文章编号:1000-5862(2013)01-0101-05

自动控制区分度作用的选题策略研究

李 萍,甘登文*,丁树良
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要:沿用引入曝光控制因子的 CAT 选题策略中所提出的曝光控制因子,同时将定长测验的测验长度以及不定长测验中被试累积信息量加入到新选题策略中,提出了在不分层的条件下自动控制区分度作用的选题策略. 蒙特卡洛模拟结果表明:新方法在项目调用均匀性、测验效率等评价指标上效果均较为理想.

关键词:项目曝光控制;选题策略;计算机化自适应测验
中图分类号:TP 391 **文献标志码**:A

0 引言

目前,针对 CAT 中项目过度曝光问题提出的选题策略至少有 5 种:随机化选题策略、条件选题策略、综合选题策略、多阶段自适应测验设计及分层选题策略[1-3]. Chang Huahua 等[4-5] 于 1999 年提出的按 a 分层法(a-STR)既能够使得题库中项目曝光相对均匀,又能使被试能力估计有较高的准确性,同时测验效率也较高. 随后,Chang Huahua 等[6] 又提出按 b 分块 a 分层法,这种方法是为了让题库在分层的条件下各个层中的难度参数分布仍然可以基本上与整个题库中项目难度分布保持一致. 2011 年,程小扬等[7-9] 提出将曝光控制因子引入到分层的 CAT 选题策略中,并将它与层数相关的区分度幂函数相结合,有效地兼顾了项目调用均匀性和测验效率. 本文加入了定长测验的测验长度以及不定长测验中被试累积信息量 2 个指标,与曝光控制因子相结合提出了在不分层的条件下自动控制区分度作用的新选题策略,对题库中项目曝光率进行有效控制.

1 预备知识

Chang Huahua 等提出的按 a 分层法实质上是想在能力估计尚不精确的初始阶段,选择区分度较低的项目施测,而在能力精估阶段则选择区分度较高的优质项目施测,也即实现选题过程中逐步"升 a". 程小扬为了达到在宏观和微观层面均实现"升 a",同时又能有效控制项目曝光率,提出了将曝光控制因子引入到 CAT 的选题策略中. 对于题库中的第 j 个项目来讲,其曝光控制因子记为 $ecf(j)$,此处 $ecf(j) = m_j/\bar{m}$. 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,$\bar{m}$ 表示题库中所有项目被前 $(i-1)$ 个考生使用的平均次数,即 $\bar{m} = \sum_{j=1}^{M} m_j/M$. M 为题库中的项目总个数. $ecf(j)$ 的曝光因子参数记为 $\lambda(j)$,$\lambda(j)$ 主要用来控制 $ecf(j)$ 在施测过程中对项目选择的影响力度(本文中 $\lambda(j)$ 取值均为 1). 另外,程小扬还引入了第 j 个项目的区分度参数 a_j 的幂函数 $a(j,T,k) = a_j^{2(T-k)/(k-1)}$,此处 T 为 CAT 施测时分层的总数,k 为被试当前所在层数. 在使用程小扬等提出的选题策略选题时,对当前被试而言,是从该被试尚未进行作答的项目集合,也即从剩余题库中选择使得下式中的 f_j 值最大的项目

$$f_j = I_j(\hat{\theta}) / [\lambda_j \cdot ecf(j) \cdot a(j,T,k)], \quad (1)$$

其中 $I_j(\hat{\theta})$ 为第 j 个项目对当前被试在能力估计值 $\hat{\theta}$ 处的 Fisher 信息量.

2 自动控制区分度作用的新选题策略

考虑到程小扬等提出的选题策略需要分层进行,本文提出了在不分层时自动控制区分度作用的新选题策略. 新方法中的控制曝光因子沿用程小扬等提出的 $ecf(j)$. 新方法中对第 i 个被试而言,对第 j 个项目的区分度参数进行调节的幂函数 $a(j,i)$ 分为 2 种情况:(ⅰ)对于以被试作答项目个数达到一

收稿日期:2012-11-16
基金项目:国家自然科学基金(30860084,31160203,31100756)资助项目.
作者简介:甘登文(1955-),男,江西奉新人,教授,主要从事智能教学软件和应用统计的研究.

定值为测验结束条件的定长测验而言,$a(j,i) = a_j^{2^{*(test\_length-L(i))/test\_length}}$,其中 a_j 为第 j 个项目的区分度,$L(i)$ 为第 i 个被试当前已作答的项目个数,$test\_length$ 为定长测验终止时被试需要作答的总项目数.(ⅱ)对于以测验信息量达到某一固定值为测验终止条件的不定长测验而言,$a(j,i) = a_j^{2^{*(Infor-infor(i))/Infor}}$,其中 $infor(i)$ 代表第 i 个被试当前已经作答了的所有项目的累积信息量,$Infor$ 代表测验终止时被试所需要达到的信息总量. 新的选题策略如下:在定长和不定长 2 种测验形式时,都始终从当前被试的剩余题库中选择使得下式最大的项目

$$f_j = I_j(\hat{\theta}) / [ecf(j) \cdot a(j,i)]. \quad (2)$$

由于在二值评分的 3PLM 中信息函数值正比于区分度参数的平方,因此在采用 Lord 提出的最大 Fisher 信息量(MIC)选题策略施测过程中会首先选择那些区分度较大的项目. 然而,在能力估计尚不准确的测验早期,如果优先选择区分度大的项目给被试作答会造成优质项目的浪费,同时也加大了这些项目的曝光率. 而较晚参加测验的被试作答这些区分度较高的项目的几率则会减少,而只能被迫选择区分度相对较低的项目,导致这些被试测验长度增加. 按照 Chang Huahua 等提出的 CAT 施测时应遵循的 rough-low、accurate-high 的策略,即能力估计较粗糙时用低区分度项目,而在能力估计较精确时,使用高区分度项目新选题策略. 借鉴程小扬的方法,在公式中引入新的区分度幂函数 $a(j,i)$ 来逐步自动调节区分度对信息函数值的影响. 在测验的早期阶段,减少区分度对信息函数值的影响. 在测验的后期阶段,慢慢增大区分度对信息函数值的影响. 所以,对已施测项目的信息量除以 $a(j,i)$,不管测验为定长测验还是不定长测验时,仔细考察 $a(j,i)$ 的计算公式,可发现 $a(j,i)$ 的指数部分的值都是从 2 渐渐逐步变化到 0,如此便可以在能力估计不精确的初始选题阶段优先调用那些区分度影响相对比较小的项目,而在能力精确估计阶段能充分选用区分度较高的优质项目来提高被试能力估计准确性和缩短被试测验长度.

3 模拟实验

3.1 被试及其题库的模拟

本文中所有试验均采用 Monte Carlo 模拟. 在实验过程中模拟生成 1 000 个被试,且所有被试的能力真值都服从标准正态分布.

在实验过程中模拟生成 4 种题库,每种题库均生成 1 000 个项目. 下文用 $N(\mu,\sigma^2)$ 表示均值为 μ,方差为 σ^2 的正态分布,$U[a,b]$ 表示在区间 $[a,b]$ 上的均匀分布,$\beta(a,b)$ 表示参数为 a,b 的 Beta 分布. 题库 1:$\ln a \sim N(0,1)$,$b \sim N(0,1)$;题库 2:$a \sim U[0.2,2.5]$,$b \sim N(0,1)$;题库 3:$\ln a \sim N(0,1)$,$b \sim U[-3,3]$;题库 4:$a \sim U(0.2,2.5)$,$b \sim U[-3,3]$. 以上所有题库中的猜测度参数 c 均出自以下分布:$c \sim \beta(5,17)$.

3.2 模拟 CAT 的施测过程

本文 CAT 施测时参与比较的 4 种选题策略为:(1)MIC 选题策略;(2)随机选题策略;(3)引入曝光控制因子的选题策略(本文记为程选题策略);(4)新的选题策略. 这里 MIC 选题策略及随机选题策略均分别作为测验效率和项目使用均匀性的参照指标.

3.2.1 能力初估阶段 CAT 的模拟 本测验为 3PLM 模型下的 0-1 评分测验. 在测验的能力初始估计阶段,随机选 3 个题目给被试作答,被试的能力估计值为被试作答正确与错误题目个数之比的自然对数. 若被试全部作答正确,则初估其能力值为 3;若被试全部作答错误,则初估被试的能力值为 -3;被试在初估阶段所作答的项目不计入定长测验时被试的测验长度,在初始阶段所作答的项目信息量不计入不定长测验时被试的累积信息量.

3.2.2 能力精估阶段CAT的模拟本文模拟 定长和不定长 2 种测验:(1)定长测验. 设定测验长度为 30. 由于在程小扬的方法中区分度 a_j 的幂函数中用到层数,因此在本定长测验中,使用程小扬的方法时,分为 5 层,每层 6 个题目. 在采用其他选题策略时不分层,测验以被试作答项目个数达到 30 时即结束.(2)不定长测验. 在采用程小扬的选题策略时,设分层数为 4. 其他选题策略时不分层,所有选题策略的测验在被试累积信息量达到 25 时即结束.

3.3 评价指标

本文中采用以下 7 个评价指标:能力估计准确性(ABS)、能力估计标准差(SD)、测验效率、项目调用均匀性、卡方(χ^2)统计量、测试重叠率、人均用题数. 具体计算方法参考文献[10]. 除测验效率越高越好之外,其他 6 个指标均是越小越好.

4 实验结果及分析

实验为定长测验时,结果见表 1 ~ 表 4;实验为不定长测验时,结果见表 5 ~ 表 8.

当测验为定长时,从表 1 ~ 表 4 中结果可以看

出,在第 1、2、3 种题库下,新方法比程小扬的方法在能力估计准确性、能力估计标准差、测验效率、测试重叠率、项目调用均匀性方面表现要好,而且在标识项目曝光率的卡方统计量上的表现只比随机选题策略差,比 MIC 选题有了较大改进,同时也比程小扬的方法表现更佳. 对第 4 种题库新方法在项目调用均匀性、卡方统计量、测试重叠率方面较程小扬的方法稍差一些,但是在能力估计准确性、能力估计标准差方面比程小扬的方法要好. 新方法在 4 种题库下,按照综合指标统一量纲[11] 来看,总体效果比程小扬的方法要好. 总之,定长测验中若要兼顾被试能力估计准确性及项目曝光率,那么新方法可认为是一种可供选择的选题策略.

表 1　定长测验 $\ln a \sim N(0,1)$,$b \sim N(0,1)$,$C \sim \beta(5,17)$ 4 种选题策略的表现

策略	*ABS*	*SD*	测验效率	项目调用均匀性	卡方统计量	测试重叠率	统一量纲
MIC 选题	0.117 6	0.078 7	1.714 5	78.259 3	204.151 0	0.233 4	4.202 5
随机选题	0.285 5	0.184 7	0.271 7	5.412 5	0.976 9	0.030 0	4.996 5
程小扬选题	0.180 1	0.124 6	0.752 9	15.748 9	8.267 6	0.037 3	3.990 3
新选题	0.158 2	0.108 2	0.919 5	15.056 9	7.557 0	0.036 6	4.315 7

表 2　定长测验 $a \sim U$ [0.2,2.5] ,$b \sim N(0,1)$,$C \sim \beta(5,17)$ 4 种选题策略的表现

策略	*ABS*	*SD*	测验效率	项目调用均匀性	卡方统计量	测试重叠率	统一量纲
MIC 选题	0.111 6	0.076 5	2.094 2	72.188 9	173.709 5	0.202 9	4.230 0
随机选题	0.236 3	0.158 9	0.379 9	5.497 5	1.007 4	0.030 0	5.135 3
程小扬选题	0.151 4	0.103 6	1.031 4	11.598 4	4.484 3	0.033 5	4.562 8
新选题	0.135 9	0.093 9	1.249 2	11.271 7	4.235 0	0.033 3	4.860 9

表 3　定长测验 $\ln a \sim N(0,1)$,$b \sim U$ [-3,3] ,$C \sim \beta(5,17)$ 4 种选题策略的表现

策略	*ABS*	*SD*	测验效率	项目调用均匀性	卡方统计量	测试重叠率	统一量纲
MIC 选题	0.129 1	0.087 3	1.440 9	86.955 3	252.040 8	0.281 3	4.172 0
随机选题	0.331 5	0.219 1	0.171 4	5.358 3	0.957 2	0.030 0	4.907 1
程小扬选题	0.191 9	0.126 2	0.633 8	22.807 6	17.340 0	0.046 4	3.741 2
新选题	0.174 7	0.118 4	0.752 1	21.422 1	15.296 9	0.044 3	3.987 3

表 4　定长测验 $a \sim U$ [0.2,2.5] ,$b \sim U$ [-3,3] ,$C \sim \beta(5,17)$ 4 种选题策略的表现

策略	*ABS*	*SD*	测验效率	项目调用均匀性	卡方统计量	测试重叠率	统一量纲
MIC 选题	0.112 8	0.075 5	1.956 1	79.527 1	210.818 9	0.240 1	4.196 8
随机选题	0.289 8	0.186 8	0.222 8	5.356 6	0.957 2	0.030 0	4.907 3
程小扬选题	0.163 9	0.110 6	0.942 7	23.862 8	18.981 2	0.048 0	3.752 1
新选题	0.148 8	0.102 5	1.102 2	24.256 8	19.613 5	0.048 7	3.943 9

当测验为不定长时,从表 5 ~ 表 8 中结果可以看出,对第 1 种和第 3 种题库新方法在项目调用均匀性、卡方统计量方面较程小扬的方法稍差些,但是在测验长度上比程小扬的方法有较大的改进,与随机选题相比有更大的改进; 在第 2 种和第 4 种题库下在项目调用均匀性、卡方统计量、测试重叠率等方面都优于程小扬的方法,同时测验长度也有缩短. 新方法在 4 种题库上,综合指标值都比程小扬的方法表现要好.

表 5　不定长测验 $\ln a \sim N(0,1)$,$b \sim N(0,1)$,$C \sim \beta(5,17)$ 4 种选题策略的表现

策略	*ABS*	*SD*	测验效率	项目调用均匀性	人均用题数	卡方统计量	测试重叠率	统一量纲
MIC 选题	0.195 8	0.136 5	1.391 3	48.357 7	12.359 0	249.868 0	0.195 7	4.214 2
随机选题	0.298 0	0.197 0	0.228 0	4.999 9	29.999 5	0.925 9	0.024 3	4.789 6
程小扬选题	0.212 4	0.147 3	0.482 4	6.616 4	29.312 5	1.663 8	0.024 2	4.793 6
新选题	0.210 8	0.148 4	0.626 1	7.154 6	25.455 0	2.279 6	0.021 0	4.889 1

表 6 不定长测验 $a \sim U$ [0.2,2.5],$b \sim N(0,1)$,$C \sim \beta(5,17)$ 4 种选题策略的表现

策略	*ABS*	*SD*	测验效率	项目调用均匀性	人均用题数	卡方统计量	测试重叠率	统一量纲
MIC 选题	0.206 7	0.144 4	1.577 1	41.887 1	11.090 0	216.877 5	0.163 5	4.185 9
随机选题	0.258 0	0.166 8	0.328 8	5.120 8	29.710 0	0.982 9	0.024 0	4.683 4
程小扬选题	0.205 2	0.144 6	0.760 1	5.620 7	21.211 0	1.735 9	0.016 3	5.194 7
新选题	0.207 3	0.144 3	0.932 2	4.539 5	18.544 5	1.325 7	0.013 3	5.920 2

表 7 不定长测验 $\ln a \sim N(0,1)$,$b \sim U$ [-3,3],$C \sim \beta(5,17)$ 4 种选题策略的表现

策略	*ABS*	*SD*	测验效率	项目调用均匀性	人均用题数	卡方统计量	测试重叠率	统一量纲
MIC 选题	0.197 8	0.137 6	1.310 5	52.952 9	13.008 0	280.177 4	0.222 7	4.211 6
随机选题	0.346 4	0.224 8	0.153 4	5.229 2	30.000 0	1.012 8	0.024 3	4.733 5
程小扬选题	0.233 3	0.153 0	0.426 0	13.202 4	29.047 5	6.691 8	0.028 5	3.921 2
新选题	0.246 6	0.163 1	0.498 0	12.805 3	27.076 0	6.811 5	0.026 6	3.977 7

表 8 不定长测验 $a \sim U$ [0.2,2.5],$b \sim U$ [-3,3],$C \sim \beta(5,17)$ 4 种选题策略的表现

策略	*ABS*	*SD*	测验效率	项目调用均匀性	人均用题数	卡方统计量	测试重叠率	统一量纲
MIC 选题	0.203 6	0.140 5	1.484 8	47.830 4	11.668 0	263.930 9	0.202 0	4.215 3
随机选题	0.303 2	0.209 5	0.200 7	5.097 1	30.000 0	0.962 3	0.024 3	4.820 1
程小扬选题	0.210 7	0.145 5	0.609 6	14.149 0	25.390 0	8.942 0	0.026 8	4.132 0
新选题	0.201 7	0.140 4	0.761 7	12.973 4	22.192 5	8.770 0	0.023 3	4.541 4

5 讨论

本文在 3PLM 的 0-1 评分模型下,在沿用曝光控制因子的条件下,提出了自动控制区分度作用的新选题,实验结果表明新方法在一定程度上有优越性. 然而,在不定长测验时新方法在缩短被试测验长度及卡方统计量方面与题库仍有着一些关联性. 从实验结果也可看出,如果人为控制项目曝光率势必在一定程度上减少了信息量大的项目的调用,反过来这样又会增加测验长度. 因此如何在这两者之间达到较好的平衡,即如何更好地在缩短被试测验长度和项目曝光率尽可能均匀方面达到均衡是一个值得深入探讨的问题.

虽然从综合指标上看,随机选题策略表现最好,但其实质上有违自适应的原则,因此任何一个自适应的测验都不会使用随机选题策略,但是随机选题策略对应最好的曝光均匀性,所以在模拟研究中常常用以作为曝光控制的对照. 使用综合指标评价选题策略的优劣时,随机选题策略竟然占优势,这说明用综合指标来评价选题策略的优劣是否合理,或者我们使用的权重是否合理(比如 7 个指标可以分成 3 大类: *ABS*,*SD* 用来刻画能力估计精度,而项目调用均匀性、卡方统计量、测试重叠率用来刻画测验安全性,测验效率、人均用题数用来刻画测验效率,是否可以每一类分别给出不一定相同的权重,而同一类别的指标再平分这一类别的权重)? 这都有待商榷. 同样 MIC 由于其对测验安全性的负面影响而受到广泛批评,因此它的综合指标上的得分的高低也不能作为对新方法取舍的标准; 而 MIC 的测验效率和人均用题数则可以作为新方法相应指标上的比较标准. 另外,新方法只是引入到 0-1 评分模型中,如果引入到多级评分模型中,效果如何还有待研究. 特别和最近引入的一些多级评分 CAT 选项策略[12-13]进行比较,是一个有趣的问题.

6 参考文献

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

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

[3] Wang Chun,Chang Huahua. Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing [J]. Journal of Educational Measurement,2011,48(3):255-273.

[4] Chang Huahua,Ying zhiling. A global information ap-

proach to computerized adaptive testing [J]. Applied psychological Measurement,1996,20(2):213-219.

[5] Chang Huahua,Ying Zhiliang. A-stratified multistage computer-ized adaptive testing [J]. Applied Psychological Measurement,1999,23(3):211-222.

[6] Chang Huahua,Jiahe Qian,Ying Zhiliang. A-stratification multistage computerized adaptive testing with b blocking [J]. Applied Psychological Measurement,2001,25,(4):333-341.

[7] 程小扬,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.

[8] 程小扬和,丁树良. 拓广分部评分模型下计算机自适应测验变加权选题策略 [J]. 心理科学,34(4):965-969.

[9] 程小扬,丁树良,严深海,等. 引入曝光因子的计算机化自适应测验选题策略 [J]. 心理学报,2011,43(2):203-212.

[10] 汤楠,丁树良,余丹. 结合优先级指标和曝光因子的多级评分 [J]. 江西师范大学学报: 自然科学版,2011,35(6):646-650.

[11] 戴海琦,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):778-783.

[12] 罗芬,丁树良,王晓庆. 多级评分计算机自适应测验动态综合选题策略 [J]. 心理学报,2012,44(3):400-412.

[13] 程小杨,丁树良,朱隆尹,等. 等级评分模型下的最大信息量分层选题策略 [J]. 江西师范大学学报: 自然科学版,2012,36(5):446-451.

The Study of Item Selection Strategy-Automatic Control for the Role of Discrimination Index

LI Ping,GAN Deng-wen*,DING Shu-liang

(Institute of Computer Information and Engineering,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: Combining the exposure control factor proposed by Cheng Xiao-yang with the ratio of the test length administered for fixed tests or the ratio of the test information administered for variable length tests to construct a new item selection strategy,and the new strategy can control the role of discrimination index automatically without stratification of the item pool during the course of computerized adaptive testing. The results of Monte Carlo simulations show that the new approach is more ideal in the performance of item exposure control and test efficiency and other indexes.

Key words: item exposure-control factor; item selection strategy; computerized adaptive testing

(责任编辑: 冉小晓)