文章编号:1000-5862(2013)01-0106-06

引入内容平衡的最大信息量组块分层选题策略

詹沛达,王立君*,杨卫敏
(浙江师范大学教师教育学院心理系,浙江 金华 321004)

摘要:在 0-1 计分下,为了解决最大信息量组块分层策略(MIS-B)中未考虑内容平衡的问题,通过加入改良多项式模型来平衡内容属性.计算机模拟试验显示:选题策略在保持 MIS-B 能力估计精准度这一前提下降低了项目重叠率,提高了题库使用均匀性和项目曝光率的均匀性.

关键词:计算机化自适应测验;内容平衡;最大信息量组块分层选题策略;改良多项式模型
中图分类号:B 841　　**文献标志码**:A

0 引言

计算机化自适应测试(computerized adaptive testing,CAT)是基于基于项目反应理论(item response theory,IRT)、计算机技术和现代教育技术的一种测验形式,它根据被试在已作答项目上的表现自适应地从题库或剩余题库中选择测验项目施测.与传统纸笔考试相比,CAT 的主要优点在于使用较少的项目而达到对被试能力值更精准的估计[1].此外的优势还有:不要求统一时间统一地点;测验之后立即呈现分数;甚至还可以施测包括视频或音频剪辑在内的新项目类型(innovative item types),这使得国内外学者对 CAT 的研究越来越多.

CAT 包括 6 个基本组成部分:采用的项目反映模式、题库、初始项目的选择、参数估计方法、选题策略和测验终止规则[2],其中选题策略作为 CAT 的重要环节之一,它的好坏与考试的信、效度,测验安全性以及测量准确性直接相关.van der Linden[3]、Cheng Ying 等[4]以及 H. Deng 等[5]提到 CAT 选题要同时考虑统计优化问题和一些非统计约束条件.统计优化主要指根据被试的反应,选择最适合其作答的项目以提高潜质估计的精度.非统计约束包括项目曝光控制、内容平衡、正确选项分布的平衡、项目长短适当、被试反应时间均衡、测验分数等值等[6].按选题策略的目的,将选题策略分为提高测量准确性的选题策略和具有非统计约束的选题策略:(1)提高测量准确性的选题策略主要包括:最大 Fisher 信息量(maximum fisher information,MFI)法[7]、最大全局信息量(maximum global-information,MGI)法[8]、全贝叶斯准则(fully Bayesian criteria)[9]等;(2)具有非统计约束的选题策略可分为具有曝光率控制的安全性选题策略和具有其他非统计约束的选题策略.安全性选题策略主要包括 S-H 法(sympson-hetter method,S-H)[10]、渐进法(Progressive Method,PG)[11]、比例法(proportional method,PP)[12]、a 分层法(a-Stratified method,a-STR)[13]、b 模块化的 a 分层策略(BAS)[14]等.具有其他非统计约束条件的选题策略主要包括:内容分块法(Constrained CAT,CCAT)[15]、加权离差法(weighted deviation model,WDM)[16]、影子测验(shadow test,ST)[17]等.内容平衡作为非统计约束中较为重要的一项,是在题库建设和选题策略中均需要考虑的问题.

1 引入内容平衡的最大信息量组块分层策略

1.1 基础知识简介

1.1.1 内容平衡问题简介　内容平衡包括题库的内容平衡和选题的内容平衡.在传统的纸笔考试中,由于所有考生的考题(选题)都是一样的,因此只要试卷(题库)是根据双向细目表中不同内容域的权重所构建的,通常是不会存在内容平衡问题的.但在 CAT 测验中,不同的考生所接触到的项目可能是不同的(自适应的),这就可能出现即使题库不存在内

收稿日期:2012-09-21
作者简介:王立君(1968-),女,辽宁大连人,博士,副教授,主要从事学科能力测量,青少年社会性发展与积极心理学等方面的研究.

容平衡问题,但选题也会出现内容不平衡的情况. 通常,产生选题的内容平衡问题的情况包括: (1) 不同内容域内的项目难度差异较大; (2) 被试对不同内容域中知识掌握的程度差异较大. 因此,有可能出现某个被试已做答项目大多数都属于较简单的内容域或掌握程度较高的内容域,那么他将获得比实际水平高的分数. 此外,当不同被试作答的项目所属的内容域比例不同时不利于不同被试之间分数的比较. 因此,在实际的 CAT 应用中,测验的内容平衡问题需要考虑. 然而内容平衡本身并不是 IRT 模型所考虑的,因此为了获得在各个内容区域平衡的测试,除在题库建设时需要考虑到内容平衡外,在 CAT 的选题策略中也需要加入内容平衡的过程.

1.1.2　3PLM 模型简介　伯恩鲍姆(A. Birnbaum)给出了 3 参数 Logistic 模型(3PLM)[18] 为

$$P_j(\theta) = c_j + \frac{1 - c_j}{1 + e^{-Da_j(\theta - b_j)}}, \quad (1)$$

其中 θ 为潜在特质,也称为能力,D 为常数(一般取值1.7),a_j、b_j、c_j 分别表示第 j 个项目的区分度、难度和猜测度. $P_j(\theta)$ 表示能力为 θ 的被试在第 j 个项目上正确作答的概率.

1.1.3　最大信息量组块分层策略(MIS-B)简介　a-STR 和 BAS 只考虑了 2 个项目参数: 项目区分度 a 和项目难度 b,而对于3PLM 特有的项目猜测度 c 参数并没有考虑在内. 由于在 3PLM 中,将题库按照区分度 a 排序与按题目最大信息量 $I_j^{\max}$ 排序得到的结果并不一致,与最大信息量 $I_j^{\max}$ 对应的能力值 $\theta_j^{\max}$ 也不会与项目难度 b 相等. 因此,在 3PLM 中 a-STR 策略和 BAS 策略并不会达到应有的效果. 为了将项目猜测度 c 参数引入到分层策略中,J. R. Barrada 等[19] 对 a-STR 和 BAS 进行了 2 个重要的修改: (1) 使用项目最大信息量 $I_j^{\max}$ 代替区分度 a; (2) 用项目信息函数达到最大值时对应的能力值 $\theta_j^{\max}$ 代替难度 b.

在 3PLM 中,项目 j 的最大信息量($I_j^{\max}$) 为[20]

$$I_j^{\max} = \frac{1.7^2 a_j^2}{8(1 - c_j^2)}\left[1 - 20c_j - 8c_j^2 + (1 + 8c_j)^{3/2}\right], \quad (2)$$

其中 a_j 为项目 j 的区分度参数,c_j 为项目 j 的猜测度参数.

项目 j 的信息函数达到最大值时的被试水平值($\theta_j^{\max}$) 为[20]

$$\theta_j^{\max} = b_j + \frac{\ln\left[1 + (1 + 8c_j)^{1/2}\right] - \ln(2)}{1.7a_j}. \quad (3)$$

1.2　新的选题策略

MIS-B 虽然是针对 a 分层法的不足而提出的改良性策略,但仍有一些问题没有得到很好解决,如阶段终止规则有待改善,未考虑内容平衡问题等. 针对 MIS-B 本身并未提出对于内容平衡的解决方法,本文提出了以 MIS-B 为基础的,通过加入改良多项式模型(the modified multinomial model,MMM) 来平衡内容属性的 4MIS-B,其步骤如下: (1) 各内容域在测试中的目标比例(和为 1.0) 形成累积分布. (2) 根据题库中项目的内容属性将整个题库分为 G 个区(即每个内容域为一个区),每区的项目数量 n_g 按照目标比例进行划分,$n_1 + n_2 + \cdots + n_g = n, g = 1,2,\cdots,G$. (3) 依据 $\theta_j^{\max}$,分别将 G 个区中项目按升序排列. (4) 确定分层后题库的层数 S. (5) 按照每 S 题为一块,将每一区划分为 P_g 个块. 由于每区中的项目数量是按目标比例分配的,当同时按照每 S 题目为一块进行划分时,每一区中块的数量仍然符合目标比例,且有 $P_1 + P_2 + \cdots + P_G = n/S$. 将每一块中的项目按 $I_j^{\max}$ 进行升序排列,再将每一块中 $I_j^{\max}$ 最小的项目放入第 1 层,次小的放入第 2 层,以此类推,$I_j^{\max}$ 最大的项目放入第 s 层,$s = 1,2,\cdots,S$. 需要注意的是,每层中的项目仍处于分区状态. 最终每层有 G 个区,共 n/S 个项目. (7) 与 S 个项目层相对应,把测验过程分为 S 个阶段. (8) 在第 s 阶段,生成一个服从均匀分布 $U(0,1)$ 的随机数,用这个随机数和内容域的累积分布比较,确定某一个区(内容域),从第 s 层中将属于这个区的项目依据 $j = \arg\min_{j \in B_q} |\hat{\theta}_j - \theta_j^{\max}|$ 挑选出来实施,其中 B_q 表示剩余题库. (9) 当一个内容域达到它的目标比例后,调整剩下内容域的比例形成新的累积分布,继续一前面的选题步骤直到满足阶段终止规则,转而进行下一阶段. (10) 重复步骤(8) 和(9) 直到测验结束.

在步骤(5) 中,如果每一区中项目数 n_g 正好被 S 整除,则 $P = n_g/S$,每一块中项目数都是 S. 如果 n_g 不能被 S 整除,当 n_g/S 的余数 $\leqslant S/3$ 时,将余下的项目归入最后一块,即 P 仍然等于 n_g/S; 当 n_g/S 的余数 $> S/3$ 时,$P = [n_g/s] + 1$,最后一块中的项目数等于 n_g/S 的余数,其余块中的项目数为 S.

从上面的步骤可以看出,分层后的题库具有 3 个特征: (1) 每一层的内容覆盖与整个题库相似; (2) 每一层 $\theta_j^{\max}$ 的分布与整个题库相似; (3) 每一层 $I_j^{\max}$ 的平均值是递增的. 这与 a-STR 和 BAS 方法想保证在能力估计较粗糙阶段使用区分度小的那些项目,而能力估计越精确的后期使用区分度较大的那一批项目的思想是一致的.

2 CAT 的模拟过程

2.1 Monte Carlo 模拟试验

2.1.1 被试及题库模拟 本试验基于3PLM,并采用 Monte Carlo 模拟方法.

(1) 模拟4个题库,每个题库包含题目数量800题,其题目参数分布为:① 区分度 $a \sim N(1.2, 0.25)$,难度 $b \sim N(0,1)$,猜测度 $c = 0$,a 与 b 相关系数 $r_{ab} = 0$;② 区分度 $a \sim N(1.2, 0.25)$,难度 $b \sim N(0,1)$,猜测度 $c \sim Beta(5,17)$,a 与 b 相关系数 $r_{ab} = 0$;③ 区分度 $a \sim N(1.2, 0.25)$,难度 $b \sim N(0,1)$,猜测度 $c = 0$,a 与 b 相关系数 $r_{ab} = 0.5$;④ 区分度 $a \sim N(1.2, 0.25)$,难度 $b \sim N(0,1)$,猜测度 $c \sim Beta(5,17)$,a 与 b 相关系数 $r_{ab} = 0.5$.

以上每个题库均包含3个内容域(比例为4∶3∶3),第1个内容域包含题库前320道项目,第2个内容域包含题库中间240道项目,第3个内容域包含题库后240道项目.

模拟的被试数量1 000人,能力参数 θ 服从标准正太分布,记为 $\theta \sim N(0,1)$.

2.1.2 试验设计 本文基于上文中4个题库,对以下7种选题策略进行对比研究:①a 分层策略(a-STR);②b 模块化的 a 分层策略(BAS);③ 引入内容平衡的分层策略(STR-C);④ 最大信息量分层策略(MIS);⑤ 最大信息量组块分层策略(MIS-B);⑥ 引入内容平衡的最大信息量组块分层选题策略(4MIS-B);⑦ 最大信息量选题策略(MFI).

本文将 a-STR、BAS 和 STR-C 统称为 A 簇策略,将 MIS、MIS-B、4MIS-B 统称为 M 簇策略. 为 4×7 交叉研究,共28个试验,每个试验均对1 000名被试进行30次 CAT 模拟全过程. 研究采用定长 CAT 测验,测验长度为40题. A 簇策略和 M 簇策略的分层数均分为4,每层200题,测验每个阶段在每层中选10道题.

2.2 CAT 模拟

2.2.1 CAT 全过程模拟 施测过程分为:初始探测阶段和正式测验阶段. 在初始探测阶段,采用从剩余题库中随机抽取试题给被试作答,直到作答题数不少于3且作答总分既不为0分也不为满分时,结束初始探测阶段. 根据被试的作答得分向量,估计能力初值,进入正式测验阶段. 在正式测验阶段,针对不同的选题策略,从剩余题库中调用与被试当前能力值匹配的项目. 被试作答完成后,根据作答得分向量估计被试当前能力值. 直到测验长度达到40,结束测验. 文中的模拟数据(包括项目参数、被试能力参数)和 CAT 的施测程序均采用 R version 2.15.0 (64-bit) 编写运行.

2.2.2 被试作答模拟和能力估计方法 通过以下方法模拟被试得分[21]:根据被试能力真值 θ 和当前所选择的项目 j 的参数,由公式(1)计算被试在第 j 个项目上的答对概率 $P_j(\theta)$,再产生一个随机数 $r(0 \leqslant r \leqslant 1)$,若 $r \leqslant P_j(\theta)$,则该被试在第 j 个项目上得1分,否则得0分. 此外,假设被试能力的先验分布为标准正太分布,采用贝叶斯期望后验估计法(EAP)估计被试能力.

2.3 评价指标

选题策略的优劣直接关系到 CAT 的质量,当其它条件固定仅改变选题策略时,对 CAT 的评价实际上就是对选题策略的评价[21],故本文采用以下评价指标:

(1) 平均偏差 $Bias = \sum_{i=1}^{N} (\hat{\theta}_i - \theta_i)/N$,

(2) 平均绝对偏差 $ABS = \sum_{i=1}^{N} |\hat{\theta}_i - \theta_i|/N$,

(3) 均方误差 $MSE = \sum_{i=1}^{N} (\hat{\theta}_i - \theta_i)^2/N$,

(4) 卡方统计量 $\chi^2 = \sum_{j=1}^{M} (A_j - \bar{A}_j)^2/\bar{A}_j$,$\bar{A}_j = Q/M$,

(5) 重叠率 $OR = Ms^2/Q + Q/M$,

其中 N 为被试总数;θ_i 为被试能力真值;$\hat{\theta}_i$ 为被试能力估计值;Q 表示测验长度;M 为题库大小;A_j 是第 j 题曝光率,其计算公式为 A_j = 第 j 题被使用的次数 $/N$;s^2 是项目曝光率的方差. 重叠率定义为任意2个随机选取的被试的重叠项目数的期望值除以考试的长度. 对于大样本,测验重叠率近似值为 OR.

以上5个指标均是越小越好. OR 越小表明该方法选题重叠率越低;ABS 和 MSE 越小,说明估计的精度越高;$Bias$ 越小,则表明该方法越接近无偏;χ^2 值越小说明曝光率越均匀,CAT 的安全性越好.

3 试验结果分析

将被试群体参加模拟的 CAT,比较7种选题策略在上述5个评价指标的表现. 综合目前有关的研究来看,在同一设定好的题库中,各种选题策略总体上并无明显优劣之分,只是相对于不同的测验要求各有优势. 此外,当测验长度为40时,各种选题策略之间的差异很小[22],这从表1 ~ 表4中的结果也可以看出. 因此在各方法总体差异较小这一前提下,综

合表 1 ~ 表 4 的试验结果对 7 种选题策略进行对比分析.

3.1　区分度 a 与难度 b 相关系数 $r_{ab}=0$、猜测度 $c=0$

MFI 在 *ABS* 和 *MSE* 方面表现最好,但在 χ^2 和 *OR* 方面表现最差. 在理想题目参数情况下,*a*-STR 在 *ABS* 和 *MSE* 方面表现是除 *MFI* 外最好的,χ^2 和 *OR* 方面在 A 簇策略中最差. 对于 *BAS* 来说,由于理想项目参数情况并不符合它提出时的前提条件(a 和 b 在实际题目中存在相关性等) ,因此并没有发挥它的优势. 当 $c=0$ 时,$I_j^{\max}=1.7^2a^2/4$,$\theta_j^{\max}=b$,*M* 簇策略和 A 簇策略对题库的排序(和组块) 是相同的,其选题效果也是一样的,见表 1.

表 1　区分度 a 与难度 b 相关系数 $r_{ab}=0$、猜测度 $c=0$ 情况的试验结果

选题策略	*ABS*	*MSE*	*Bias*	χ^2	*OR*
a-STR	0.128	0.026	0.001	9.926	0.062
BAS	0.129	0.027	-0.001	5.857	0.057
STR-C	0.129	0.026	0.002	5.019	0.055
MIS	0.128	0.025	-0.003	9.967	0.068
MIS-B	0.129	0.028	0	5.773	0.061
4MIS-B	0.130	0.027	-0.003	4.930	0.059
MFI	0.114	0.020	0.003	145.960	0.399

3.2　区分度 a 与难度 b 相关系数 $r_{ab}=0.5$、猜测度 $c=0$

在实际的题目编制中,难度 b 和区分度 a 之间是正相关的. 这种相关性可能造成实际的选题情况与 a 分层法的预计不同,这也是 blocking(组块) 策略提出的原因. 从表 2 可以看出,当 $r_{ab}=0.5$ 时,A 簇策略中的 BAS 在 6 个评价指标上均优于 *a*-STR; M 簇策略中的 MIS-B 同样在 6 个评价指标上优于 MIS. 同样在 $c=0$ 的情况下,M 簇策略和 A 簇策略对题库的排序(和组块) 是相同的,其选题效果也是一样的. MFI 的情况与本文 3.1 节情况类似.

表 2　区分度 a 与难度 b 相关系数 $r_{ab}=0.5$、猜测度 $c=0$ 情况的试验结果

选题策略	*ABS*	*MSE*	*Bias*	χ^2	*OR*
a-STR	0.138	0.029	0.002	22.725	0.079
BAS	0.132	0.028	-0.001	6.377	0.058
STR-C	0.131	0.028	0.001	5.956	0.055
MIS	0.139	0.030	-0.002	30.303	0.095
MIS-B	0.131	0.030	-0.002	7.129	0.064
4MIS-B	0.131	0.028	-0.001	6.163	0.060
MFI	0.117	0.022	0	140.025	0.379

3.3　区分度 a 与难度 b 相关系数 $r_{ab}=0$、猜测度 $c\sim Beta(5,17)$

在题库 3 中,当项目猜测度 c 不为 0 时,M 簇策略在 ABS 和 MSE 方面略优于 A 簇策略,这与 J. R. Barrada 等[19] 的结论(不包括 4MIS-B) 一致. 但他们没有明确的是,在 χ^2 本文方面 M 簇策略较 A 簇策略稍差. MFI 的情况与本文 3.1 节情况类似.

表 3　区分度 a 与难度 b 相关系数 $r_{ab}=0$、猜测度 $c\sim Beta(5,17)$ 情况的试验结果

选题策略	*ABS*	*MSE*	*Bias*	χ^2	*OR*
a-STR	0.161	0.042	0.002	12.683	0.066
BAS	0.162	0.043	-0.002	6.936	0.059
STR-C	0.163	0.043	0.002	6.010	0.056
MIS	0.159	0.041	-0.003	14.616	0.066
MIS-B	0.159	0.042	0	7.317	0.065
4MIS-B	0.161	0.043	-0.001	6.333	0.062
MFI	0.132	0.028	-0.001	152.130	0.422

3.4 区分度 a 与难度 b 相关系数 $r_{ab}=0.5$、猜测度 $c\sim Beta$ (5,17)

题库 4 的项目参数分布情况是这 4 个题库中最不理想的. 从表 4 可以看到所有选题策略的 6 个评估指标值都是最大的. 从整体上看,M 簇策略在 *ABS* 和 *MSE* 方面仍优于 A 簇策略,在 χ^2 和 *OR* 方面仍稍劣于 A 簇策略. MFI 的情况与本文 3.1 节情况类似.

表 4 区分度 a 与难度 b 相关系数 $r_{ab}=0.5$、猜测度 $c\sim Beta$ (5,17) 情况的试验结果

选题策略	*ABS*	*MSE*	*Bias*	χ^2	*OR*
a-STR	0.170	0.048	-0.001	25.784	0.083
BAS	0.168	0.046	-0.001	7.656	0.060
STR-C	0.167	0.046	0	7.008	0.059
MIS	0.165	0.044	-0.002	21.987	0.078
MIS-B	0.163	0.043	-0.003	8.975	0.068
4MIS-B	0.162	0.043	-0.001	7.210	0.062
MFI	0.136	0.030	-0.004	150.887	0.418

3.5 整体分析

从本研究整体的结果来看:(1) 在能力估计精准度方面,MFI 比 6 种分层选题策略都好:(1) 区分度 a 与难度 b 的相关性会增大所有选题策略的 5 种评估指标的值. 其中 a-STR 和 *MIS* 的 χ^2 指标值增幅较大;(2) 猜测度 c 的引入会增大所有选题策略的 5 种评估指标值;(3) 在平均偏差 *Bias* 方面,7 种策略的 *Bias* 指标值基本接近 0;(4) 在 χ^2 方面,分层选题策略中的 a-STR 和 *MIS* 表现较差,说明他们的项目曝光率均匀性较差;(4) Blocking(组块) 策略在 $r_{ab}=0$ 和 $r_{ab}=0.5$ 时均能降低 χ^2 和 *OR* 指标值;(5) M 簇策略和 A 簇策略相比,其在能力估计精准度上稍有提升,在项目曝光率均匀性和重叠率方面稍有下降;(6) 4 个题库中,4MIS-B 在 5 种评价指标上表现均较好.

4 讨论

首先,本文从内容平衡方面考虑,优化了 MIS-B 策略,使得选题更均匀、合理. 将新的选题策略与 a-STR、BAS、STR-C、MIS、MIS-B、MFI 比较,实验数据显示,引入内容平衡的最大信息量组块分层策略(4MIS-B) 在保持 MIS-B 原有能力估计精准度前提下降低了题目重叠率、提高了题库使用的和题目曝光率的均匀性. 这与 YiQ [23] 的结论一致.

其次,从本研究的数据结果上讲,M 簇策略和 A 簇策略在整体上差异很小,可针对不同的测验要求酌情选择. 但从观念上讲,M 簇策略考虑到了 3PLM 独有的猜测度 c 参数,这是值得提倡的.

最后,本文仍有些许值得改进的地方,如:(1) 假设的内容域分布和实际情况是有差异的;(2) 相比于 Q. Yi 等 [23] 和 J. R. Barrada 等 [19] 的研究,本研究每个题库中题量稍大,导致每层中题目数量较充足,可能也是导致各分层策略之间的差异性较小的原因之一;(3) 本研究仍为定长 CAT,且子题库为等题量划分,而程小杨等 [24] 曾指出递减的各层子题库划分方法能有效地提高 a-STR 的测验效率,如果比例划分合适还能降低项目曝光率. 因此,还有必要探究如何修改 M 簇策略使其适用于不定长 CAT 以及其在子题库题量不平衡情况下的表现;(4) 本研究仍基于二级评分情况,不适用于多级评分的情况等. 这些都有待做进一步研究和改进.

5 参考文献

[1] Weiss D J. Improving measurement quality and efficiency with adaptive testing [J]. Applied Psychological Measurement,1982,6(4):473.

[2] Weiss D J,Kingsbury G. Application of computerized adaptive testing to educational problems [J]. Journal of Educational Measurement,1984,21(4):361-375.

[3] van der Linden W J,Glas C A W. Computerized adaptive testing: theory and practice [C]. MA: Kluwer,2000:27-52.

[4] Cheng Ying,Chang Huahua. The maximum priority index method for severely constrained item selection in computerized adaptive testing [J]. British Journal of Mathematical and Statistical Psychology,2009,62:369-383.

[5] Deng Hui,Ansley T,Chang Huahua. Stratified and maximum information item selection procedures in computer adaptive testing [J]. Journal of Educational Measurement,2010,47:202-226.

[6] 毛秀珍,辛涛. 计算机化自适应测验选题策略述评[J]. 心理科学进展,2011,19(10):1552-1562.

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

[8] Chang Huahua, Ying Zhiliang. A global information approach to computerized adaptive testing [J]. Applied Psychological Measurement, 1996, 20: 213-229.
[9] van der Linden W J. Bayesian item selection criteria for adaptive testing [J]. Psychometrika, 1998, 63: 201-216.
[10] Sympson J B, Hetter R D. Controlling item-exposure rates in computerized adaptive testing [C]. CA: Navy Personnel Research and Development Center, 1985: 973-977.
[11] Revuelta J, Ponsoda V. A comparison of item exposure control methods in computerized adaptive testing [J]. Journal of Educational Measurement, 1998, 35: 311-327.
[12] Segall D O. A sharing item response theory model for computerized adaptive testing [J]. Journal of Educational and Behavioral Statistics, 2004, 29: 439-460.
[13] Chang Huahua, Ying Zhiliang. *a*-stratified multistage computerized adaptive testing [J]. Applied Psychological Measurement, 1999, 23(3): 211.
[14] Chang Huahua, Qian Jiahe, Ying Zhiliang. *a*-Stratified multistage computerized adaptive testing with b blocking [J]. Applied Psychological Measurement, 2001, 25(4): 333.
[15] Kingsbury G G, Zara A R. Procedures for selecting items for computerized adaptive tests [J]. Applied Measurement in Education, 1989, 4: 359-375.
[16] Stocking M L, Swanson L. A method for severely constrained item selection in adaptive testing [J]. Applied Psychological Measurement, 1993, 23: 277-292.
[17] van der Linden W J, Reese L M. A model for optimal constrained adaptive testing [J]. Applied Psychological Measurement, 1998, 22: 259-270.
[18] 漆书青,戴海琦,丁树良. 现代教育与心理测量学原理 [M]. 北京: 高等教育出版社, 2002: 89.
[19] Barrada J R, Mazuela P, Olea J. Maximum information stratification method for controlling item exposure in computerized adaptive testing [J]. Psicothema, 2006, 18: 156-159.
[20] Hambleton R K, Swaminathan H. Item response theory: Principles and applications [M]. MA: Kluwer, 1985.
[21] 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.
[22] Barrada J R, Olea J, Ponsoda V, et al. A method for the comparison of item selection rules in computerized adaptive testing [J]. Applied Psychological Measurement, 2010, 34(6): 438.
[23] Yi Q, Chang Huahua. *a*-Stratified CAT design with content blocking [J]. British journal of mathematical and statistical psychology, 2003, 56(2): 359-378.
[24] 程小杨,丁树良. 子题库量不平衡的按 *a* 分层选题策略 [J]. 江西师范大学学报: 自然科学版, 2011, 35(1): 5-9.

The Maximum Information Stratification Method with Content Balancing in Computerized Adaptive Testing

ZHAN Pei-da, WANG Li-jun*, YANG Wei-min
(Department of Psychology, Zhejiang Normal University, Jinhua Zhejiang 321004, China)

Abstract: In 0-1 scored CAT, a new item selection strategy is proposed to improve the MIS-B method by introducing the Modified Multinomial Model. The results of Monte Carlo simulations show that compared with MIS-B, the approach proposed in this paper can reducing item overexposure rate, balancing item usage within the item bank, and maintaining measurement precision.

Key words: CAT; content balancing; MIS-B; modified multinomial model

(责任编辑: 冉小晓)