文章编号:1000-5862(2021)04-0384-06

区分度与测验进程相匹配的 CAT 选题策略

李　佳[1],丁树良[1*],况天昊[2]
(1. 江西师范大学计算机信息工程学院,江西 南昌　330022;2. 江西科技学院信息工程学院,江西 南昌　330098)

摘要: 选题策略是计算机化自适应测验(CAT)的核心. 该文提出了一种新的选题策略,是一种相对严格的"升 a"方法,它选择区分度参数的百分等级尽可能接近测验进程的项目,而且还可以通过调整控制参数的取值来满足不同测验场景的需求. Monte Carlo 实验结果表明:该方法在测验精度、项目曝光率控制和题库利用率等方面均表现良好.

关键词: 计算机化自适应测验;选题策略;项目曝光控制;题库利用率;控制参数
中图分类号: B 841　　**文献标志码:** A　　**DOI:** 10.16357/j.cnki.issn1000-5862.2021.04.10

0　引言

计算机化自适应测验(Computerized Adaptive testing,CAT)是项目反应理论(Item Response Theory,IRT)的重要应用之一[1]. CAT 作为一种连续性测验,被试在完成测验后能立即得出能力水平值,因此 CAT 被认为是最有效的测验形式之一,并且被应用到各种测验领域中. 从 20 世纪 90 年代开始,CAT 已经被广泛应用于美国中小学、政府机关、专业机构中的各种不同考试中. E. S. Quellmalz 等[2]指出,对于高风险的测验,需要慎重使用 CAT. 其最主要的原因是传统的最大信息量选题策略[3]过于关注被试能力估计精度,每次都从题库中选择具有最大信息量的项目,而项目信息量与项目区分度的平方成正比. 因此,高区分度的项目被频繁选出,导致该类项目的曝光率过高,给整个 CAT 带来较高的风险. 事实上,最大信息量选题策略也会导致大量低区分度的项目不能被较好地利用,造成题库中低区分度项目的浪费.

a 分层方法[4]可以有效阻止项目的过度曝光,而且该方法操作简单,只要降低一点能力估计精度,就可以降低高风险项目的曝光率. 该方法事先将题库分成 K 层(一般来说,K 为固定参数,在各模拟实验中一般取 $K=4$),第 1 层项目的区分度参数最小,第 2 层项目的区分度参数次之,…,第 K 层的项目具有最大的区分度参数. 在题库中的项目必须根据项目区分度参数被分匹到且仅被分匹到某个层中,不同层的项目没有交集,每层项目合并起来正好是整个题库,这是题库项目集合的一个划分[5]. 但是常数 K 的选择并没有固定的标准,一般在每一层选取固定数量的项目或者固定信息量的项目,且这也没有固定标准. 通过模拟实验发现:a 分层方法仍然会存在 50% 以上过低曝光的项目,造成题库的浪费;特别是当题库中项目发生改变时,a 分层方法还必须将题库重新进行分层后才能使用. 因为随着计算机化自适应测验过程的进行,被试能力估计越来越准确,所以随着被试施测项目的增多,只有使用区分度更大的项目才能得到更高的测验效率. a 分层方法的初衷是按照项目区分度从小到大的顺序选择项目,但在 a 分层方法实施过程中只能保证区分度参数按照所划分的层次逐步上升,不能保证在整个测验过程中每一个项目的区分度参数呈现一个上升状态.

李萍等[6]提出的自动控制区分度作用的选题策略可以将项目信息量、曝光因子和区分度函数巧妙地结合在一起,可以较好地控制项目曝光率,且具有较高的测验精度. 但该方法没有关注题库利用率,

收稿日期: 2021-03-18
基金项目: 国家自然科学基金(62067004,62067005,61967009)资助项目.
作者简介: 李　佳(1979—),女,江西南昌人,讲师,主要从事计算机辅助教学和心理测量方面的研究. E-mail:1276676143@qq.com
通信作者: 丁树良(1949—),男,江西樟树人,教授,主要从事计算机辅助教学和心理测量方面的研究. E-mail:ding06026@163.com

即没有提升低曝光率的项目所占比例,并且对于当自动控制区分度因子中指数部分系数取值为 2 时没有给出具体的理由.

1　项目反应理论

在 IRT 中,假定同一被试对各个项目的作答是相互独立的(即局部独立性). 被试 α 在 CAT 施测过程中作答反应向量为 $\boldsymbol{U}=(u_{\alpha_1},u_{\alpha_2},\cdots,u_{\alpha_L})$,$L$ 为施测项目数; 似然函数为 $L(\boldsymbol{U}|\theta_\alpha)=\prod_{j=1}^{L}P_{\alpha_j}^{u_{\alpha_j}}(1-P_{\alpha_j})^{1-u_{\alpha_j}}$,其中 u_{α_j} 表示被试 α 对项目 j 的反应,取值为 0 或 1,分别表示答对或答错该项目. 在 IRT 框架下,能力为 θ_α 的被试正确作答项目 j 的概率 P_{α_j} 可以取不同的计算式,常见的模型是 3 参数 Logistic 模型,其项目反应函数为 $P_{\alpha_j}=c_j+(1-c_j)/(1+\exp(-D\alpha_j(\theta_\alpha-b_j)))$,其中 α_j 为项目 j 的区分度参数,b_j 为项目 j 的难度参数,c_j 为项目 j 的猜测度参数,$D=1.7$. 在该条件下,3PLM 的项目 Fisher 信息量为 $I_j(\theta_\alpha)=D^2a_j^2(1-c_j)/((c_j+e^{Da_j(\theta_\alpha-b_j)})(1+e^{-Da_j(\theta_\alpha-b_j)})^2)$. 由局部独立性假设可得测验信息量为 $I(\theta_\alpha)=\sum_{j=1}^{L}I_j(\theta_\alpha)$. Fisher 测验信息量(也简称为测验信息量)等于测量误差方差的倒数,这是最大信息量选题策略的理论基础. 但因为项目信息量与区分度平方成正比,所以最大信息量选题策略在实践中会造成高区分度项目频频使用,而低区分度项目较少使用甚至不用,从而造成题库浪费,且危及考试安全. 研究人员一方面知道测验信息量的重要性,另一方面也意识到一味追求使用高信息量项目对题库安全性的威胁,自从 a 分层方法被提出以来,一些研究人员就从对测验信息量既使用又限制这种思路入手进行了相当多的讨论. 本文特别介绍其中 2 种策略: a 分层方法和自动控制区分度作用的选题策略,并且沿着这个思路引入一种新的选题策略.

2　控制项目曝光率的选题策略

2.1　a 分层方法

题库根据项目的区分度参数被划分成 K 层,具有最小的区分度项目被划分在第 1 层,具有第 2 小的区分度项目被划分在第 2 层,以此类推,具有最大的区分度项目被划分在第 K 层. CAT 过程也被分成相应的 K 个阶段,在定长测验中 L_j 个项目来自第 j 层,$L_1+L_2+\cdots+L_K$ 正好为测验总长度 L. 在不定长测验中,从第 j 层中选出的项目信息量总和为 I_j,$I_1+I_2+\cdots+I_L$ 不小于预先指定的测验信息量. 每层项目的选取方法有 2 种: (i) 根据当前被试的能力值,选取最大信息量的项目 $\max_{j\in R_a}(I_j(\theta_\alpha))$,其中 R_a 为被试当前层的剩余题库; (ii) b 匹配法,即选取当前能项目难度参数最接近的项目 $\min_{j\in R_a}\{|b_j-\theta_\alpha|\}$.

由于严格地曝光率限制了高区分度项目的使用,会降低能力估计值精度. 近年来有学者提出的动态分层方法[7] 既能够尽量控制曝光率,又能尽量保证能力估计精度,不失为一种权衡曝光率和测验精度的好方法. 但是该方法划分区块的大小等于测验长度,对于每一个被试而言,这实质上等于一个分块仅仅包含一个项目,也就是等于不分块.

2.2　自动控制区分度作用的选题策略

$e_{cf}(j)=m_j/\overline{m}$ 被称为题库中第 j 个项目的曝光因子,其中当第 i 个考生参加考试时,m_j 表示前 $i-1$ 个考生使用第 j 个项目的总次数,$\overline{m}$ 表示题库中所有项目被前 $i-1$ 个考生使用的平均次数,即 $\overline{m}=(\sum_{j=1}^{M}m_j)/M$,$M$ 为题库中的题数.

在定长测验中,自动区分度因子 $a(j,i)=a_j^{2(L-L(i))/L}$; 在不定长测验中,自动区分度因子 $a(j,i)=a^{2(I_{nfor}-i_{nf}(i))/I_{nfor}}$,其中 L 表示测验长度,$L(i)$ 表示第 i 个被试当前已作答的项目个数,$i_{nf}(i)$ 表示第 i 个被试当前的测验信息量,I_{nfor} 表示总的测验信息量. 选择项目,即考虑 $\max_{j\in R_a}(I_j/(e_{cf}(j)\,a(j,i)))$.

2.3　测验进程比与项目区分度百分等级相匹配的选题策略

先介绍 2 个名词: (i) 被试的作答进程比,即对于定长 CAT,设测验长度为 L,被试已经作答 j 题,作答进程比为 j/L; 而对于不定长 CAT,被试作答对应的信息量与预定的终止信息量之比为作答进程比; 被试作答进程比简称为进程比. (ii) 项目区分度的百分等级,即将题库中的所有项目按区分度参数从小到大进行排序,确定每个项目的排序位置(次序),项目的次序与题库中题目量(容量)之比,被称为该项目的区分度的百分等级,简称为区分度百分等级. 如题库中共有 520 个项目,若项目 j 的区分度按照从小到大排在第 26 位(次序),则项目 j 的区分度百分等级为 $26/520=0.05$.

本文提出的新选题策略的本质是使被试测验进程比与被选项目区分度百分等级尽量匹配,这种匹配的程度作为 Fisher 信息量的调节因子,以达到均匀调用题库中的项目,尽可能平衡项目的曝光率,提高项目的利用率,降低题库建设的成本,提升题库的

安全性的目标. 而使用 Fisher 信息量的目的是提高被试能力估计精度.

J. R. Busemeyer 等[8] 指出当前测验阶段和项目区分度参数在题库中的百分等级"相似"(similarity)时,具有更有效的测量结果. 将区分度参数的百分等级和当前测验进程尽量接近,就是这种相似性. 本文分定长试验和不定长试验 2 种情况进行讨论.

情况 1 测验长度为 L 的定长测验. 设在 CAT 过程中被试 α 已经做了 j 个题目,则第 $j+1$ 个题目将选择 $\max\limits_{j\in R_a}(I(\theta_\alpha)\exp(-\lambda\mid j/L-r(a_j)\mid))$ 的项目(若有多个项目同时达到最大值,则随机选取 1 个),其中 $I_j(\theta_\alpha)$ 是被试在当前能力估计值下该项目的信息量,$r(a_j)$ 是该项目的区分度相对秩,其值为 0 ~ 1; L 是指测验长度,j/L 是被试作答进程比,其取值为 0 ~ 1,当项目区分度相对秩和被试作答进程比越接近(即项目区分度相对秩和被试作答进程比一致)时,$\exp(-\lambda\mid j/L-r(a_j)\mid)$ 的值越接近 1,测验效率越高. 此时,最大化 $(I(\theta_\alpha)\exp(-\lambda\mid j/L-r(a_j)\mid))$ 既可以提高测验精度,又可以较好地控制项目曝光率. 称 λ 为控制参数,其取值必须大于 0; λ 是比较项目区分度相对秩和被试作答进程比的灵敏度参数,λ 的取值应根据不同的测验模型和不同的测验要求,由测验专家进行确定. 一般而言,λ 值越大,$\exp(-\lambda\mid j/L-r(a_j)\mid)$ 对整个选题过程的影响更占优势,曝光控制也就越严格,项目曝光率控制得更好,题库更安全; 反之,λ 值越小,$\exp(-\lambda\mid j/L-r(a_j)\mid)$ 的值越接近 1,项目信息量 $I_j(\theta_\alpha)$ 更占优势,测验精度越高,能力估计越准确. 事实上,当 $\lambda=0$ 时,该选题策略就退化为最大信息量选题.

情况 2 不定长测验. 测验信息量为 I_{nfor},设在 CAT 过程中被试 α 已经做了 j 个题目,则第 $j+1$ 个题目选择 $\max\limits_{j\in R_a}(I(\theta_\alpha)\exp(-\lambda\mid i_{nf}(\alpha)/I_{nfor}-r(a_j)\mid))$ 的项目,其中 $i_{nf}(\alpha)$ 表示被试 α 当前的测验信息量,I_{nfor} 表示总测验信息量.

总体来说,在测验初期,被试作答进程比的值比较小,CAT 会自动选择低区分度的项目,随着测验过程的进行,选择项目的区分度会逐渐升高,所以,新的选题策略实质也是一种升 a 方法,且是一种连续的升 a 过程.

3 比较不同选题策略的 CAT 表现

3.1 被试及题库模拟

模拟实验 1 蒙特卡罗模拟产生 1 000 个被试,被试能力真值均服从均值为 0、方差为 1 的标准正态分布; 在 3PLM 模型下设计题库,所有试验模拟条件同参见文献[9]. 题库结构为模拟生成 520 个项目且满足条件 $\ln a\sim N(0,1)$, $b\sim N(0,1)$, $c\sim \mathrm{Beta}(5,17)$, $0.2<a<2.5$, $-3.5<b<3.5$, $|a-b|<4.0$, $c<0.4$. 题库的项目数据见表 1.

表 1 题库的项目数据

项目数据	区分度	难度 b	猜测度 c
平均值	1.001 30	-0.006 464 7	0.223 380
标准差	0.608 37	0.979 380 0	0.080 761

3.2 模拟 CAT 的施测过程

本文不考虑内容平衡[10] 和机会红利[11] 对 CAT 的影响,简化 CAT 设计为被试的能力初值为 0,参与比较的 6 种选题策略为

(i) 最大 Fisher 信息量选题策略(MFI);
(ii) 随机化选题策略(RS);
(iii) a 分层最大信息量选题策略(AST);
(iv) a 分层难度参数匹配选题策略(ASB);
(v) 自动控制区分度选题策略(DB);
(vi) 新选题策略(MPD),控制参数 $\lambda=2$.

根据不同的选题策略采取 NMLE 方法[12] 对能力进行估计,分定长测验和不定长测验 2 种情况. 定长测验的测验长度为 40; 不定长测验的测验信息量为 16. 在 a 分层中,题库被分为 4 层: 定长测验每层选 10 题; 不定长测验各层的累积信息量的比例为 1∶1∶1∶1,即在每层测验信息量达到 4 时退出.

4 比较不同控制参数的取值对新选题策略的影响

4.1 被试及题库模拟

模拟实验 2 蒙特卡罗模拟产生 1 000 个被试,被试能力真值均服从均值为 0、方差为 1 的标准正态分布; 在 3PLM 模型下设计 4 种题库,均模拟生成 520 个项目.

题库 1: $\ln a\sim N(0,1)$, $b\sim N(0,1)$, $c\sim \mathrm{Beta}(5,17)$;

题库 2: $a\sim U(0.2,2.5)$, $b\sim N(0,1)$, $c\sim \mathrm{Beta}(5,17)$;

题库 3: $\ln a\sim N(0,1)$, $b\sim U(-3,3)$, $c\sim \mathrm{Beta}(5,17)$;

题库 4: $a\sim U(0.2,2.5)$, $b\sim U(-3,3)$, $c\sim \mathrm{Beta}(5,17)$.

4.2　模拟 CAT 的施测过程

取被试的能力初值为 0,采用新选题策略(MPD)进行施测,控制参数 λ 分别取 1、2、3.

采取 NMLE 方法对能力进行估计,定长测验的测验长度为 40,不定长测验测验信息量为 16.

5　评价指标

测验平均绝对离差(A_{BS}):

$$A_{BS} = \sum_{i=1}^{N} |\hat{\theta}_i - \theta_i| / N,$$

测验均方根误差(R_{MSE}):

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

卡方检验统计量(χ^2):

$$\chi^2 = \sum_{j=1}^{M} (e_{r_j} - L/M)^2 / (L/M),$$

项目过低曝光率(L_E)[11]: $L_E = \sum_{i=1}^{M} L_{E_i} / M$,测验效率($T_E$): $T_E = \sum_{i=1}^{N} i_{nf}(i) / \sum_{i=1}^{N} L_i$,其中 N 为被试总人数,θ_i 为第 i 个被试的能力真值,$\hat{\theta}_i$ 为第 i 个被试的能力估计值,M 为题库中的项目数,L 为测验长度(在不定长测验中,L 为平均测验长度),e_{r_j} 为第 j 个项目的曝光率,L_{E_i} 为题库中曝光率小于 0.05 的项目数,$i_{nf}(i)$ 为被试 i 的测验信息量,L_i 为被试 i 的测验长度.

测验平均绝对离差(A_{BS})和测验均方根误差(R_{MSE})都表明测验能力估计的准确性,A_{BS} 为能力估计的系统偏差,R_{MSE} 为能力估计值和真实值的随机误差,它们都是评价测验准确性的常用指标,A_{BS} 和 R_{MSE} 越接近 0,表示能力估计越准确;χ^2 为整个题库的使用均匀性,χ^2 越大表明项目调用越不均匀,存在大量高曝光率项目,测验存在高风险. 理想的项目曝光场景是题库中所有的项目具有相同的项目曝光率,即每个项目有 L/M 的概率被选中,理论上此时卡方值为 0. 否则,若所有被试做同样的题目,则卡方值最大. 如 $L = 40$,$M = 520$,若每个被试都是做这 40 个相同的题目,则这 40 个题目的 e_{r_j} 值为 1,而其他 480 个题目的 e_{r_j} 为 0,L/M 约为 0.077,$\chi^2 = 480 \times (0 - 0.077)^2 / 0.077 + 40 \times (1 - 0.077)^2 / 0.077 \approx 479.52$. 项目曝光率 L_E 的值越小表明题库利用率越高,低区分度项目被充分利用,这不会造成题库的浪费. 在定长测验中每个被试的测验长度一定,测验精度越高,测验效率 T_E 越大; 在不定长测验中每个被试的测量精度类似,早达到测验精度的被试所需测验长度更短,测验效率 T_E 值更大,而晚达到测验精度的被试所需测验长度就更长,测验效率 T_E 值更小. 因此,测验效率 T_E 的值越大越好.

6　实验结果及其分析

在实验 1 中的定长测验结果如表 2 所示; 在实验 1 中的不定长测验结果如表 3 所示; 在实验 2 中定长测验结果如表 4 所示; 在实验 2 中不定长测验结果如表 5 所示.

表 2　在定长测验中 6 种选题策略的表现

选题策略	A_{BS}	R_{MSE}	χ^2	L_E	T_E
MFI	0.118	0.222	204.200	0.649	1.390
RS	0.286	0.385	0.977	0.000	0.228
AST	0.133	0.314	66.620	0.543	1.293
ASB	0.148	0.336	58.240	0.561	0.760
DB	0.121	0.260	17.580	0.315	1.126
MPD	0.123	0.265	12.890	0.109	1.141

从表 2 可以看出,在定长测验中,最大信息量(MFI)选题具有最高的测验精度,但是项目曝光率非常高,题库利用率也是最低的,题库安全性差; 随机化选题(RS)策略虽然项目曝光率最低,题库利用率最高,但是测验精度太低,事实上在正常的 CAT 中是不会采用这种选题策略的,不过可以作为项目曝光控制的对照. a 分层方法的最大信息量(AST)选题策略和 b 匹配选题(ASB)策略的测量精度高于随机选题策略,但低于不分层的自动控制区分度(DB)的选题策略和新选题(MPD)方法. 其原因是在每一层选题时,并不能做到严格按升 a 的方式选题,测量精度不会逐步提高; 自动控制区分度选题策略和新选题策略的测量精度相当,都是牺牲测量精度来提高题库安全性和题库利用率. 在题库安全性方面,因为最大信息量选题策略每次只会选择信息量高的项目,所以这些高信息量的项目在被试中频繁出现,导致 χ^2 值较高,题库安全性较差,并且存在大量曝光率低于 5% 的项目,题库利用率较低;

随机化选题策略由于不考虑测验精度,所以题目使用均匀,χ^2 值较小,不存在曝光率低于 5% 的项目,题库利用率较高. a 分层的 2 种选题策略因没有和测验过程相结合,χ^2 值偏高,低曝光率的项目也超过 50%,题库利用率偏低. 在新选题策略(MPD)中被试的作答进程比和项目区分度的相对秩高度一致,由于题库中项目的使用非常均匀,所以既可有效降低高曝光率项目的曝光度,又能提高低曝光率项目的使用机会,因此 χ^2 值较自动控制区分度选题策略更低,曝光率过低的项目也能被较好地利用,题库利用率更高. 在测验效率方面,最大信息量选题策略最大,其次是 a 分层最大信息量选题,最后是新选题策略. 综合能力估计准确性、题库安全性和题库利用率,新的选题方法在这 6 种选题策略中表现更好一些.

表 3 在不定长测验中 6 种选题策略的表现

选题策略	A_{BS}	R_{MSE}	χ^2	L_E	T_E
MFI	0.176	0.237	249.900	0.639	1.751
RS	0.305	0.397	0.926	0.000	0.272
AST	0.233	0.323	55.930	0.554	1.249
ASB	0.258	0.350	64.640	0.608	0.752
DB	0.211	0.248	16.630	0.279	1.120
MPD	0.217	0.251	12.280	0.100	1.153

从表 3 可以看出,在不定长测验中,试验结果和定长试验类似,但是测验精度总体来说更差一些. 因为不定长测验的测验平均长度短于定长测验的固定测验长度,所以在测验效率方面不定长测验总体表现比定长测验更好. 因此,在 CAT 中采用不定长测验也是合情合理的.

表 4 新选题策略当控制参数取不同值时在定长测验中的表现

控制参数	题库 1		题库 2		题库 3		题库 4	
	A_{BS}	χ^2	A_{BS}	χ^2	A_{BS}	χ^2	A_{BS}	χ^2
$\lambda = 1$	0.107	14.56	0.106	16.36	0.138	17.25	0.111	20.99
$\lambda = 2$	0.123	12.89	0.108	14.23	0.142	15.30	0.113	19.61
$\lambda = 3$	0.125	11.91	0.120	13.57	0.151	14.03	0.127	18.65

表 5 新选题策略当控制参数取不同值时在不定长测验中的表现

控制参数	题库 1		题库 2		题库 3		题库 4	
	A_{BS}	χ^2	A_{BS}	χ^2	A_{BS}	χ^2	A_{BS}	χ^2
$\lambda = 1$	0.197	16.64	0.205	17.35	0.233	17.01	0.201	18.94
$\lambda = 2$	0.217	12.28	0.207	13.26	0.247	16.81	0.202	18.77
$\lambda = 3$	0.228	12.27	0.218	11.98	0.264	16.69	0.230	16.00

从表 4 和表 5 的纵向结果来看: 在相同题库下,控制参数 λ 的不同取值对新选题策略在 CAT 中的实践效果是不一样的,控制参数 λ 越大能力估计的准确性越低,但 χ^2 值越小项目曝光控制越好,考试越安全; 控制参数 λ 越小能力估计的准确性越高,但是 χ^2 值较大项目曝光率控制越差,测验精度越高. 从表 4 和表 5 还可以看出: 相同的控制参数和相同的选题策略对于不同的题库试验结果也是不一样的,即新选题策略和题库具有一定的关联性. 因此,对于不同的 IRT 模型、不同的题库类型,应根据测量专家的意见选取不同的控制参数,以得到更好的测量结果.

7 讨论

在 CAT 考试过程中,特别是高风险的考试,试题的安全性至关重要. 所以,题库安全性是在计算机化自适应测验中很重要的一个评价指标. 对被试而言,降低信息量丰富项目的出现次数,即控制高质量项目的过度曝光是题库安全性的重要一环. 现有的 a 分层方法尽管可以较好地控制项目曝光率,但是题库需要事先分层,分层数目和每层选题数目均不确定,并且当题库中的项目发生更改时需要重新划分题库等问题. 通过以上 4 组实验表明,新选题策略具有如下优点: (i) 只需提前计算好每个项目的区分度相对秩,而不需要事先对题库进行分层,就可以降低系统误差. (ii) 适用于不同的 CAT 考试场景,根据不同的测验需求选择的控制参数,如对测量精度要求更高,则控制参数取值小一点; 若对题库的安全性要求更高,则将控制参数的值调高一点. (iii) 该方法简单灵活,适用于多种不同的 IRT 测量模型,可直接应用于多级评分模型中,可操作性强 新方法虽然

和按 a 分层策略一样,都是升 a 的选题策略,但是新方法与按 a 分层选题策略不同的是: 按 a 分层策略是块与块之间的项目区分度上升,不能保证题与题之间的区分度参数上升,新方法按照区分度大小的秩次选题,保证题与题之间区分度上升. 新方法的 χ^2 值只是比随机化选题方法的更大,这可以较好地佐证本文的说法. 另外,在整个蒙特卡罗模拟过程中,假设被试能力服从标准正态分布,这是因为若假设被试服从均匀分布,则在端点处也会存在大量被试,这并不符合实际情况.

把新选题策略应用于多维 CAT(multidimensional CAT,MCAT) 模型[13] 中还需要做进一步的讨论. 特别是迁移到具有认知诊断功能的计算机化自适应测验(CD-CAT) ,因为认知诊断模型的区分度参数或者作用类似于 IRT 的区分度参数是什么,还需要仔细斟酌.

8　参考文献

[1] 漆书青,戴海崎,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.

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

[3] Lord F M. Application of item response theory to practical testing problems [M]. Hillsdale NJ: Erlbaum Associates,1980: 392-449.

[4] Chang Huahua,Ying Zhiliang. a-stratified multistage computerized adaptive testing [J]. Applied Psychological Measurement,1999,23(3):211-222.

[5] 左孝凌,李为鑑,刘永才. 离散数学 [M]. 上海: 上海科学技术文献出版社,1982: 128-130.

[6] 李萍,甘登文,丁树良,等. 自动控制区分度作用的选题策略研究 [J]. 江西师范大学学报: 自然科学版,2013,37(1):101-105.

[7] Chen Jyun-Hong,Chao Hsiu-Yi,Chen Shuying. A dynamic stratification method for improving trait estimation in computerized adaptive testing under item exposure control [J]. Applied Psychological Measurement,2020,44(3): 182-196.

[8] Busemeyer J R,Diederich A. Cognitive modeling [M]. CA: Sage Pubns,2009.

[9] 李佳,丁树良. 多种分层方法在 CAT 校准误差中的应用研究 [J]. 江西师范大学学报: 自然科学版,2016,39(1):69-72.

[10] 李佳,丁树良,方剑英. 基于平均数形式的选题策略比较 [J]. 江西师范大学学报: 自然科学版,2015,39(1): 17-20.

[11] Cheng Ying,Jeffrey M Patton,Can Shao. a-stratified computerized adaptive testing in the presence of calibration [J]. Educational and Psychological Measurement,2015,75(2):260-283.

[12] 李佳,丁树良. 计算机化自适应测验中能力估计新方法 [J]. 江西师范大学学报: 自然科学版,2019,43(2): 111-115.

[13] 毛秀珍,王娅婷,杨睿. 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 [J]. 心理学探新,2019,47(1):47-56.

The Item Selection Strategy on Composing the Discrimination with the Test Process in CAT

LI Jia[1],DING Shuliang[1*],KUANG Tianhao[2]

(1. School of Computer Information Engineering,Jiangxi Normal University,Nanchang Jiangxi 330022,China;
2. College of Information Engineering,Jiangxi University of Technology,Nanchang Jiangxi 330098,China)

Abstract: The item selection strategy is an important content in computerized adaptive testing(CAT) . A new item selection strategy about composing the percentile rank of the discrimination parameter for an item in the bank with the particular examinee's test process is introduced in this paper. The control parameter can be changed to meet the different test situations. New method has better performance through the Monte Carlo simulation method on improving the test precision,controlling item exposure and the utilization of the item bank.

Key words: CAT; item selection strategy; item exposure control; item bank utilization; control parameter

(责任编辑: 冉小晓)