文章编号:1000-5862(2016)01-0056-05

多种分层方法在 CAT 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:若研究或使用 CAT 时将项目参数的估计值认为是真实值,则会产生所谓的机会红利(capitalization on chance),国内尚未报道该方面的研究.在定长和不定长 CAT 测验中考察引入曝光因子的多种分层化选题策略和随机选题策略,综合比较它们在测验精度、题库利用率和机会红利等评价指标中的表现,发现引入曝光因子的 2 种动态分层方法有更好的表现.

关键词:分层方法;曝光因子;题库利用率;机会红利
中图分类号:B 841 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2016.01.10

0 问题的提出

计算机化自适应测验(CAT)的最大特点就是有效性,相比自适应测验和线性测验,它能用更少的题目得到更精确的能力估计值.当前,CAT 已被广泛地应用于美国研究生入学考试、美国医生护士资格考试和美国军队职业倾向成套测验[1].选题策略是 CAT 的核心内容,其中应用最广的是 1980 年 F. M. Lord 提出的最大信息量选题策略(MFI)[2],然而 MFI 方法有许多不足,如为了选择"最优"试题,高区分度题目被频繁选取,而其它题目则较少用到或几乎用不到,造成题库中题目的曝光不均匀从而危及题库的安全性和造成题库的极大浪费.为了改善这种情况,Chang Huahua 等[4-5]提出了 a-分层法和 a-分层 b-分块的选题策略(AST),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,但该方法题库的利用率并不高,有 40% 以上题目是属于过低曝光的(曝光率小于0.02)[6-7];R. B. Juan[8] 认为 AST 方法没有考虑到3PLM 中的 c 参数,应将 c 参数融入到题库分层和选题策略当中,提出了最大信息量分层方法(MIS),用项目 j 的最大信息量 $I_{\max}(j)$ 和该项目取得最大信息量时对应的能力点 $\theta_{\max}(j)$ 分别替代 AST 方法中的 a 参数和 b 参数进行分层,每层中的选题方法是考虑能力估计值和难度的最小距离.但上述 2 种分层方法都需要事先分好层,比较浪费时间,因此罗芬等[9]在等级反应模型下提出了关于难度的动态选题方法(DB),以及李萍等[10]提出了区分度自动分区的选题策略(DA),这 2 种方法均为自动分层方法,不需要改造题库,操作起来更为简洁.

然而,在实施 CAT 的过程中项目参数的真实值事实上是未知的,只能用参数的估计值,这就有一个更深层次的问题需要讨论.本文将用参数估计值参与 CAT 选题产生的影响称为机会红利(capitalization on chance)[6,8],即使用能力估计值的前提下,先定义相对测验效率(relative test efficiency,RTE),它是利用项目参数估计值计算的所有被试的测验信息量之和,除以利用项目参数真值计算的所有被试的测验信息量之和.注意到 Fisher 信息量约等于能力估计的方差的倒数以及与不同被试反应相互独立,因此机会红利可以理解为真实的项目参数对应的"真实"方差(记为 D_1)除以估计项目参数对应的"估计"的方差(记为 D_2),这个比值越大,说明"估计"的方差越小.一方面,"真实"方差不应大于"估计"方差;另一方面,方差越小表明估计越准确,于是这个比值越大,说明"估计"的方差获得不应得到的"好评"越明显,通俗地说,机会红利就是由于项目参数估计的误差额外获得的好处.因此,比值 D_1/D_2 越接近 1 越好;如果比值远大于 1,就产生了大量的机会红利.注意到,R. Bowles 等[11]指出校准误差(calibration error)来自于随机测量误差,如项目的相关影响和项目参数的漂移,由于是随机误差,所以项目参数估计是渐近无偏的,如果是来自系统误差,项

收稿日期:2015-11-17
基金项目:国家自然科学基金(31500909,31360237,31300876,31160203,31100756,30860084,11401271),教育部人文社会科学研究青年基金(13YJC880060)和江西省教育科学 2013 年度一般课题(13YB032)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.

目参数估计会随样本的变化而变化,就不是渐近无偏的了. 由于 Fisher 信息量和项目参数中的区分度参数的平方成正比,所以选择区分度大的选题策略相应的机会红利一般会更大一些,而随机选题策略的机会红利应该比较小. var der Linden 等[12] 发现 MFI 方法有较低的标准误差,原因是 MFI 的选题具有高区分度参数,所以具有较小的估计误差,他们将这种现象称为校准误差的随机性或者是机会红利,在定长的 CAT 模拟中采用估计的项目参数已经证明了这种现象. J. M. Patlon 等[13] 也证明了在不定长 CAT 中机会红利的发生,在已选的项目中测验信息量是根据参数估计值计算的,会导致能力估计值虚假的低标准误,从而使测验过早结束. 2014 年,Cheng Ying [6] 讨论过 AST 方法在测验初期用低 a 项目,项目选择根据项目难度和当前能力估计值相匹配,所以机会红利的影响得到缓和,对校准误差而言比 MFI 方法更稳健(robust) . 因为在 2000 年 van der Linden 等[12] 已经研究过单纯的 MFI 方法,2015 年,Cheng Ying 等[6] 研究过单纯的 AST 方法在校准误差中的表现,它们有一个共同的不足就是题库利用率不高,而程小扬[14] 提出的引入曝光因子的选题策略可以较好地提高题库利用率,平衡题目的被选用次数,所以本文将上述 4 种方法常见的分层方法以及最大信息量选题策略分别和曝光因子相结合,得到含曝光因子的 a 分层 b 分块选题策略(E_AST) 、含曝光因子的最大信息量分层策略(E_MIS) 、含曝光因子的难度自动分区选题策略(E_DB) 、含曝光因子的区分度自动分区选题策略(E_DA) 、含曝光因子的最大信息量选题策略(E_MFI) ,再加上随机选题策略(作为比较的基准) 一起在定长测验和不定长测验下分别进行模拟比较. 本研究共设计了 3 类评价指标: (i) 测验精度,(ii) 题库利用率,(iii) 机会红利,对以上 6 种选题策略进行综合评价比较.

0.1　曝光因子

称 $ecf(j) = m_j/\bar{m}$ 为题库中的第 j 个项目的曝光因子[14] ,当第 i 个考生参加考试时,m_j 表示前面 $i-1$ 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 $\bar{m} = \sum_{j=1}^{M} m_j/M$,$M$ 为题库中的题数.

0.2　项目参数

在 3PLM 下,项目参数的真实值用 $\gamma = (a,b,c)$ 表示,用来生成项目反应; 用 MMLE/EM 算法估计得到项目参数的估计值用 $\hat{\gamma} = (\hat{a},\hat{b},\hat{c})$ 表示,参与项目选择和能力估计.

0.3　β 参与比较的 6 种选题策略

以下 R_α 表示被试 α 的剩余题库,即从题库中扣除被试 α 已测项目后的项目集合.

(i) 随机化选题策略(RS) : $rand\{\hat{I}_j\}$;

(ii) 含曝光因子的最大信息量选题策略(E_MFI) : $\max_{j\in R_a}(\hat{I}_j/ecf(j))$;

(iii) 含曝光因子的 a 分层 b 分块选题策略(E_AST) : 该方法是先让题库按难度参数 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,而每层中难度 b 分布和整个题库的分布相似,每层中选择满足 $\min_{j\in R_a}(|\hat{b}_j-\hat{\theta}|/ecf(j))$ 的项目;

(iv) 含曝光因子的最大信息量分层策略(E_MIS) : 该方法是先让题库按 $\theta_{\max}(j)$ 排序,相类似的 $\theta_{\max}(j)$ 形成为一个 $\theta_{\max}(j)$ 块,在每个块中按 $I_{\max}(j)$ 排序后,再按 $I_{\max}(j)$ 进行分层,这种方法使题库分为 K 层,每层中选择满足 $\min_{j\in R_a}(|\hat{b}_j-\hat{\theta}|/ecf(j))$ 的项目,其中

$$\theta_{\max}(j) = \hat{b}_j + \frac{\ln(1+(1+8\hat{c}_j)^{1/2}) - \ln 2}{1.7\hat{a}_j},$$

$$I_{\max}(j) = \frac{1.7^2\hat{a}_j^2}{8(1-\hat{c}_j)^2}(1-20\hat{c}_j-8\hat{c}_j^2+(1+8\hat{c}_j)^{3/2});$$

(v) 含曝光因子的关于难度的动态分区选题方法(E_DB) : 选取测验信息量平方根的倒数(记为 ε) 作为度量测量的精度,即 $\varepsilon = (\sum_{j=1}^{m_\alpha}\hat{I}_j)^{-1/2}$,选题时考虑 $\underset{j}{rand}\{|\hat{b}_j-\theta|/ecf(j) < \varepsilon\}$;即从该题目集合中随机选取;

(vi) 含曝光因子的区分度自动分区选题方法(E_DA) : 在定长实验中,自动区分度因子 $a(j,i) = a^{2(l_{\text{test}}-L(i))/l_{\text{test}}}$,在不定长测验中,自动区分度因子 $a(j,i) = a_j^{2(Infor-\inf(i))/Infor}$,其中 l_{test} 为测验长度,$L(i)$ 为第 i 个被试当前已作答的项目个数,$inf(i)$ 表示第 i 个被试当前的测验信息量,$Infor$ 为预先确定的总测验信息量,选题时考虑 $\max_{j\in R_a}(\hat{I}_j/(ecf(j)\,a(j,i)))$.

1　模拟实验

1.1　被试及其题库的模拟

为了方便比较,下文所有试验模拟条件同参考文

献[6]:(i)在实验过程中模拟生成题库共520个项目且满足条件 $\ln a \sim N(0,1)$,$b \sim N(0,1)$,$c \sim Beta(5,17)$,且 $0.2 < a < 2.5$,$-3.5 < b < 3.5$,$\text{abs}(a-b) < 4$,$c < 0.4$. 题库的项目数据见表1(真实值).(ii)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,每次估计均重新生成2 500名真值服从正态分布的被试进行估计[15]. 项目估计准确性见表2.

表1 题库的项目数据

项目数据	区分度 a	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	0.807 610

表2 题库的项目参数估计的准确性

项目数据	区分度 a	难度 b	猜测度 c
ABS	0.128 23	0.174 46	0.149 94
RMSE	0.214 14	0.243 13	0.239 93

猜测度的估计误差较大是很难避免的[1],所以基于3PLM的CAT更容易受到机会红利的影响.

1.2 模拟CAT的施测过程

实验过程中模拟生成1 000个被试,且被试能力真值服从标准正态分布. 本测验为3PLM模型下的0-1评分测验. 设被试的能力初值为0,然后根据不同的选题策略采用EAP方法对能力进行估计. 分定长和不定长2种测验:(i)定长测验中设定测验长度为40,分层测验中题库分成4层,每层选10题;(ii)不定长测验中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达到4时退出.

1.3 评价指标

(i)测验精度指标为能力估计准确性(ABS):$ABS = \sum_{i=1}^{N} |\theta_i - \hat{\theta}_i| / N$,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. 该指标反映了被试能力真值与其能力估计值的平均偏差,指标值越小说明能力估计越准确.

(ii)题库利用率指标为卡方检验统计量(χ^2)、项目从未曝光率(UE)和项目过低曝光率(NE). 卡方检验统计量(χ^2):

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

其中 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. 该指标用来衡量题库中项目曝光的均匀性,指标值越小说明题库中项目曝光越均匀,测验安全性越高.

(iii)项目从未曝光率(UE):$UE = \sum_{i=1}^{M} UE_i / M$,其中 UE_i 表示题库中曝光率等于0的项目数;项目过低曝光率(NE)[5-6]:$NE = \sum_{i=1}^{M} NE_i / M$,其中 NE_i 为题库中曝光率小于0.02的项目数,这项指标能更清楚地体现题库的利用率. 显然,UE 和 NE 的值越小越好.

(iv)机会红利指标为相对测验效率(relative test efficiejncy,RTE)[6],

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

其中 N 为被试总人数,$\hat{\theta}_i$ 为被试 i 的最终能力估计值,$\hat{\gamma}_i$ 是对被试 i 项目的估计参数值,γ_i 是对被试 i 施测项目的参数真实值. 分子、分母均为Fisher信息量的和. RTE 的统计意义如前所述,可以指示是否有机会红利发生,它越接近1越好,越大于1越不好(出现更大的机会红利).

(v)不定长测验的测验平均长度(AL):由于不定长测验中每个被试的测量精度类似,所以早达到测验精度的被试所需测验长度更短,而晚达到测验精度的被试所需测验长度就更长,这项指标体现了测验效率.

1.4 实验结果及其分析

实验为定长测验时,结果见表3;实验为不定长时,结果见表4.

当测验为定长时,从表3可以看出,和预期的一致,由项目估计参数计算的测验信息量比用真实参数计算的信息量高,这表明机会红利确有发生,RS方法是随机选题,机会红利最小,而E_MFI方法得到了更大的随机误差,也就是机会红利值更大. 特别地,用估计参数算得的最大信息量项目有大的区分度估计就会导致正的校准误差,这些结果有一个重要的暗示,就是用E_MFI方法能力估计的标准误更

低是假的,各分层方法的机会红利更小,这是因为它们没有过于看重高区分度值的项目,成功地降低了机会红利的影响. 加入曝光因子之后的各种方法的题库利用率都有显著提高,特别是含曝光因子的自动分层方法 E_DB 和 E_DA,不仅机会红利比较小,而且题库利用率比较高. 各分层方法对校准误差更为稳健. 而 E_MFI 方法更易受机会红利的影响,采用了高的区分度估计值,得到虚假的高测验信息量,而分层方法更倾向于发现测验信息量接近真实测验信息量的项目,对机会红利而言更为稳健,以前的研究表明[4-5,7-9],用真实参数值时,各分层方法能力恢复情况和 MFI 方法差不多,在用估计参数时,各分层方法表现都比较好,只有轻微的失利,但是有更好的题库利用率.

表 3　定长测验($L=40$) 6 种选题策略的表现

分层方法	*ABS*	χ^2	*UE*	*NE*	*RTE*
RS	0.225 2	4.881 4	0.000 0	0.000 0	1.139 3
E_MFI	0.049 8	47.021 0	0.407 2	0.402 4	2.657 2
E_AST	0.108 6	43.489 0	0.251 5	0.200 0	1.254 5
E_MIS	0.114 7	46.672 0	0.273 4	0.134 6	1.236 0
E_DB	0.104 8	16.353 0	0.211 5	0.115 4	1.234 9
E_DA	0.129 9	14.287 0	0.192 3	0.000 0	1.234 8

表 4　不定长测验 6 种分层方法的表现

分层方法	*ABS*	χ^2	*UE*	*NE*	*RTE*	*AL*
RS	0.342 10	5.017 6	0.000 0	0.000 0	1.166 1	35.531
E_MFI	0.137 40	46.732 0	0.454 5	0.438 5	2.657 5	21.246
E_AST	0.252 41	21.758 3	0.287 5	0.253 4	1.254 2	29.846
E_MIS	0.269 12	20.226 5	0.265 6	0.243 3	1.239 6	28.438
E_DB	0.228 70	14.524 0	0.128 8	0.115 4	1.239 3	28.455
E_DA	0.221 90	13.893 3	0.097 6	0.000 0	1.236 6	27.541

当测验为不定长时,从表 4 可以看出,实验结果和定长测验类似,但测验精度更差一些,测验平均长度都短于定长测验的测验长度,这也说明了不定长测验更有利于提高测验效率,所以在 CAT 中采用不定长测验是有道理的,同时还可以发现,含曝光因子的自动分层方法题库利用率比较高,题目使用均匀性更好,能力估计精度比较高.

2　讨论

含曝光因子控制的分层选题策略中,最大信息量选题策略表面上估计精度比较高,但是机会红利比较大,这说明能力估计精度高是一种“假象”,这是由于项目参数估计的误差引起的. 在本文讨论的其他几种含曝光因子控制的分层化选题策略中,固定分层选题策略中的分层退出规则的制定(本文是测验信息量的平均分配,其实也可以按 1∶4∶9∶16 分配测验信息量到各层中[16],或者其它的分配方法) 也会影响到机会红利,综合比较还是自动分层选题策略表现最优,机会红利比较小,题库使用率比较高,题目使用均匀性比较好,只是能力估计精度表面上比 E_MFI 差一些.

对于不定长 CAT,由于预先指定的信息量为 16,对应的平均测验长度均小于定长 CAT 的 40 题,所以不定长 CAT 的能力估计精度要稍微差一些. 如果考察表 3 和表 4 中随机选题策略,定长和不定长的测验长度之比约等于 40/35 = 1.142 8,而相对应的能力估计精度(*ABS*) 之比为 0.225 2/0.342 1 = 0.658 3. 后面增加的 5 题对于提高能力估计的精度的贡献远远大于测验开始时的 5 题,这是因为在执行 CAT 的过程中,能力的估计是一个不断精确化的过程.

3PLM 因含有猜测参数 c,其估计准确性不高,所以更易受机会红利的影响,但可操作的 CAT 非计量学指标中的内容平衡约束,可以降低机会红利发生的风险,因为每次选题时只考虑题库的子集,这可以做实验进一步地验证; 另外,多级评分模型机会红利的研究未见报道,原则上 0-1 评分的机会红利计算公式中项目难度参数修改为难度向量(或者阈值参数向量) 就可以推广为多级评分的机会红利计算

公式,这就可以进行更加深入地探讨.

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.

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

[3] Chang Huahua, Ying Zhiliang. To weight or not to weight? balancing influence of initial items in adaptive testing [J]. Psychometrika, 2008, 73(3): 441-450.

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

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

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

[7] 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): 69-72.

[8] Ramón B J, Paloma M, Julio O. Maximum information stratification method for controlling item exposure in computerized adaptive testing [J]. Psicothema, 2006, 18(1): 156-159.

[9] 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(3): 400-412.

[10] 李萍,甘登文,丁树良. 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.

[11] Bowles R, Wise S, Kingbury G. A report on position effects in the NCLEX RN Examination [R/OL]. [2015-11-19]. www. ncsbn. org/Positoion_Effects_NCLEX_RN. pdf.

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

[13] Patton J M, Cheng Ying, Yuan Kehai, et al. The influence of item calibration error on variable-length computerized adaptive testing [J]. Applied Psychological Measurement 2013, 37(1): 13-22.

[14] 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.

[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] 胡姗. 基于 GPCM 和 3PLM 的 CAT 研究 [D]. 南昌: 江西师范大学,2015.

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

LI Jia, DING Shuliang

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

Abstract: Previous studies of CAT have treated item parameter estimates as if they are the true population parameter values, but capitalization on chance may occur. In this article, it is examined that the performance of several stratified methods under more realistic conditions where item parameter estimates instead of true parameter values are used in the CAT. To improve the bank utilization and reduce the captalization on chance, Monte Carlo simulations manipulated some key factors based on 3-parameter Logistic model: fixed or variable test, stratified item pool methods, use or not use of the exposure control factor. The results showed that the two dynamical stratified methods in conjunction with exposure control factor are better.

Key words: stratified item select method; exposure control factor; bank utilization; capitalization on chance

(责任编辑: 冉小晓)