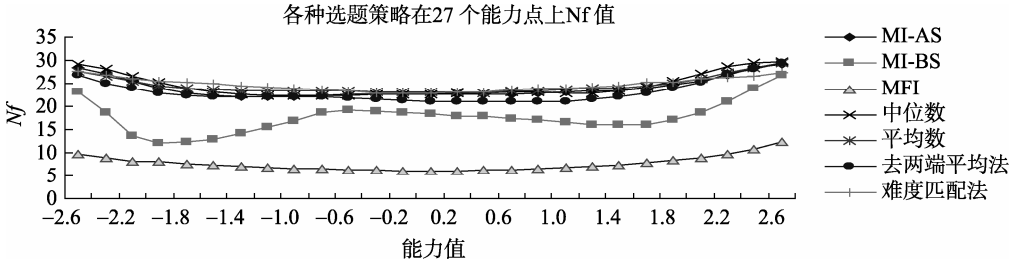

图2　各种选题策略在27个能力点上*Nf*值

从图2可知，对不同能力水平的被试，*MI-BS*的测验是除MFI方法外长度最小的，MI-AS与另外4种传统的按*a*分层法的长度非常接近.

综合安全性和精度2个方面，在不定长CAT中，MI-BS方法在这2个方面表现都较优异.

3.2 定长 **CAT** 的实验情况

被试群体 1 参加模拟的定长 CAT，各种选题策略在各评价指标上的总体表现如表 2 所示.

对于定长测验来说，在相同的作答题数下，对于 2 个精度指标 *RMSE* 和 *ABS,* MI-AS 和 MI-BS 方法比 4 种按 *a* 分层下的方法都要好(尤其是 MI-BS 方法表现更突出)，而比 MFI 要差一此.

从表 2 中反映安全性的 2 个指标来看，难度匹配法最好，MI-BS 和 MI-AS 次之，其它几种选题策略更差一些. 在定长 CAT 下详细的 400 个项目曝光率分布情况如图 3 所示.

表 2 定长 **CAT** 下各种选题策略比较

选题策略	*Bias*	*RMSE*	*ABS*	测验效率	χ^2	*Rt*
MI-AS	0.005	0.234	0.184	1.17	8.9	0.062
MI-BS	0.003	0.230	0.182	1.21	6.2	0.055
MFI	0.000	0.143	0.113	3.45	139.8	0.389
中位数	–0.003	0.237	0.187	1.15	12.4	0.071
平均数法	0.006	0.240	0.189	1.05	19.0	0.087
去两端平均数法	–0.002	0.236	0.185	1.11	17.7	0.084
难度匹配法	0.004	0.240	0.189	1.13	3.6	0.049

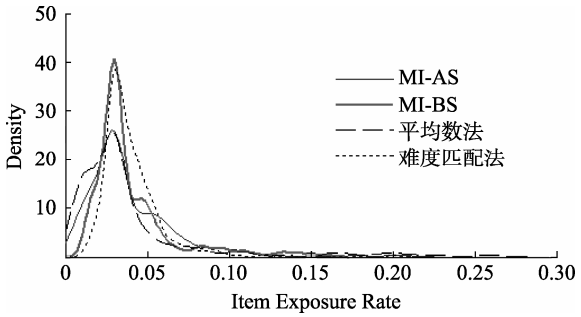

图 3 定长 **CAT** 下项目曝光率密度分布图

从图 3 可知, MI-BS 与难度匹配法的曝光率基本接近，曝光率落在 0.035~0.045 的项目数最多(项目的平均曝光率为 16/400=0.04)，曝光率落在两头的较少，说明这 2 种方法使低曝光率和高曝光率的项目都有减少，中等曝光率的项目增多，测验的安全性最好. MI-AS 方法与平均数等方法的结果基本接近，比 MI-BS 和难度匹配法差一些.

在定长测验中，测验长度一定时，*ABS* 和 *RMSE* 指标可反映测验的精度. 将被试群体 2 加入模拟的 CAT 进行比较研究，在此只取 *ABS* 作为比较，27 个能力特征点在 *ABS* 上的表现如图 4 所示.

从图 4 可知，除 MFI 外, MI-BS 在高水平和低水平的被试中测量精度(*ABS*)值都是最高的(即误差最小)，在中等水平的被试中与另外 5 种按 a 分层法(包括 MI-AS)的 *ABS* 值基本接近.

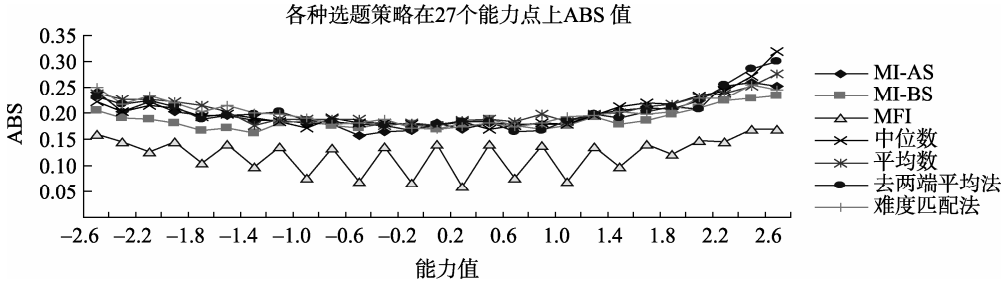

图 4 各种选题策略在 **27** 个能力点上 ***ABS*** 值

4　结论与展望

通过比较可以得出，在等级反应模型中，采用最大信息量分层法(特别是 MI-BS)作为 CAT 选题策略，比传统的 MFI 方法的曝光率要好很多，而与陈平等或戴海琦等提出的几种按 a 分法比较，测验效率和项目曝光率都有一定的提高．还有，本文在多级评分中考虑到难度的平衡问题而引入的 MI-BS，这在多级评分中是否有必要，值得进一步探索．

当然，对于在实际题库中可能还要考虑内容等约束条件，本文并没有做相应探讨．还有，将 MI-AS、MI-BS 方法与传统的 *SH* 等曝光率控制方法相结合，在 GRM 模型下会有怎样的效果，也有待于进一步研究．

注意到拓广分部评分模型(GPCM)项目中的步骤参数不一定单调上升，这一点和 GRM 的项目难度参数不同，所以这里提出的 2 种新的多级评分模型的选题策略是否适用于 GPCM，值得考虑．

5　参考文献

[1] Quellmalz E S, Pellegrino J W. Perspective: technology and testing [J]. Science, 2009, 323(2): 75-79.

[2] 张华华，程莹．计算机化自适应测验(CAT)发展前景和展望 [J]．考试研究, 2005(4): 12-24.

[3] Lord F M. A broad-range tailored test of verbal ability [J]. Applied Psychological Measurement, 1977(1): 95-100.

[4] Chang Huahua, Ying Zhiliang. *a*-stratification multistage computerized adaptive testing [J]. Applied Psychological Measurement, 1999, 23: 211-222.

[5] Chang Huahua, Qian Jiahe, Ying Zhiliang. *a*-stratified multistage computerized adaptive testing with b blocking [J]. Applied sychological Measurement, 2001, 25: 333-341.

[6] 李铭勇，张敏强，简小珠．计算机自适应测验中测验安全控制方法评述 [J]．心理科学进展, 2010, 18(8): 1339-1348.

[7] Juan R B, Paloma M, Julio O. Maximum information stratification method for controlling item exposure in computerized adaptive testing [J]. Psicothema, 2006, 18: 156-159.

[8] 陈平，丁树良，林海菁，等．等级反应模型下计算机化自适应测验选题策略 [J]．心理学报, 2006, 38(3): 461-467.

[9] 戴海崎，陈德枝，丁树良，等．多级评分题计算机自适应测验选题策略比较 [J]．心理学报, 2006, 38(5): 78-783.

[10] 刘珍，丁树良，林海菁．基于 GPCM 的计算机自适应测验选题策略比较 [J]．心理学报, 2008, 40(5): 618-625.

[11] 罗芬，丁树良，王晓庆．多级评分计算机化自适应测验动态综合选题策略[J]．心理学报, 2012, 44(4): 400-412.

[12] 罗照盛，欧阳雪莲，漆书青，等．项目反应理论等级反应模型项目信息量 [J]．心理学报, 2008, 40 (11): 1212-1220.

The Stratified Item Selection Strategy with Maximal Information under Graded Response Model

CHENG Xiao-yang[1], DING Shu-liang[2], ZHU Long-yin[1], WU Hua-fang[1]

(1.College of Mathematics and Computer, Gannan Normal University, Ganzhou Jiangxi 341000, China;
2.College of computer Information and Engineering, Jiangxi Normal University, Nanchang Jiangxi 330027, China)

Abstract: The stratified item selection strategy with maximal information proposed by Juan R. B. is applied to graded response model. Item j's maximal information $I_{\max}(j)$ is taken as the basis of stratificaton, and the ability value $\theta_{\max}(j)$ where the item j reaches its maximal information is taken as item j's comprehensive difficulty. The parameter a and b in the a-stratified method(AS) and a-stratified method with b blocking (BS) proposed by Chang Huahua et al. are replaced by $I_{\max}(j)$ and $\theta_{\max}(j)$, respectively, so as to form MI-AS method and MI-BS method. The simulation experiments demonstrate that MI-AS and MI-BS are better than traditional AS and BS.

Key words: computerized adaptive testing; Graded Response Model; Item Selection Strategies

(责任编辑：冉小晓)