文章编号:1000-5862(2018)04-0374-05

基于 GRM 模型的 CAT 分层方法在校准误差中的应用研究

李 佳,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:在计算机化自适应测验(CAT)中,0-1 评分模型下 b 组块 a 分层的方法(BASTR)可以提高测量准确性的同时平衡项目的曝光率,但在多级评分模型中项目难度/步骤参数有多个,无法直接使用该方法;又因为信息函数可以较好地综合被试能力和项目参数,但最大信息量选题策略的测验安全性太低. 因此,将多级评分模型中的多个参数综合成一个指标作为 b 分块的依据,模仿 BASTR 方法,提出 5 种新的 B 分块 a 分层方法,并且采用"影子题库"下最大信息量的选题方法. 在等级反应模型(GRM)下蒙特卡洛实验结果表明,新方法在测验精度、题库利用率和机会红利等评价指标中总体表现良好,B_max - min 分块方法表现最优.

关键词:计算机化自适应测验;GRM 模型;B 分块 a 分层方法;机会红利;影子题库
中图分类号:B 841.7 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2018.04.09

0 问题的提出

被人们誉为"测验领域的新天地"[1]的计算机化自适应测验(computerized adaptive testing,CAT)具有测验精度高、长度短、成本低、实时反馈考试成绩等优点,被广泛应用于美国医生护士资格考试、美国研究生入学考试和中国汉语水平考试当中[1]. 根据 CAT 采用的测量模型,可分为基于项目反应理论(item response theory,IRT)的单维 CAT(unidimensional CAT,UCAT),基于多维项目反应理论(multidimensional item response theory,MIRT)的多维 CAT(multidimensional CAT,MCAT)以及基于认知诊断理论的认知诊断 CAT(cognitive diagnostic CAT,CD-CAT). 在 UCAT 中根据评分模型的不同又可分为 0-1 的 2 级评分 CAT 和多级评分 CAT. 虽然 0-1 评分在国外大受推崇,但是它处理不了诸如计算题、论述题、作文题等多值评分的项目. 为了实际测量的需要,为了符合我国考试现状,提高考试质量,研究多级评分模型下的 CAT 是很有必要的.

选题策略决定了被试作答的测验项目,关系到测验结果的准确性、测验的安全性和测验的可信度,是 CAT 的重要环节之一. 常见的多级评分选题策略主要有 2 种:(i)能力与项目难度/步骤参数的某个综合指标相适应的选题策略[2-4],这类方法为了与单维的被试能力值相匹配,将参数向量进行降维,但这样做可能会浪费一些有效信息;(ii)将选题过程的多个重要部分如信息函数、项目参数以及被试当前能力值综合在一起,得到一个综合指标进行选题[5-7].

由于项目难度和区分度通常存在正相关[8],Chang Huahua 等[9]在 1999 年提出的 a 分层法的基础上于 2001 年提出基于 0-1 评分的 b 组块 a 分层的选题策略(BASTR)[10],它可以较好地控制项目曝光率,提高题库的安全性,以及抵消测验初期能力估计值的不确定性. 在等级反应模型(graded response model,GRM)下,大量蒙特卡洛实验表明项目难度向量降维后与区分度也存在较高的相关性(具体请参见下文模拟实验部分),模仿 BASTR 方法将难度参数进行降维后作为 B 分块的依据,本文共提出了 5 种 B 分块方法:(i)按照难度参数取平均值分块,(ii)按照难度参数取中位数分块,(iii)按照难度参

收稿日期:2017-08-16
基金项目:国家自然科学基金(31500909,31360237,31300876),教育部人文社会科学研究青年基金(BYJC880060)和江西省教育厅科学技术 2017 年一般项目(GJJ170212)资助项目.
作者简介:李 佳(1979-),女,江西南昌人,副教授,主要从事计算机辅助教学和心理测量方面的研究. E-mail:1276676143@qq.com

数去掉最大值和最小值后取平均值分块,(iv)随机取一个难度参数分块,(v)按照难度参数的最大值和最小值取平均值后分块. 又因为在多级评分中,信息函数可以较好地综合项目参数和能力参数,是平衡能力测量准确性和题库使用安全性的重要途径,并且"影子题库"可以明显提高项目调用的均匀性[2],因此采用"影子题库"下的最大信息量选题方法,这样既保证了测验的准确性又兼顾了项目曝光的均匀性.

在 CAT 施测过程中,项目选择、被试得分的计算、能力的估计以及测验的终止,归根到底都依赖项目参数. 然而,在现实中,只有估计的项目参数可以提供,机会红利就会发生[11]. 这种现象在 0-1 评分定长 CAT 中已有证明[12-13];又因为自适应项目挑选标准倾向于选择虚假的估计的大区分值的项目,这会带来假的大信息量和假的低能力估计标准误,机会红利对不定长 CAT 测验长度的影响更大[14]. 然而,在国内外还未见文献报道机会红利对多级评分 CAT 的影响. 本文将研究 CAT 在 GRM 模型下,机会红利对定长 CAT 和不定长 CAT 的影响.

0.1　GRM 模型简介

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,每个等级难度要求严格递增,记 $P^*_{\alpha j,t}$ 为被试 α 在第 j 个项目得 t 分或 t 分以上的概率,则 $P^*_{\alpha j,t} = 1/(1+\exp(-Da_j(\theta_\alpha - b_{jt})))$,记 $P_{\alpha j,t}$ 为被试 α 在第 j 个项目恰得 t 分的概率,则 $P_{\alpha j,t} = P^*_{\alpha j,t} - P^*_{\alpha j,t+1}$,而 Fisher 信息量公式为 $I_j(\theta_\alpha) = \sum_{t=0}^{f_j} D^2 a_j^2 P_{\alpha j,t}(1 - P^*_{\alpha j,t} - P^*_{\alpha j,t+1})^2$,其中 a_j 为题库中第 j 个项目的区分度,b_{jt} 为第 j 个项目等级 t 的难度,第 j 个项目共有 f_j+1 个等级,D 取值 1.7.

0.2　项目参数

在 GRM 中,项目参数的真实值用 $\gamma = (a, \vec{b})$ 表示,用来生成被试对项目的作答反应;用 MMLE/EM 算法[15] 估计得到项目参数的估计值用 $\hat{\gamma} = (\hat{a}, \hat{\vec{b}})$ 表示,参与题库的组块和分层,项目的选择以及被试能力的估计.

0.3　影子题库下最大信息量选题方法

在剩余题库即未作答题库中计算各个项目在被试当前能力估计值上的信息量,从中选出 5 个信息量最大的项目,然后在这 5 个项目中随机选用一题作为被试的下一题.

0.4　*B* 分块的新方法

1) 难度参数取平均值作为分块依据:$B_j(ave) = (b_{j1} + b_{j2} + \cdots + b_{jf_j})/f_j$;

2) 难度参数取中位数作为分块依据:

$$B_j(mid) = \begin{cases} \text{取}\{b_{j1}, b_{j2}, \cdots, b_{jf_j}\}\text{ 的中间值,当 } f_j \text{ 为奇数,} \\ \text{取}\{b_{j1}, b_{j2}, \cdots, b_{jf_j}\}\text{ 中间 2 个值的算术平} \\ \text{均值,当 } f_j \text{ 为偶数;} \end{cases}$$

3) 难度参数去掉最大值和最小值后取平均值作为分块依据:$B_j(ave\_\max - \min) = (b_{j2} + \cdots + b_{jf_j-1})/(f_j - 2)$;

4) 随机取一个难度参数作为分块依据:$B_j(rand) = random\{b_{j1}, b_{j2}, \cdots, b_{jf_j}\}$;

5) 难度参数的最大值和最小值取平均值后作为分块依据:$B_j(ave\_\max + \min) = (b_{j1} + b_{jf_j})/2$.

具体的 B 分块 a 分层方法如下:先让题库按降维后的难度参数 B 排序,相类似的 B 参数形成为一个 B 块,在每个块中按区分度 a 排序后,再按 a 参数进行分层. 这种方法使题库分为 K 层,每层是按升 a 的在每层中按照影子题库下最大信息量方法进行选题.

0.5　参与比较的 8 种选题策略 / 方法

影子题库下最大信息量选题(MFI),作为各方法的准确率和测验长度比较基础;随机化选题(RS),作为各方法的曝光率方面比较基础;a 分层下影子题库下最大信息量选题(A_MFI):题库仅按升 a 排序,每层中按照影子题库下最大信息量方法选题;难度参数向量取平均值分块方法(B_ave);难度参数向量取中位数分块方法(B_mid);难度参数向量去掉最大值和最小值后取平均值分块方法(B_max - min);难度参数随机取值分块方法(B_rand);难度参数向量的最大值和最小值取平均值分块方法(B_max + min).

1　模拟实验

1.1　题库模拟及题库项目参数估计结果

本文在 GRM 模型下设计 2 种题库[16],题库结构如下:题库 1,模拟产生 520 个项目,项目的每个难度参数都服从标准正态分布;题库 2,模拟产生 520 个项目,项目的每个难度参数都服从 [-3,3] 之间的均匀分布. 以上 2 种题库的区分度都服从对数正态分布,取值范围为 [0.2,2.5],每个项目的难度等级数都为 5. 区分度参数与难度向量参数降维后指标的相关性见表 1(其中相关系数计算见文献 [8]).

表 1 题库的区分度参数和难度参数降维后指标的相关性

难度向量参数	取平均值	取中位数	去最大、最小值后取均值	随机取值	最大值和最小值取平均值
题库 1	0.763	0.694	0.796	0.773	0.766
题库 2	0.735	0.690	0.838	0.712	0.756

采用 MMLE/EM 算法估计题库中的项目参数,共估计 50 次取平均值,每次估计都重新生成 2 500 名真值服从正态分布的被试进行估计. 项目估计准确性见表 2(其中 ABS 和 RMSE 计算公式请见文献[15]).

表 2 题库的项目参数估计的准确性

项目数据	题库 1 区分度 a	题库 1 难度 b	题库 2 区分度 a	题库 2 难度 b
ABS	0.120 5	0.182 4	0.136 2	0.194 2
RMSE	0.206 9	0.263 3	0.183 7	0.284 3

1.2 模拟 CAT 的施测过程

实验过程中模拟产生 1 000 个被试,被试能力真值均服从标准正态分布. 本测验为 GRM 模型下的多级评分测验,设被试的能力初值为 0,采用 EAP 方法对能力进行估计. 定长和不定长 2 种测验:定长测验,定长测验测验长度为 32,分层测验中题库分成 4 层,每层选 8 题;不定长测验,所有选题策略的测验在被试累积信息量达到 16 时结束,分层测验中每层信息量按 1∶1∶1∶1 分配,即在每层信息量达到 4 时退出.

1.3 评价指标

1) 能力估计准确性(ABS)

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

2) 卡方检验统计量(χ^2)

$$\chi^2 = \sum_{j=1}^{M} \left(\left(A_j - \left(\sum_{j=1}^{M} A_j / M \right) \right)^2 / \left(\sum_{j=1}^{M} A_j / M \right) \right);$$

3) 测验效率[13] (test efficiency,TE)

$$TE = \sum_{i=1}^{N} I(\hat{\theta}_i, \hat{\gamma}_i) / N;$$

4) 相对测验效率(relative test efficiency,RTE)

$$RTE = \sum_{i=1}^{N} I(\hat{\theta}_i, \hat{\gamma}_i) / \sum_{i=1}^{N} I(\hat{\theta}_i, \gamma_i);$$

5) 不定长测验的测验平均长度(AL)

$$AL = \left(\sum_{i=1}^{N} test\_length(i) \right) / N;$$

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 为题库中项目数,A_j 为第 j 题的曝光率,即 A_j = 第 j 题的使用次数 /N,γ_i 为被试 i 所考项目的参数真实值,$\hat{\gamma}_i$ 为被试 i 所考项目的参数估计值. I_i 为第 i 个被试的测验 Fisher 信息量,$test\_length(i)$ 为被试 i 的测验长度.

能力估计准确性(ABS)表明了测验的准确性,它越接近零,表示越接近无偏,能力估计越准确;卡方统计量反映了题库项目的曝光率,值越小说明项目调用越均匀,CAT 的安全性越好;测验效率和相对测验效率体现了机会红利的影响程度,测验效率是在项目参数估计值下求被试平均测验信息量,这往往是个虚高的值,而相对测验效率值越接近 1,表明机会红利影响越小;因为在不定长 CAT 中每种方法的 TE 值均为 16,所以在不定长测验中用评价指标 AL 代替 TE,且更直观. 本文采用测验精度,题库利用率和机会红利影响这 3 种评价指标对以上 8 种策略进行综合评价比较.

1.4 实验结果及其分析

实验为定长测验,结果见表 3;实验为不定长,结果见表 4.

表 3 定长测验 8 种选题策略的表现

方法	ABS		χ^2		TE		RTE	
	题库 1	题库 2	题库 1	题库 2	题库 1	题库 2	题库 1	题库 2
MFI	0.081 4	0.104 8	67.730	65.590	18.95	19.85	1.189	1.191
RS	0.298 5	0.299 3	7.765	7.738	12.08	11.56	1.081	1.074
A_MFI	0.103 3	0.102 7	45.630	39.580	17.57	17.46	1.151	1.157
B_ave	0.113 9	0.114 5	14.960	14.410	16.83	16.49	1.133	1.123
B_mid	0.115 0	0.116 1	16.880	16.930	16.19	16.95	1.128	1.125
B_max − min	0.106 2	0.107 2	16.090	14.940	16.88	16.75	1.122	1.126
B_rand	0.116 8	0.117 0	13.640	13.190	15.91	15.17	1.120	1.122
B_max + min	0.113 3	0.113 5	14.050	13.810	15.76	15.99	1.124	1.129

在定长测验中,5 种新方法的测验精度略低于最大信息量选题策略,但明显高于随机化选题策略,

和 a 分层下最大信息量选题策略测验精度相当;项目曝光控制不仅要降低过度曝光项目的使用率而且应提高曝光过低项目的使用率,即提高项目曝光的均匀性. 因此,在题库利用率方面的影响,新方法比最大信息量选题和 a 分层最大信息量选题的卡方值都要小很多,表明新方法的题库利用率更高;在大多数文献资料中,测验效率的计算公式[16]为 $\left(\sum_{i=1}^{N} I(\hat{\theta}_i,\gamma_i)\right)/(NL)$,其中 L 为测验长度,N 为被试总人数,该测验效率值越大表明项目为被试提供的信息量越多. 注意到项目参数的真实值未知,所以这里的测验信息量应该是 $I(\hat{\theta}_i,\hat{\gamma}_i)$,而不是 $I(\hat{\theta}_i,\gamma_i)$,由相对效率 TRE 和 TE 一起可以计算出被试的真实测验信息量. 以 MFI 方法为例,在题库 1 下,平均测验估计信息量即测验效率 TE 值为 18.95,而相对测验效率 RTE 为 1.189,那么平均测验真实信息量为 18.95/1.189 = 15.93,并没有想象中的那么高,也证明了机会红利在 GRM 模型下 CAT 中确实会发生,并且 $\left(\sum_{i=1}^{N} I(\hat{\theta}_i,\gamma_i)\right)/(NL)$ 表示的高测验效率也是虚假偏高的. 这 5 种方法指标整体都差不多,但 B_ave和 B_max - min 略好一些的原因是参数构成更稳定,B_max - min 因去掉两头的极端数据,表现更好一些,而 B_mid、B_rand 和 B_max + min 方法中 B 取值更随机一些,受本身项目参数影响较大,所以评价指标有好有差. 综合而言,题库 1 和题库 2 下结果基本相当.

表 4　不定长测验 8 种方法的表现

方法	ABS		χ^2		RTE		AL	
	题库 1	题库 2	题库 1	题库 2	题库 1	题库 2	题库 1	题库 2
MFI	0.121 1	0.139 1	60.510	58.740	1.183	1.197	10.32	9.812
RS	0.202 5	0.220 9	4.745	4.488	1.029	1.063	29.76	28.690
A_MFI	0.166 8	0.163 0	39.280	34.140	1.153	1.158	14.36	12.280
B_ave	0.174 5	0.185 5	11.320	11.560	1.127	1.125	15.67	16.290
B_mid	0.180 3	0.175 2	11.660	11.870	1.124	1.126	15.28	16.150
B_max - min	0.178 4	0.172 5	11.150	11.170	1.121	1.128	14.15	13.460
B_rand	0.182 5	0.186 2	11.120	11.070	1.123	1.124	19.29	18.040
B_max + min	0.172 3	0.178 2	11.620	11.380	1.126	1.127	14.85	15.140

定长与不定长的比较;当测验为不定长时,从表 4 的结果可以看出,实验结果和定长测验类似,但测验精度更差一些,原因是在选题时使用虚假的高 a 值项目,使测验过早结束;而且测验平均长度都短于定长测验的测验长度,卡方值在同等条件下小于定长实验下的卡方值,这也说明了机会红利对不定长 CAT 影响更大. 具体以 FMI 方法为例,在题库 1 下要求测验信息量达到 16 结束测验,但 RTE 值为 1.183,所以真实的测验信息量为 16/1.183≈13.52,所以平均测验长度仅为 10.32 就结束了测验.

2　讨论

因为测验信息量和项目区分度正相关,项目选择方法会过分依赖虚假的高 a 值,在项目估计值下的测验信息量会比在项目真实值下的测验信息量更大,所得高测验效率其实是虚假偏高的. 因此,本文通过分层组块的方法,降低机会红利的影响,尽可能还原测验本质. 尽管如此,在定长实验中可以看到 MFI 方法提供的平均测验信息量仍然是最多的,最大信息量选题策略结合影子题库在各种分层方法下不仅可以降低机会红利影响,还可以提高项目曝光均匀性,所以是一种比较好的 CAT 选题策略.

本文中题库 1 和题库 2 对 CAT 的施测过程影响不大,但文献[16]指出在对项目参数降维时,如果要得到较好的测量效果,就必须适量增加综合较难或者较容易的项目,改善测验精度和降低项目曝光率,本文没有做这方面的考虑,且是固定分成 4 层,不定长测验信息量平均分配,以及项目参数难度等级数为 5,这些都有可能影响实验结果. 但是,本文提出的按难度参数的 5 种情况下的 B 分块 a 分层方法简单明了,很容易直接应用到 GPCM 模型当中. 若在选题方法中加入曝光因子[17],是否可以进一步提高题库利用率,降低机会红利的影响,这是一个非常有趣的问题,值得进一步探讨.

3　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,2002:154-155.

[2] 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.

[3] 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.

[4] 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.

[5] 程小扬,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,2011,34(4):965-969.

[6] 罗芬,丁树良,王晓庆. 多极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(3):400-412.

[7] 王晓庆,罗芬,丁树良,等. 多极评分计算机化自适应动态调和平均选题策略 [J]. 心理学探新,2016,36(3):270-275.

[8] Lord F M,Wingersky M S. An investigation of methods for reducing samplingerror in certain IRT procedures [J]. Applied Psychological Measurement,1983,8(2):347-364.

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

[10] Chang Huahua,Jia heqian,Ying Zhiliang. *a*-stratified multistage computerized adaptivetesting with *b* blocking [J]. Applied Psychological Measurement,2001,25(4):333-341.

[11] 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.

[12] van der Linden W J,Glas C A W. Capitalization on item calibration error in adaptive testing [J]. Applied Measurement in Education,2000,13(1):35-53.

[13] 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.

[14] Jeffrey M Patton,Cheng Ying,Yuan Kehai,et al. The influence of item calibration error on variable-length computerized adaptive testing [J]. Applied Psychological Measurement,2013,75(1):1-17.

[15] 陈青,丁树良,朱隆尹,等. 3 参数等级反应模型及其参数估计 [J]. 江西师范大学学报:自然科学版,2010,34(2):117-122.

[16] 程小扬,丁树良,巫华芳,等. 多级评分模型下的题库结构对 CAT 的影响分析 [J]. 心理学探新,2014,34(5):452-456.

[17] 程小扬,丁树良,严深海. 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.

The Several Stratified Methods of CAT in the Presence of Calibration Error on GRM

LI Jia,DING Shuliang

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

Abstract: For dichotomous scoring,the *a*-stratified method with *b* blocking(BASTR) is an effective and safe method for computerized adaptive testing(CAT). But it could not be applied to the polytomous scoring CAT,because there are too many parameters in the polytomous item response model. It is well known that the Fisher information function is a good comprehension of all item parameters as well as the ability parameter,but the maximum Fisher information (MFI) method derogates the security of CAT. Five new stratified methods are proposed in this paper. The new methods are comprehension of all information of the item parameters for polytomous items and play the role of BASTR. Because "shadow pool" can improve the uniformity of item bank, so the item select strategy is MFI under "shadow pool". The results of Monte Carlo study of graded response model (GRM) show that the new methods has better effect, and B_max - min method is the best one.

Key words: CAT;GRM;a-stratified method with B blocking;capitalization on chance;shadow pool

(责任编辑:冉小晓)