

计算机化自适应测验中原始题项目参数的估计*

游晓锋^{1,2} 丁树良¹ 刘红云³

(¹江西师范大学信息工程学院, 南昌 330027) (²外语教学与研究出版社, 北京 100089)

(³北京师范大学心理学院, 北京 100875)

摘 要 计算机化自适应测验(Computerized Adaptive Testing, 简称 CAT)其安全性面临着新的挑战, 小题库的安全更受威胁。如何建设一个大型、优质的题库成为 CAT 研究中一个非常重要的课题。目前 CAT 题库的建设存在一些问题, 如成本高且保密性较差。尤其是等值技术较复杂且锚题重复使用容易造成泄露。如能在实施 CAT 过程中插入未经过参数估计的项目(原始题), 同时对原始题项目参数进行估计, 这对建设大型、优质的 CAT 题库来说其意义是不言而喻的。本文基于 1PLM 和 2PLM 对此进行研究, 提出了原始题在线估计的新方法以及推导出了求区分度参数 a 迭代初值的计算公式。研究表明: 无论是模拟研究还是实证研究, 原始题被作答的次数对项目参数估计结果都会产生不同的影响, 并且原始题作答人数越多项目参数估计精度也越高。

关键词 原始题; CAT; 夹逼平均法; 多重迭代极大似然估计法; 蒙特卡洛模拟

分类号 B841

1 引言

随着计算机技术的发展和测量理论的成熟, 计算机化自适应测验(Computerized Adaptive Testing, 简称 CAT)在教育领域中的应用越来越广泛, 它不同于传统纸笔测验“千人一卷”, 而是“因人而异”的“量体裁衣”式测验。众所周知, 实施 CAT 需要一个优质、大型的题库。建设一个题库需要做一系列工作, 首先命题专家根据测验目的命制出一批试题, 然后通过实测收集被试对试题的作答反应、给每个被试评分, 再进行项目参数估计, 通过模型资料拟合检验对已估出项目参数的试题进行筛选, 挑选认为合格的试题入库。

上述某些过程还必须满足一定的条件, 比如参加测验的被试总数不能太少, 且分布要相当合理。Baker 和 Kim (2004) 综合前人研究成果, 认为对二参数 Logistic 模型(2PLM)欲估准未知项目参数和被试能力值, 至少要 30 个项目 500 人; 而对三参数 Logistic 模型(3PLM)则至少要 60 个项目 1000 人。

此外, 若要建立一个大型题库, 所有项目不可能对同一批被试进行施测, 而需要对不同批次被试或在不同时间点进行施测, 这时为了使不同测验获得的参数具有可比性, 即放在同一量尺上, 则必须进行等值(Kolen & Brennan, 2004); 而要进行等值, 在测试之前就必须要进行等值设计。

由此可以发现建立一个大型题库是十分复杂的一项工作。有资料报导, 美国一些高风险(High stake)计算机化自适应测验的项目, 从专家命制试题到最后入库, 平均一道试题需要花费大约 1000 美元(张华华, 私人通讯)。我国卫生部考试中心在题库建设中也呈现优良项目的获取成本高且建库周期长的特点。特别是题库还需要不断更新, 即不断地淘汰一些质量不够好的试题, 让一些使用次数较多的试题休眠, 以及加入一些新的试题, 这样更加提高了题库的建设与维护成本。于是对于欲建或已建题库的考试机构都希望能够花较少的时间, 较少的费用创建、维护和扩展基于项目反应理论(IRT)的题库。因为只有基于 IRT 的题库才可能在科学定

收稿日期: 2009-04-10

* 国家自然科学基金(30860084, 60263005), 教育部高校博士点基金 (8020070414001)资助。

通讯作者: 丁树良, E-mail: ding06026@163.com

义上使不同的测验之间能进行比较,才可能为 CAT 服务,而基于经典测验理论基础上的题库却做不到这一点(漆书青,戴海崎,丁树良,2002)。

据我们对题库建设的实际工作调查发现,试题编制后收集实测数据是使建库成本上升的重要原因。因为它包含组织被试实施测验,阅卷评分等诸多环节。被试的测验动机是影响作答反应的重要因素,从而也是影响题库质量的一个重要问题。于是,考虑从收集实测数据这一环节入手,以降低建库成本,提高被试考试动机,这成为当前 CAT 讨论的一个热点(Wainer, 2000; Han, 2007; Chang & Lu, 2009; Makransky, 2009)。

Wainer 等 8 位专家学者(2000)指出在实施 CAT 时,对每个参加 CAT 的被试可以随机指派固定数量的原始题,让被试对原始题进行作答,原始题不带项目参数,并且原始题得分不记入成绩也不参与能力估计。在 CAT 中植入(Seeding)原始题的目的是收集原始题的作答反应数据,然后按边际极大似然估计(MMLE)方法或贝叶斯方法获得项目未知参数(Wainer et al., 2000)。Wainer 等人(2000)称这种估计方法为在线校准(Online calibration),由于估计项目参数时使用了题库中已经校准过的项目信息,故这种在线校准不需要等值(Wainer et al., 2000)。这种方法通过在 CAT 施测时获得被试对原始题作答反应的数据,省略了组织被试对原始题单独施测的时间和费用,也节省了等值的费用,同时作答时被试无法知道哪些是不计入总分的原始题,故保证了其有较高的考试动机,因此被认为是一种值得推广的

好办法。

但是对原始题采用 MMLE 方法进行项目参数估计对被试人数有一定要求,即每个原始题必须接受一定量被试的作答。假设参加测验的人数 n 固定,每个项目若至少需要有 m 个人作答,显然 m 值越大,这场测验可以植入原始题数就越少,因此题库的更新成本就越高。寻找一种对每个原始题作答的被试人数 m 相对较少的前提下,却能保证原始题项目参数的估计精度的估计方法,这就是本文要讨论的问题。本文首先介绍对原始题在线进行项目参数估计的新方法及其原理,然后再介绍如何进行估计,以及用 Monte Carlo 模拟方法考查估计结果的精度,并采用实测数据对该项目参数估计方法进行验证,最后还对此问题作进一步讨论。

2 研究方法

2.1 用夹逼平均法求取项目的难度

若采用 0-1 评分方式,在实施 CAT 时植入原始题,考虑到答对该项目的被试能力值一般高于该项目的难度值,而答错该项目的被试能力值一般低于该项目的难度值,当被试的能力值与项目的难度值大致相等的时候,根据单参数逻辑斯蒂克模型(1PLM)可以知道,被试答对该项目的概率为 50%。由此可以知道只要某一项目作答人数足够多,在答对该项目和答错该项目的被试中总是存在着一些被试的能力值与该项目的难度值非常接近。而且通过实验我们可以得到如图 1 所示的结果,这一结果也支持我们的这一看法。

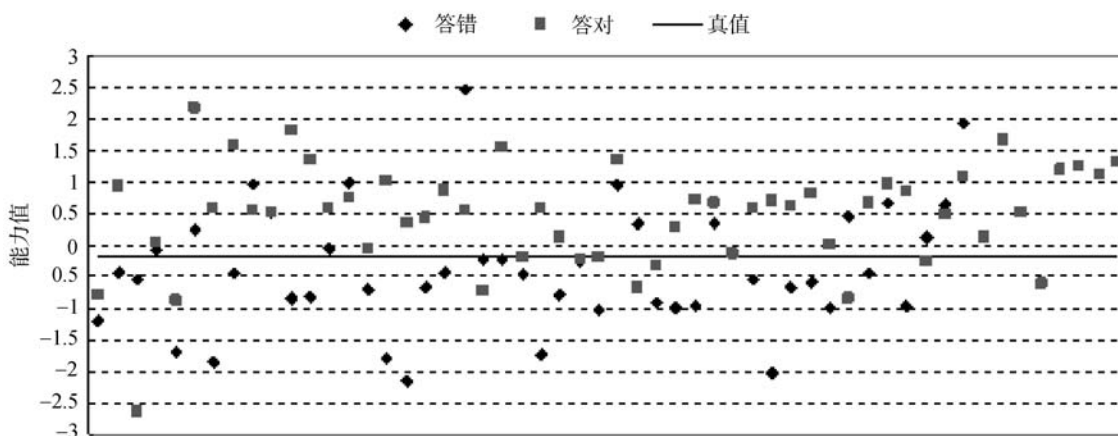


图 1 1PLM 下 100 个被试在难度真值为-0.16661 的原始题作答散点图

由此启发我们采用“夹逼平均法”方式来求取项目的难度值 b_j , 记 $correct(j) / wrong(j)$

为答对/答错第 j 个项目所有被试能力值 θ 集合。即:

$$\begin{cases} \text{correct}(j) = \{\theta_i | \theta_i \text{ 为被试 } i \text{ 的能力值且 } U_{ij} = 1\}, \text{ 记 } \text{correct}(j) \text{ 中人数为 } c_j \\ \text{wrong}(j) = \{\theta_i | \theta_i \text{ 为被试 } i \text{ 的能力值且 } U_{ij} = 0\}, \text{ 记 } \text{wrong}(j) \text{ 中人数为 } w_j \end{cases}$$

将集合 $\text{correct}(j)$ 中被试能力值 θ 从小到大排序得到 $\text{csort}(i, j)$, 即 $\text{csort}(1, j) \leq \text{csort}(2, j) \leq \dots \leq \text{csort}(c_j, j)$, 将集合 $\text{wrong}(j)$ 中被试能力值 θ 从大到小排序得到 $\text{wsort}(i, j)$, 即 $\text{wsort}(1, j) \geq \text{wsort}(2, j) \geq \dots \geq \text{wsort}(w_j, j)$, 令

$$b_j = \left(\sum_{i=\text{cnum}/2}^{\text{cnum}} \text{csort}(i, j) + \sum_{i=\text{wnum}/2}^{\text{wnum}} \text{wsort}(i, j) \right) / 2 \quad (1)$$

其中 cnum 、 wnum 随着项目被作答的次数不同而有所不同, 并且 $\text{cnum}/2$ 和 $\text{wnum}/2$ 都是取整, 首先为 cnum 、 wnum 设定一个初始值, 如果答对第 j 个项目所有被试的人数 c_j 小于设定的初始值, 那么 cnum 值则等于答对第 j 个项目所有被试人数 c_j ; 如果答错第 j 个项目所有被试的人数 w_j 小于设定的初始值, 那么 wnum 则等于答错第 j 个项目所有被试人数 w_j 。通过大量的模拟实验发现 cnum 、 wnum 初值的选取对项目难度 b 值的估计有较大影响, 本文所涉及实验的 cnum 、 wnum 初始值设定为 18, 是我们根据大量模拟实验而得到的一个经验值。

2.2 单参数条件极大似然估计

1PLM 原始题项目参数估计方法是建立在被试能力值 θ 已知的一般假设基础之上, 1PLM 为:

$$P(u_{ij} = 1 | \theta_{ij}) = \frac{1}{1 + \exp[-(\theta_{ij} - b_j)]} \quad (2)$$

其中参数 b_j 为项目的难度参数; 特别值得注意的是: 式(2)中其他变量与传统单参数逻辑斯蒂克模型表示有所不同, 在式(2)中 θ_{ij} 为第 j 个项目第 i 次被作答时对应被试能力值; u_{ij} 表示第 j 个项目第 i 次被作答时的作答反应; $P(u_{ij}=1|\theta_{ij})$ 表示能力为 θ_{ij} 被试答对难度为 b_j 的项目的概率;

由于所有作答原始题的被试都参加了自适应测验, 因此就可以获得每个作答原始题被试的能力值 θ_{ij} , 对项目难度可以采用条件极大似然估计。由局部独立性假设可以推出似然函数为:

$$L_j = \prod_{i=1}^{n_j} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}} \quad (3)$$

其中 n_j 表示项目 j 被作答的次数, $Q_{ij}=1-P_{ij}$ 。与式(3)相应的对数似然函数为:

$$\ln L_j = \sum_{i=1}^{n_j} (u_{ij} \ln P_{ij} + (1-u_{ij}) \ln Q_{ij}) \quad (4)$$

(4)式对项目参数 b_j 求导后, 再通过 Newton-Raphson(N-R)迭代求解(可参见漆书青等, 2002; Baker & Kim, 2004)可以获得该原始题的难度参数

b_j , 其中 b_j 的初值则由(1)式给出。

2.3 双参数条件极大似然估计

2PLM 原始题项目参数估计方法同样也是建立在被试能力值 θ 已知基础之上, 其模型为:

$$P(u_{ij} = 1 | \theta_{ij}) = \frac{1}{1 + \exp[-Da_j(\theta_{ij} - b_j)]} \quad (5)$$

其中参数 a_j 为项目的区分度参数; 其他的变量与(2)式中表示相同。

同样也由自适应测验可以得到每个作答原始题被试的能力值 θ_{ij} , 再通过夹逼平均法获得原始题难度的估计值 b_j , 然后通过采用条件极大似然函数对原始题的区分度进行估计。由此可以得到与(4)式相仿(但 P_{ij} 由(5)式给出)对数似然函数, 进一步再对项目参数 a_j 求导后, 通过 Newton-Raphson 迭代求解可以获得该项目的区分度参数 a_j 。

如果 a_j 的极大似然估计是用迭代求解, 则初值的选取就非常重要(可参见王能超, 1984), 并且初值的选取是否恰当直接影响到 a_j 的估计精度和迭代次数。在本文实验中 a_j 初值选取方法如下:

先将被试在原始题中的通过率 p_j 转化为标准正态分数 z_j , 也就是设随机变量 ε 服从标准正态分布 $N(0,1)$, 要寻找 z_j 使得随机事件 $\varepsilon \geq z_j$ 的概率等于 p_j

$$p(\varepsilon \geq z_j) = p_j$$

实际上 z_j 可以由公式近似给出, 即由

$$\frac{1}{\sqrt{2\pi}} \int_{z_j}^{+\infty} e^{-\frac{t^2}{2}} dt = p_j, \text{ 可以近似计算出 } z_j;$$

由于 b_j 已知, 因此可以求到二列相关系数 r_b , 其公式为:

$$r_{b_j} = z_j / b_j \quad (6)$$

最后我们可以通过式(7)求到 a_j 的初值 $a_j^{(0)}$, 其公式为:

$$a_j^{(0)} = r_{b_j} / \sqrt{1 - r_{b_j}^2} \quad (7)$$

2.4 多重迭代条件极大似然估计

多重迭代条件极大似然估计是为了估计原始题的项目参数而综合前面所述诸多方法的一种迭代估计方法。这种估计方法分为三个阶段:

第一阶段: 已知被试能力值, 求项目难度 b_j 的条件似然估计。

第二阶段: 将已获得难度参数 b_j 作为已知, 采

用条件极大似然估计法获得项目区分度参数 a_j 。

第三阶段：将项目区分度参数 a_j 看作已知，第二阶段获得难度参数 b_j 作为条件极大似然估计中项目难度的迭代初值，再次采用条件极大似然估计得到新的 b_j ，重复二、三阶段，直到似然函数的值稳定为止，最后获得的 a_j 和 b_j 的值作为原始题项目参数的估计值。

2.5 评价指标

本研究采用衡量估计值精度的指标有平均绝对偏差 ABS 和平均偏移均方根 RMSD 两个指标。

平均绝对偏差 ABS 的计算公式为：

$$ABS(X) = \sum_{r=1}^R \left(\sum_{i=1}^K |x_i - \hat{x}_{ir}| / K \right) / R \quad (8)$$

平均偏移均方根 RMSD 的计算公式为：

$$RMSD(X) = \sqrt{\sum_{r=1}^R \sum_{i=1}^K (x_i - \hat{x}_{ir})^2 / (K \times R)} \quad (9)$$

其中： $\hat{x}_{ir}$ 是 x_i 的第 r 次试验中的估计值， x_i 是模拟真值， K 为原始题的个数， R 表示重复试验的次数。ABS 指标反映了真值与估计值绝对偏差的平均，RMSD 指标反映了真值与估计值偏移均方根的平均。ABS、RMSD 值越小，估计准确性越高。所以，ABS、RMSD 值是越小越好。在本研究中，需要考虑项目参数 a_j 和 b_j 的估计精度。

3 实验

本文所有实验都是模拟在 CAT 过程中植入原始题，依据被试在测试过程中计算机记录的被试作答信息和自适应测验估计得到的被试能力值后，再对原始题项目参数进行在线校准。对被试能力真值和项目参数真值是采用蒙特卡罗模拟方法给出，其中被试能力值 θ 和项目难度参数 b 和项目区分度参数 a 取对数后服从标准正态分布 $N(0,1)$ 。被试的作答反应和作答反应后的选题均采用陈平等(2006)使用的方法，即“以能力的估计值选题，以能力真值做题”。更详细一点说，即设被试的能力真值为 θ_i ，而当前其能力估计值为 $\hat{\theta}_i$ ，从题库中选取被试 i 尚未使用过的且项目难度与 $\hat{\theta}_i$ 相接近的项目让被试 i 来做。设项目 j 的参数向量为 ζ_j ，依据给定的项目反应模型(本文是 2PLM)，计算项目反应函数 $P(\theta_i, \zeta_j) = P_{ij}$ ，并且在 $[0,1]$ 上的均匀分布中抽取随机数 r ，如果 $r \leq P_{ij}$ 则令 $X_{ij}=1$ ，否则 $X_{ij}=0$ 。而对于原始题的分派方法请参见下文。

在本实验中 CAT 能力估计方法是采用贝叶斯

后验期望估计方法，我们假设能力先验分布为标准正态分布。当测验长度不大于 20 时，积分结点数设为 13 个；测验长度大于 20 时，积分结点数设为 16 个。这样假设主要是为了方便运算。模拟程序始终将能力控制在 $[-3,3]$ 之间。

3.1 原始题植入方法

将原始题与 CAT 题库中正在使用的题目混在一起呈现给被试，让被试在不知情的情况下对原始题进行作答，这种方法称为植入原始题(Wainer, 2000)。对被试能力进行估计时，不使用被试对原始题的作答反应，只是利用其作答反应与经过校准的项目(题库中已有的试题)的关系对原始题的项目参数进行估计。从理论上来说，以这种方式植入原始题不会对被试能力估计造成影响。但从实践的角度来看，影响依然还会存在。因为植入原始题的数量、植入的位置以及原始题的属性都可能会对整个测验有所影响，因此需要特别加以考虑。在施测的过程中，首先要保证植入原始题的数量不应该给被试造成负担，并且在测试过程中要随机植入。但是植入原始题时试题的内容分布、难度分布却很难控制，特别是要不让被试有所察觉就更难于控制(Wainer, 2000)。每个被试都是随机的从原始题库中选取项目作答，同一被试不能作答相同的原始题，并且当某道原始题被作答次数达到了系统设定的次数以后，该道原始题在后面测试中就不会再被选中。

在本文中所有实验采用的是定长的 CAT，系统设定每个被试必须完成 30 道试题(不包括原始题)，且设定每个被试要求作答原始题数为 6。因此植入原始题所采用的方法是：对于每个被试，在测试开始时都产生 6 个互不相同的随机正整数，且 6 个正整数的取值都在 5 至 28 之间，然后在实施 CAT 过程中当被试已完成的试题数(不包括原始题)等于 6 个随机数中任一数值时，则就对该被试植入原始题。原始题选取方法也是产生一个 1 至原始题总数的一个随机正整数，然后选取对应的原始题给被试作答，如果被试已做该原始题则重新选取。当所有被试都完成了自适应测验以后，再对原始题项目参数进行估计。为了保证每个原始题都能达到系统设定的作答次数，则要求(被试的人数 $\times$ 6)必须大于(原始题个数 $\times$ 系统设定原始题被作答次数)，例如：被试人数为 1000，原始题个数 100，系统设定每个原始题被作答次数为 100。100 $\times$ 100 $>$ 1000 $\times$ 6 就无法做到让每一个原始题作答次数达到 100。在对所有原始题被作答次数都达到系统设定的次数以后，本

系统设定后面的被试不需要再作答原始题。本文实验分成两大部分, Monte Carlo 模拟包括实验一、二、三, 其中实验一为 1PLM, 实验二为 2PLM, 实验三为在 2PLM 下采用实测数据的实验。

3.2 实验一

本实验模拟被试人数为 5000 人, CAT 题库容量为 1000 道, 原始题数为 100 道, 这个实验条件只是为了叙述上的方便, 而不是本质上的限定(请参见第 5 节综合讨论)。在 1PLM 模型下原始题项目参数进行估计, 并且对原始题被作答次数(即对某道原始题参与作答的人数)的不同和和原始题难度参数 b 选取方法的不同分别进行实验分析。

实验采用的方法如下: 首先采用夹逼平均法获得的每个原始题项目参数 b , 然后将已经获得参数 b 的值作为条件极大似然估计(CMLE)的初值, 最后估计得到原始题项目参数 b 。每个实验都独立重复模拟 30 次, 结果见表 1。

表 1 采用 CMLE 估计初值的估计结果比较

原始题数	作答次数	ABS(b)	RMSD(b)
100	50	0.2621	0.3333
100	60	0.2497	0.3148
100	70	0.2268	0.2852
100	80	0.2074	0.2638
100	90	0.1985	0.2507
100	100	0.1897	0.2397
100	110	0.1779	0.2260
100	120	0.1713	0.2168
100	130	0.1669	0.2111
100	140	0.1614	0.2033
100	150	0.1538	0.1941

从本实验的结果我们可以看出, 原始题不同作答次数对项目参数估计结果产生不同的影响, 表 1 的数据显示 ABS、RMSD 都随着原始题被作答次数增加而减少, 这说明原始题被作答次数越多估计值就越接近真实值。

3.3 实验二

本实验模拟被试人数为 5000 人, CAT 题库容量为 1000 道, 原始题数为 100 道, 在 2PLM 模型下采用多重迭代极大似然估计法获得原始题项目难度和区分度参数。其中在第一阶段采用夹逼平均法求得原始题项目参数 b 的值作为迭代初值, 实验独立重复模拟 30 次, 结果见表 2。

表 2 采用多重迭代极大似然估计结果的比较

原始题数	作答次数	区分度 a		难度 b	
		ABS	RMSD	ABS	RMSD
100	50	0.2786	0.3829	0.2931	0.4384
100	60	0.2419	0.3363	0.2557	0.3712
100	70	0.2403	0.3319	0.2349	0.3460
100	80	0.2191	0.3033	0.2201	0.3201
100	90	0.2108	0.2930	0.2178	0.3208
100	100	0.1996	0.2722	0.2064	0.3047
100	110	0.1850	0.2551	0.2046	0.3075
100	120	0.1834	0.2545	0.2012	0.3018
100	130	0.1749	0.2397	0.1849	0.2777
100	140	0.1678	0.2324	0.1805	0.2684
100	150	0.1665	0.2273	0.1798	0.2671

从实验数据可以看出, 采用多重迭代极大似然估计法获得原始题项目难度参数和区分度参数结果的估计精度随着原始题被作答的次数的增加而增加。特别是当作答次数达到 100 时, 区分度参数估计结果 ABS 值都在 0.2 以下, 而项目难度参数估计结果 ABS 的值也在 0.2 左右。Baker 和 Kim (2004) 指出要获得较满意的参数估计值样本容量要达到 500, 可见本文的方法对样本容量要求就相当宽松。

3.4 实验三

由于我们开发的 CAT 系统暂时还没有广泛应用, 因此要想在实际应用中来验证原始题在线校准项目参数则比较困难。本实验根据计算自适应测试时植入原始题原理, 从 2007 年“中小学生学习质量分析、反馈与指导系统”(SAAE)英语学科中抽出了 6000 个考生对 80 道选择题真实作答数据进行实证研究, 具体方法如下:

第一步: 对抽中的 6000 个学生的作答数据采用 EM 算法对被试的能力和项目参数进行估计, 获得每个被试的能力值和每个项目的参数值。

第二步: 将第一步估计得到的每个被试能力值看作为是被试在参加自适应测验后估计得到的能力值, 试题项目参数看作为原始题项目参数的真值。

第三步: 把 80 道试题看作为原始题, 模拟被试参加自适应测验时植入原始题的方法, 即: 对每一道试题随机从 6000 个被试中选出若干个被试在试卷上的真实作答情况为其反应结果。由此可以得到对每个原始题的真实作答反应矩阵。而被试的能力值则取第一步估计得到的能力值。

第四步: 通过运用夹逼平均法求得每道试题的

难度值。再采用多重迭代条件极大似然估计方法求得原始题的项目参数。

对于规定的每个项目不同作答次数的实验都独立重复模拟 30 次, 估计结果见表 3。

表 3 采用真实作答数据估计结果的比较

原始题数	作答次数	区分度 a		难度 b	
		ABS	RMSD	ABS	RMSD
80	50	0.2694	0.4055	0.3685	0.5337
80	60	0.2278	0.3426	0.3328	0.4744
80	70	0.2132	0.3179	0.3060	0.4382
80	80	0.1999	0.3046	0.2961	0.4284
80	90	0.1967	0.2961	0.2783	0.4175
80	100	0.1801	0.2713	0.2701	0.4042
80	110	0.1806	0.2756	0.2635	0.3929
80	120	0.1729	0.2682	0.2582	0.3839
80	130	0.1610	0.2514	0.2463	0.3658
80	140	0.1570	0.2428	0.2392	0.3659
80	150	0.1520	0.2344	0.2270	0.3543

从表 3 数据可以看出, 采用真实作答数据模拟自适应测验时植入原始题对其项目参数估计精度也是随着作答次数的增加而增加。但是与模拟实验结果表 2 的数据相对, 难度参数的估计精度都要逊色一些。对于实证实验和模拟实验估计精度存在一定的误差, 其原因可能是因为实证实验中原始题项目参数的真值是通过估计得到(它本身至少带有一定的估计误差)。

4 综合讨论

2007 年在日本东京召开的国际心理测量学大会上, 在日本工作的中国人韩太哲先生(Han Taizhe)提出了自适应参数估计方法(Adaptive Estimating Method, 简称 AEM), 即在被试参加自适应测试的时候插入一些新的试题从而收集被试对新题的作答反应, 再对试题进行参数估计(Han, 2007), 但该方法只是对 1PLM 适用。本文模拟在实施 CAT 过程中植入原始题, 分别在 1PLM 和 2PLM 模型下对原始题目的项目参数进行估计。研究结果表明: 1) 无论是在 1PLM 还是 2PLM 模型下, 原始题不同作答次数对项目参数估计结果都会产生不同的影响, 并且对每个原始题作答人数的增加也会使项目参数估计精度提高。2) Hulin 等人在 1982 年模拟结果表明: 应用联合似然估计手段, 在 2PLM 模型下, 项目数为 30, 至少需要 500 人, 才能保证估计结果精度比较高(转引自 Baker & Kim, 2004)。而在本文

的模拟实验中, 由表 2 的结果可以看出人数不少于 100 人时, 项目参数的估计结果 ABS 都在 0.2 左右。但需要说明的是, 本文仅讨论能力已知时项目参数的估计问题, 显然这是一个条件估计。

在本文已列举实验的基础上, 通过大量的模拟实验还发现, 采用多重迭代条件极大似然估计方法获取原始题的项目参数, 对已有题库的题量没有太多的限定, 同时对于每次估计的原始题个数也没有限定。因此它对于在一个小型题库基础上扩充题库的意义就不言而喻。同时本方法也适用于计算机化考试(Computer-Based Test, 简称 CBT)中题库的扩充。

另外, 如果不仔细考虑, seeding items 的方法事实上可能对考生反应构成一定的不利影响, 必须想办法加以克服。

荷兰测量学家 Verschoor 和 Straetmans 在 MATHCAT: A flexible testing system in mathematics education for adults 一文中讲到在实施 CAT 时植入项目(seeding items), 还讲了实施办法: 可见 seeding items 已应用于荷兰成人数学教育使用的 CAT 之中。仿照荷兰测量学家的办法, 我们也有如下一个初步的想法:

在题库建设过程请命题专家为每个项目提供一个 CTT 难度值。

1) 命题专家可以大体感觉得到各个项目的通过率, 即估计出 CTT 意义上的难度, 作为原始题一个附加指标; 同样在题库建设过程也请命题专家为每个项目提供一个 CTT 难度值。

2) 对每个考生而言在 CAT 实施前期, 植入的原始题为 CTT 意义上的中等难度题, 中期, 比照所估计的能力, 找相应的项目。比如: 计算植入原始题前被试已经作答的 2~3 道试题的 CTT 难度的一个均值, 然后从原始题题目集合中找到与均值接近的原始题给被试作答; 或者比照 CAT 估计的能力值, 确定高、中、低能力值的范围, 分别对应 CTT 相应的难度。在 CAT 实施后期, 也可以参照上述方法实施。

当然, 我们以上所说的方法不一定成熟, 但这至少可以避免能力较低的考生考太难的项目而引起焦虑, 也不会因为高能力考生考太易的项目而粗心。在实际实施 CAT 时, 还可以不断改进方法。

Chang 和 Lu (2009)一文中整合自适应设计、序贯估计和测量误差模型以解决在线项目校准问题, 使用了很多概率统计的理论和方法。在没有一点先验的信息可以利用(即题库中完全没有一个项目)的

条件下, Makransky (2009) 的文章中讨论自适应测验中自动在线校准设计问题。这和本文的条件不同, 因此其效果(估计精度)也没有本文的理想。然而该文提供了一个思路, 即先设原始题集中所有项目的难度参数都为 0, 对若干个人施测后, 再用本文提出的新方法, 重估项目参数。这个思路是否可行, 如何将 Makransky 和本文的方法进行有效结合, 如何对 3PLM 和多级评分模型下的在线参数估计这都有待进一步的研究和探讨。

参 考 文 献

- Baker, F. B., & Kim, S. H. (2004). *Item Response Theory: Parameter estimation techniques*, (2nd eds.). New York, Marcel Dekker.
- Chang, H. H., & Ying, Z. L. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 211–222.
- Chen, P., Ding, S. L., Lin, H. Q., & Zhou, J. (2006). Item selection strategies of computerized adaptive testing based on Graded Response Model. (in Chinese). *Acta Psychologica Sinica*, 38(3), 461–467.
- [陈平, 丁树良, 林海菁, 周婕. (2006). 等级反应模型下计算机化自适应测验选题策略. *心理学报*, 38(3), 461–467.]
- Chang, Y-c. I., & Lu, H. (2009). Online calibration via variable length computerized adaptive testing. *Psychometrika*, 75(1), 140–157
- Han, T. Z. (2007). *Estimating Item Parameter Adaptively in CAT-With one Parameter Model*. IMPS2007, Tokyo.
- Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling and Linking*, (2nd ed.). Springer Verlag.
- Makransky, G. (2009). An automatic online calibration design in adaptive testing (published online). The 2009 GMAC Conference on Computerized Adaptive Testing.
- Qi, S. Q., Dai, H. Q., Ding, S. L. (2002). *Principles of Modern Educational and Psychological Measurement* (pp.141–142). Beijing, China: Higher Education Press. (in Chinese).
- [漆书青, 戴海崎, 丁树良. (2002). *现代教育与心理测量学原理*(pp.141–142). 北京: 高等教育出版社.]
- Ross, S. M. (2007). *Simulation*, (4th ed.), Beijing, China: Posts & Telecom Press.
- Verschoor, A. J., & Straetmans, G. J. J. M. (2000). MATHCAT: A flexible testing system in mathematics education for adults. In W. J. van der Linden & C. A. W., Glas. (eds.). *Computerized adaptive testing: theory and practice*, (pp.101–116). Kluwer Academic Publishers. Netherlands.
- Wainer, H., Dorans, N. J., Eignor, D., Flanger, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A Primer*, (2nd ed.) (pp.85–86, 121–122, 161–185). NJ: Lawrence Erlbaum Associates, Publishers.
- Wang, N. C. (1984). *Brief introduction to numerical analysis*. High Educational Press. (in Chinese).
- [王能超. (1984). *数值分析简明教程*. 高等教育出版社.]

Parameter Estimation of the Raw Item in Computerized Adaptive Testing

YOU Xiao-Feng^{1,2}; DING Shu-Liang¹; LIU Hong-Yun³

(¹ Jiangxi Normal University, Nanchang 330027, China)

(² Foreign Language Teaching and Research Press, Beijing 100089, China)

(³ School of Psychology, Beijing Normal University, Beijing 100875, China)

Abstract

Along with the development of computer technology and increasing needs of individual learning, Computerized Adaptive Testing (CAT) has received more and more attention. However, the problem of the test security is becoming a new challenge to the CAT, for example, excessive exposure of the test items might weaken the efficiency and equity of the CAT. Therefore, constructing a large-scale and high-quality CAT item bank is highly demanded

The traditional method of constructing a CAT item bank involves four steps: first, the items are developed by the experts according to the test specification and blueprint; second, a representative sample of examinees is recruited to take the test; Then, the calibration of item parameters and testing of the model-data fit will be conducted after-ward. Finally, items of high-quality and its item parameters will be added into the item bank based on the pre-analysis.

However, several problems exist in the above traditional method. First, a large number of test takers are needed, which is time-consuming and expensive. Second, the security of the items cannot be guaranteed in the

pilot test. Third, complicated techniques of equating may have to be used to construct a good item bank, which may also affect the security of the items, especially when the anchor items are repeatedly used in the test equating process.

Thus, if the raw items can be seeded in the CAT process and the item parameters combined with the examinee abilities estimated at the same time, it will be significant for the construction of CAT item bank. The research in this area has not been widely conducted and reported in the domestic journals, although this issue has been a big topic for foreign researchers. Thus, This study aims to explore how to insert raw items and estimate the item parameters in CAT and investigate the efficient strategies with high security for enlarging the item bank. A new online calibration method is proposed based on the principle of adaption in CAT, provided there is a small-scale item pool. And the formula of initial value during the stage of the iteration is determined. When the ability parameter has been estimated, the conditional maximum likelihood estimation(CMLE) will be implemented to estimate item parameters in the raw items. First, the difficulty parameter could be obtained through CMLE. Second, the discrimination parameter could be estimated with the abilities and difficulty parameter and the initial value of the discrimination parameter could be gained through CMLE. Third, the difficulty parameter could be estimated when the discrimination parameter and abilities are known .Repeat the second and the third steps until the stop condition is satisfied.

Simulation of Monte Carlo has been employed to estimate the parameters of raw items with the one-parameter Logistic and two-parameter Logistic models in the study and good result has been gained.

Key words raw items; CAT; squeezing average method; multi-iterative maximum likelihood estimate method; simulation of Monte Carlo