文章编号:1000-5862(2018)01-0062-05

基于 GRM 的在线校准研究

熊建华,罗　慧,王晓庆,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:计算机化自适应测验的题库面临建设成本高且更新、扩充技术较复杂等问题. 在线校准技术,可以将新题和旧题置于同一参数量尺上,降低了题库扩充成本. 已有若干关于两级评分下新题的在线校准研究,但多级记分项目的在线校准却鲜见报道. 该文先由拓展的夹逼平均法求取难度初值,并用多序列相关系数法求取区分度初值,再采用多步 EM 算法估计项目参数. Monte Carlo 模拟结果表明:新题项目参数的估计值返真性较好,且参数的估计精度随着作答次数的小幅度增加而保持着逐渐提高的趋势.

关键词:在线校准;夹逼平均法;多步 EM 算法;多序列相关系数法
中图分类号:B 841.7　　文献标志码:A　　DOI:10.16357/j.cnki.issn1000-5862.2018.01.11

0　引言

近年来,随着计算机技术的发展和教育测量理论的日渐成熟,计算机自适应测验(Computerized Adaptive Testing,CAT)凭借测验时间灵活、题型多样化、测验长度相对短等优势,从而在许多大型评价项目中得到广泛的应用. CAT 在使用一段时间后,题库中某些试题可能由于过度曝光、存在缺陷或过时等情况而不再适用[1],此时需要对题库进行淘汰旧题和补录新题的操作. 新题需要估计出相应的项目参数(下文称为新题校准),且保证新题和题库中的题目(下文称为旧题)的参数在同一量尺上,再经过筛选后,才能入库并正式投入使用.

H. Wainer 等[1]阐述了 CAT 中 2 种校准新题的策略:(i)传统校准策略,先进行锚测验设计,在新、旧题之间设置一些锚题,再通过实测数据进行等值转换以保证新题和旧题的项目参数在同一量尺上. 这种校准策略涉及等值设计问题,这可能带来大量的人力、物力的耗费,而且可能导致新题信息的曝光. (ii)在线校准策略,在被试对旧题进行自适应作答过程的同时,将新题指派给其作答,并通过收集的被试作答反应进行新题的校准,测验过程中被试既作答旧题又作答新题,被试实际上充当着锚人的作用[2],省去了进行等值的繁琐步骤,将旧题和新题置于同一量尺上. 近年来,国内学者也对在线校准技术进行了深入研究[3-7],他们的研究成果表明在线校准是一种有效的方式,避免了传统的锚测验设计离线校准的缺点.

H. Wainer 等[1]认为在对新题进行在线校准时,通常有以下 2 种设计方式可动态地在实施 CAT 过程中将新题植入:(i)随机设计,即当每位被试到达预设的植入新题位置时,从新题集合中随机挑选 1 个,植入被试 CAT 测验中以收集其作答反应,并保证该被试不重复答该题. (ii)自适应设计,利用自适应的特点,依据新题的初估值选择与之适应的被试来作答,以达到减少参数估计所需样本量的目的. 因自适应设计的方法与新题的初估值息息相关,并且初估所需的样本量难以确定,而随机设计法实施更为简便,且在某些条件下随机设计法返真性较好[8],故本文的新题校准采用随机设计植入 CAT 方式.

随着计算机和其他电子设备的普及使得点击流以及被试与计算机交互的持续时间等反应数据可直接利用,使更多的新题型纳入现今和未来的任务中[9]. 随着非选择题(如简答分析题)日益受重视,多级记分项目也变得越来越重要. 目前,在线校准研究基本上集中在 0-1 评分模型,很少考察多级评分模型,特别是等级反应模型[9](GRM)的在线校准未见研究报道. 本文开发基于 GRM 模型对新题的项目参数进行在线校准的方法,以项目参数的返真结果为主要评价标准,用 Monte Carlo 模拟探究这种在线校准方法的表现.

收稿日期:2017-10-12
基金项目:汉考国际科研基金项目"基于 GRM 的项目参数在线校准"(CTI2017B06)的研究成果.
作者简介:熊建华(1977-),女,江西樟树人,副教授,主要从事智能教育软件的研究. E-mail:270281168@qq.com

1　研究方法

新题的参数估计分为 2 个过程:(i) 项目参数的初估阶段;(ii) 项目参数的精估阶段,主要采用经典的多步 EM 算法[10].

1.1　新题的项目参数的初估阶段

1.1.1　**去两端极值夹逼平均法计算新题的难度初值**　在 0-1 评分模式下,根据项目反应理论(IRT)的基本思想[11],考虑到被试答对某项目的概率与其能力值和该项目的难度值的比值有关,比值大于 1 则答对的概率大,否则答对的概率小. 根据单参数逻辑斯蒂克模型(1PLM)可知,该比值等于 1 时,被试答对该项目的概率为 0.5. 因此,对于同一项目来说,在作答次数足量的前提下,通常存在着能力与项目难度值十分相近的被试[7],因此可以利用这些被试的能力来初估新题的难度. 下面称这种获得项目难度初值的方法为"能力值夹逼法",而能力值和项目难度值接近的被试集合为标本组.

在多级评分下采用等级间被试能力值夹逼法计算难度等级的初值. 游晓锋等[7]在 0-1 评分下定义了 c_{num} 和 w_{num},分别表示被试在某题正确作答的人数和错误作答的人数,并建议每个夹逼区间的样本量为 18. 在多级评分下,用 $c(t)$ 表示被试在某题得 t 分的人数. 通过预研究发现,若样本组的样本量仍设为 18,结果不稳定. 因此本文在夹逼平均法的基础上,提出了一种更加宽泛的范围选取方式,称为去两端极值法. 即收集第 j 题得 t 分的被试的能力值 θ,$t=1,2,\cdots,f_j$,将这 f_j 个集合中的能力值按照从小到大的顺序排列存入 $cap(t,i,j)$,再去掉每个集合中排列在前 5% 以及后 5% 的被试的能力值,即 $cap(t,i,j)=\{\theta_i,\theta_i$ 为在第 j 题得 t 分的被试 i 的能力值$\}$,记 $cap(t,i,j)$ 中的人数为 $c(t)$,

$$b_{jt}=\Big(mean\Big(\sum_{i=c(t)\times 0.05}^{c(t)\times 0.95} cap(t,i,j)\Big)+mean\Big(\sum_{i=c(t+1)\times 0.05}^{c(t+1)\times 0.95} cap(t+1,i,j)\Big)\Big)/2. \quad (1)$$

去两端极值法是抹去最高分和最低分再取平均值. 这个 5%,是通过反复实验获得的,它既容易实施又可以保持一定准确性.

1.1.2　**多序列相关系数法计算新题的区分度初值**　新题的区分度参数可利用多序列相关系数法[12]求取,具体过程如下[13]:

(i) 对每个新题 j,利用被试对该题的作答反应计算各等级的通过率 $P_t^*=n/N$,$t=1,2,\cdots,f_j$,其中 N 为总被试人数,n 为在新题 j 上得分大于等于 t 的人数;

(ii) 将 P_t^* 转化成标准正态分数 Z_t,并求出 Z_t 对应的正态密度函数值 $h(Z_t)$;

(iii) 求出新题 j 得分的标准差 σ_j,及新题 j 得分与总分(此处总分指的是被试经过 CAT 后的累计总分及其对该题的作答反应的和)的题总相关系数 r_{xy},得出点多序列相关系数 $rpp_j=r_{xy}\times\sigma_j/\sum_{k=1}^{f_j}h(Z_k)$;

(iv) 再将点多序列相关系数转化成多序列相关系数 $rp_j=rpp_j\times\sigma_j/\sum_{k=1}^{f_j}h(Z_k)$;

(v) 计算得出新题 j 的区分度初值 $a_j=rp_j/\sqrt{1-rp_j^2}$.

1.2　评价指标

采用平均绝对偏差 ABS 来衡量新题的项目参数的精度,其计算公式为

$$ABS(X)=\sum_{r=1}^{R}\Big(\sum_{i=1}^{K}|x_i-x_{ir}|/K\Big)/R, \quad (2)$$

其中 K 为新题数量,R 为重复实验的次数,x_{ir} 为 x_i 的第 r 次试验中新题项目参数的估计值,x_i 为新题项目参数的模拟真值. ABS 指标反映估计值与真值绝对偏差的平均,ABS 值越小,说明估计值与真值越靠近.

2　Monte Carlo 模拟实验

在 CAT 中动态地随机挑选被试作答新题,将自适应测验得到的被试估计能力值视为被试能力值的真值,并基于被试对新题的作答反应,对新题项目参数进行在线校准. 假定被试能力值服从标准正态分布 $\theta\sim N(0,1)$、项目区分度的对数 $\ln a_j\sim N(0,1)$、项目难度参数 $b_{jt}\sim N(0,1)$ 且 $b_{j1}<b_{j2}<\cdots<b_{jf_j}$,即等级难度之间按升序排列. 使用最大信息量选题策略作为 CAT 中的选题策略. 采用贝叶斯后验期望估计方法估计其能力值,终止规则为定长 30 题(不包括新题).

在模拟实验中,取被试 5 000 人,CAT 题库容量 1 000题,新题 100 题. 难度参数采用 2 种初值计算方法:(i) 特定区间,通过反复的预研究取 2 个等级的夹逼样本量为 $pk_1=36$、$pk_2=10$(得 0,2 分的标本组使用夹逼样本量为 36,得 1 分的标本组使用夹逼样本量为 10);(ii) 扩展区间,即去两端极值法剔除前后 5% 的区间;(iii) 2 种方法均重复实验 30 次取

其平均值.

2.1 区分度相同的 GRM 在线校准

本实验讨论新题区分度相同时,对新题的难度参数在线校准精度,并讨论其在不同等级数和不同作答次数时的表现. 基本思路为: 先利用 2 种计算初值的方法得到新题的各个等级难度初值,再采用 MEM 算法进一步估计等级难度参数.

2.1.1 2 个等级的 GRM 实验结果 实验结果如表 1 和表 2 所示.

表 1 2 个等级难度的夹逼平均法的实验结果

新题数	作答次数	$ABS(b_{j1})$	$ABS(b_{j2})$	难度平均偏差
100	100	0.205 1	0.212 9	0.209 0
100	200	0.143 3	0.146 2	0.144 8
100	300	0.122 0	0.120 8	0.121 4

表 2 2 个等级难度的去两端极值法(5%)的实验结果

新题数	作答次数	$ABS(b_{j1})$	$ABS(b_{j2})$	难度平均偏差
100	100	0.210 8	0.207 3	0.209 1
100	200	0.144 2	0.140 7	0.142 5
100	300	0.120 2	0.119 4	0.119 8

表 1 和表 2 的数据表明,*ABS* 均随着新题的被答次数的递增而递减. 值得注意的是,当作答次数达到 200 时,难度参数的估计值与模拟真值的偏差均小于 0.15,说明新题的被答次数越多,参数估计的准确性越高. 比较表 1 和表 2 可以看出,2 种方法的迭代效果较好,由于特定区间夹逼平均法样本量的选取需要耗费更多的时间和精力且选取范围不稳定,因此在后续实验中,求取等级难度的初值方法就不再用此初值计算方法,而使用一种夹逼范围更宽泛的去两端极值 5% 的夹逼平均法代替.

2.1.2 多个等级难度的 GRM 实验结果 表 3 表明,随着难度等级数和新题的被答次数(即校准样本量)的递增,估计值均有向真值靠近的趋势. 随着等级数的递增,每道新题的被作答次数并不是成倍增长,只是小幅度地增长(增加 100 次),但是从表 3 中可以明显看到,当作答次数为 200 时,等级数 $f=4$ 与等级数 $f=3$ 的等级难度参数估计的精度基本持平在 0.15 左右; 当作答次数为 300 时,等级数 $f=5$ 与等级数 $f=4$ 的等级难度参数估计的精度相差不大. 这说明在作答次数足量的条件下,不同等级之间的难度参数估计的精度基本趋于稳定.

表 3 多个等级难度的去两端夹逼平均法的实验结果

等级数	作答次数	100	200	300	400	500	600
$f=3$	$ABS(b_{j1})$	0.207 3	0.152 3	0.121 3	0.105 7	–	–
	$ABS(b_{j2})$	0.207 8	0.145 0	0.113 6	0.099 4	–	–
	$ABS(b_{j3})$	0.208 2	0.149 5	0.120 3	0.104 2	–	–
	难度平均偏差	0.207 8	0.148 9	0.118 4	0.103 1	–	–
$f=4$	$ABS(b_{j1})$	–	0.149 9	0.124 3	0.107 4	0.095 6	–
	$ABS(b_{j2})$	–	0.139 6	0.113 6	0.097 8	0.087 6	–
	$ABS(b_{j3})$	–	0.138 2	0.112 8	0.099 4	0.089 9	–
	$ABS(b_{j4})$	–	0.151 7	0.121 6	0.108 7	0.096 0	–
	难度平均偏差	–	0.144 9	0.118 1	0.103 3	0.092 3	–
$f=5$	$ABS(b_{j1})$	–	–	0.128 9	0.112 2	0.098 7	0.089 6
	$ABS(b_{j2})$	–	–	0.116 5	0.102 0	0.089 6	0.082 8
	$ABS(b_{j3})$	–	–	0.114 2	0.100 2	0.088 7	0.080 6
	$ABS(b_{j4})$	–	–	0.117 9	0.103 1	0.089 3	0.083 4
	$ABS(b_{j5})$	–	–	0.129 3	0.110 6	0.101 3	0.091 1
	难度平均偏差	–	–	0.121 4	0.105 6	0.093 5	0.085 5

2.2 区分度不同的 GRM 在线校准

本实验讨论区分度不同的新题参数校准. 实验基本思路为: 先利用去两端极值夹逼平均法,结合被试在新题上的作答反应得到新题的各个等级难度初值,再基于被试的作答反应及其总分,利用多序列相关系数法得到新题的区分度初值,最后使用 MEM 方法联合估计出新题的区分度参数及各个等级难度.

由表 4 可知,新题参数的估计精度随着作答次数和等级数的增加而提高. 值得注意的是,当作答次数为 400 时,区分度参数的 *ABS* 可以达到 0.2 以下,而 2、3 个等级的难度整体偏差在 0.2 左右,4、5 个等级的难度整体偏差都在 0.2 以下. 从实验结果来看,随着新题的作答次数增加(但是并没有成倍增加),新题的等级难度和区分度的估计精度保持着提高的趋势.

表 4　多个等级难度的去两端夹逼法且区分度不同的实验结果

等级数	作答次数	200	300	400	500	600	700
$f=2$	$ABS(a_j)$	0.239 1	0.217 3	0.194 7	0.190 8	0.192 4	–
	$ABS(b_{j1})$	0.312 7	0.255 7	0.220 4	0.205 7	0.192 4	–
	$ABS(b_{j2})$	0.325 8	0.268 2	0.231 8	0.201 9	204 6	–
	难度平均偏差	0.319 3	0.262 0	0.226 1	0.203 8	202 0	–
$f=3$	$ABS(a_j)$	0.245 5	0.202 9	0.189 3	0.195 6	0.183 0	–
	$ABS(b_{j1})$	0.320 8	0.262 7	0.214 2	0.193 9	0.177 5	–
	$ABS(b_{j2})$	0.311 2	0.219 5	0.190 8	0.167 2	152 5	–
	$ABS(b_{j3})$	0.333 5	0.250 3	0.228 2	0.207 8	192 8	–
	难度平均偏差	0.321 8	0.244 2	0.211 1	0.189 6	174 3	–
$f=4$	$ABS(a_j)$	–	0.185 8	0.174 3	0.174 0	0.171 6	0.166 4
	$ABS(b_{j1})$	–	0.228 8	0.200 0	0.187 5	0.178 0	0.169 8
	$ABS(b_{j2})$	–	0.188 3	0.163 2	0.149 4	0.141 1	0.133 3
	$ABS(b_{j3})$	–	0.196 6	0.160 5	0.153 4	0.143 6	0.133 6
	$ABS(b_{j4})$	–	0.248 2	0.202 6	0.202 0	0.190 0	0.178 2
	难度平均偏差	–	0.215 5	0.181 6	0.173 1	0.163 2	0.153 7
$f=5$	$ABS(a_j)$	–	0.188 1	0.180 2	0.175 1	0.168 1	0.170 1
	$ABS(b_{j1})$	–	0.236 1	0.205 4	0.194 3	0.178 5	0.172 3
	$ABS(b_{j2})$	–	0.176 8	0.157 3	0.147 6	0.133 3	0.128 5
	$ABS(b_{j3})$	–	0.174 8	0.150 8	0.138 6	0.125 3	0.119 5
	$ABS(b_{j4})$	–	0.193 7	0.167 4	0.155 9	0.138 7	0.134 8
	$ABS(b_{j5})$	–	0.241 7	0.217 9	0.209 7	0.185 7	0.181 9
	难度平均偏差	–	0.204 6	0.179 8	0.169 0	0.152 3	0.147 4

3　结论与讨论

本文分 2 种情况探究了基于 GRM 模型多级记分项目的在线校准以及扩展的 MEM 方法的表现. 新题的项目参数的估计精度随着作答次数的小幅度增加而保持着逐渐提高的趋势. 在保证足量的作答次数的情况下,不同等级的难度参数和区分度的估计精度基本趋于稳定. 这足以说明在多级评分下的在线校准保持着两级评分下参数估计的优势.

本文在多级评分下采用去两端极值夹逼法计算等级难度的初值. 由游晓锋等[6]的研究可知,适量的样本量是需要通过大量模拟实验才能得到,它的取值范围也是不稳定的,而夹逼样本量的选取对新题的等级难度的估计有着不可忽视的影响. 要将夹逼平均法求取项目的难度应用于多级评分,夹逼样本量的选取是一个难点. 在区分度相同的 GRM 实验中,有使用夹逼平均法求取 2 个等级的新题难度参数,其中夹逼样本量是通过大量预研究得到. 是否存在更合理的样本量或更为简便的等级难度初值方法是十分值得探究的.

另外,本文植入新题的方式是随机设计,CAT "自适应"的特点并没有体现出来,Zheng Yi[8]在基于 GPCM 的多级记分项目的在线校准中探究了 GPCM 模型下,拓展的算法及程序在随机设计和自适应设计下植入新题的参数估计效果. 在其研究中,新题项目参数的模拟真值的分布不够广泛而固定于特定的几种取值,这可能是使得随机设计的表现优于自适应设计的原因. 因此在后续的研究中加入自适应思想的新题植入方式是必要的.

再者,本文是以将 CAT 的估计能力值看作是具体被试的能力真值来探究多级记分项目在 GRM 模型下的在线校准. 但是这样的做法可能将估计能力值与真实能力值的偏差传递到新题的在线校准过程中. 为减小这种偏差对估计效果的影响,He Yinhong 等[4]提出了一种新方法——MLE_LBCI_Method A. 如何将这种新方法(MLE_LBCI_Method A) 和本文的方法进行有效结合,将有待进一步的研究和探讨.

4　参考文献

[1] Wainer H, Mislevy R J, Wainer H, et al. Item response theory, item calibration, and proficiency estimation [J]. Bioresource Technology, 1990, 98(1): 218-20.

[2] 陈平,张佳慧,辛涛. 在线标定技术在计算机化自适应测验中的应用 [J]. 心理科学进展, 2013, 21(10): 1883-1892.

[3] 陈平. 2 种新的计算机化自适应测验在线标定方法[J]. 心理学报,2016,48(9):1184-1198.
[4] He Yinhong,Chen Ping,Li Yong,et al. A new online calibration method based on Lord's Bias-correction (online first)[J]. Applied Psychological Measurement,2017,41(6):456-471.
[5] 汪文义. 认知诊断评估中项目属性辅助标定方法研究[D]. 南昌: 江西师范大学,2012.
[6] 汪文义,丁树良,游晓锋. 计算机化自适应诊断测验中原始题的属性标定[J]. 心理学报,2011,43(8):964-976.
[7] 游晓锋,丁树良,刘红云. 计算机化自适应测验中原始题项目参数的估计[J]. 心理学报,2010,42(7):813-820.
[8] Zheng Yi. Online calibration of polytomous items under the generalized partial credit model[J]. Applied Psychological Measurement,2016,40(6):434-450.
[9] Samejima F. Estimation of latent ability using a response pattern of graded scores[J]. Psychometrika,1969,34(1):1-97.
[10] Ban J C,Hanson B A,Wang T,et al. A comparative study of on-line pretest item: calibration/scaling methods in computerized adaptive testing[J]. Journal of Educational Measurement,2001,38(3):191-212.
[11] 漆书青. 现代教育与心理测量学原理[M]. 北京: 高等教育出版社,2002.
[12] 张厚粲,徐建平. 现代心理与教育统计学[M]. 北京: 北京师范大学出版社,2004.
[13] 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.

The Online Calibration Based on Graded Response Model

XIONG Jianhua,LUO Hui,WANG Xiaoqing,DING Shuliang
(College of Computer Information Engineering,Jiangxi Normal University,Nanchang Jiangxi 330022,China)

Abstract: Item bank of computerized adaptive test is faced with high construction cost and updated,expanded technology more complex. Not only can the application of the online calibration technology reduce cost,but also put the calibrated parameter values of the new items and the old in the same scale. Online calibration of new items under dichotomously scored models has achieved good results,but under polytomously scored items is reported rarely. To explore the performance of online calibration for polytomously scored items,a method to calculate the initial values of the multiple EM cycle method (MEM) is proposed based on graded response model(GRM),which is focus on the extended squeezing average method and the multiple-sequence correlation coefficient method to calculate as the initial parameters of the new item,then use the multiple EM cycle method to estimate parameters. Results of Monte Carlo simulation show that the parameter estimation of new items can get acceptable estimation accuracy and the parameters of new items are more accurate with a small increase of the sample size for calibration.

Key words: online calibration; squeezing average method; the multiple EM cycle method; multiple-sequence correlation coefficient method

(责任编辑: 冉小晓)