心 理 学 报 2006,38(3):461~467
Acta Psychologica Sinica

等级反应模型下计算机化自适应测验选题策略*

陈 平[1],丁树良[1],林海菁[1,2],周 婕[1]

([1]江西师范大学计算机信息工程学院,南昌 330027) ([2]江西工业贸易职业技术学院,南昌 330100)

摘 要 计算机化自适应测验(CAT)中的选题策略,一直是国内外相关学者关注的问题。然而对多级评分的 CAT 的选题策略的研究却很少报导。本研究采用计算机模拟程序对等级反应模型(Graded Response Model)下 CAT 的四种选题策略进行研究。研究表明:等级难度值与当前能力估计值匹配选题策略的综合评价最高;在选题策略中增设"影子题库"可以明显提高项目调用的均匀性;并且不同的项目参数分布或不同的能力估计方法都对 CAT 评价指标有影响。

关键词 等级反应模型,计算机化自适应测验,选题策略,影子题库。

分类号 B841

1 引言

计算机化自适应测验(CAT)依靠大型题库,采用项目反应理论(IRT),自行去适应被试水平,灵活施测难度最恰当而性能优良的项目,从而实现对被试的高效测量[1]。目前,CAT 在很多领域得到了广泛的应用,TOEFL、GRE 等世界范围的大型考试都是采用这种考试形式。美国的医生护士资格考试以及军事服役职业能力倾向成套测验也相继推出了 CAT 版本[2]。在我国,CAT 的研究与应用已有不少,国家汉语考试中心举办的汉语水平考试 HSK、第四军医大学对应征者进行的文化水平测验都是使用 CAT 形式;江西师范大学二十年来成功编制了多个 CAT。

国外的考试与测验,一个时期过分偏好客观题(选择题),这就造成 CAT 的许多研究与应用都是建立在 0-1 评分模型的基础上。而在我国,考试测验在重视客观题的同时,还十分重视主观题的运用,比如说,填空题、计算题、论述题等。许多实测资料都是多等级评分资料,等级数有的还超过十个或二十个(例如作文题)。事实证明,使用多级评分项目比使用 0-1 评分项目可以获取更多的被试信息[2]。所以,为了适应我国的考试现状,更为了提高考试质量,很有必要研究多级评分模型下的 CAT。

选题策略是 CAT 的一个重要组成部分,选题策略的好坏直接影响到测验的安全性、测验的准确性以及测验的效率。在 0-1 评分模型 CAT 中,选题策略已经比较成熟,Lord 提出的最大费舍信息量方法、Sympson 和 Hetter 提出的 SH 方法以及张华华提出的 a-stratified 方法都很具有代表性。而对多级评分模型下 CAT 选题策略的研究,国内刊物未见公开报导,国外也还不太多。于是,多级评分模型下 CAT 的选题策略理应成为学者们关注的热点之一,但是由于多级评分模型的复杂性,使得这一研究成为一个难点;同时又由于目前 CAT 使用的题型多为 0-1 评分题型,使得对这一问题的研究相对于应用来讲具有超前性。尽管 Dodd,De Ayala 和 Koch 对多级评分 CAT 的题库、选题策略、能力估计方法和终止规则都作了一些研究[2,3],但是,对于不以信息量为衡量标准的选题策略,对于多级评分项目题库中项目参数的分布以及对于等级反应模型(GRM)中能力估计使用 EAP 方法,特别是迭代时使用 Fisher-Score方法等等问题均未涉及。本文研究了 Samejima 等级反应模型下 CAT 的几种选题策略以及影响 CAT 评价的一些因素。

收稿日期:2005-05-09

* 本研究受国家自然科学基金项目(60263005)、江西省自然科学基金(0411021)和江西省教育厅科技项目资助,并部分受教育科学规划重点课题(DBB010501)和江西省分布计算工程技术研究中心开放课题基金(江西师范大学)资助。

通讯作者:丁树良,E-mail:ding06026@163.com,电话:0791-8786468

2 GRM 下不同选题策略

2.1 GRM 简介

Samejima 在逻辑斯蒂克(Logistic)模型的框架下,建立了等级反应模型 GRM,突破了过去项目反应模型只能用于 0 -1 评分项目的限制[1]。

GRM 假设每个项目只有一个区分度值、有多个难度等级值,而且每个项目在各个等级上的难度值是严格单调递增的。若项目 j 有 f_j 个等级(即有 f_j 个分值),则

$$b_{j1} < b_{j2} < \cdots < b_{j,fj}$$

Samejima (1989)提出分两步获得能力为 θ 的被试在某个项目上恰得某个等级分的概率。

第一步:能力为 θ_α 的被试 α 在第 j 个项目上得分不低于(等于或高于)t 分的概率表示为

$$\left.\begin{array}{l} P^*_{\alpha j0}=1,\quad P^*_{\alpha j,fj+1}=0; \\ \text{对于 } t=1,2,\cdots,f_j, \\ P^*_{\alpha jt}=\{1+\exp[-Da_j(\theta_\alpha-b_{jt})]\}^{-1} \end{array}\right\} \quad (1)$$

其中,D 为量表因子,一般取 1.7,a_j 为第 j 个项目的区分度。

第二步:能力为 θ_α 的被试 α 在第 j 个项目上恰得 t 分的概率 $P_{\alpha jt}$ 表示为:

$$P_{\alpha jt}=P^*_{\alpha jt}-P^*_{\alpha j,t+1} \quad (t=0,1,2,\cdots,f_j) \quad (2)$$

称(2)为 GRM 的运算特征函数(Operating Characteristic Function)。

2.2 GRM 选题策略

为了表述简洁,我们先引入剩余题库这一概念。在 CAT 实施过程中,对被试 α 而言,所谓剩余题库是指题库中该被试尚未作答的项目的全体。用 $L\alpha$ 表示当前剩余题库中所有项目编号的集合。

CAT 是根据被试当前的能力估计值,从剩余题库中选择最适合被试作答的项目对被试施测。在 0 -1 评分模型中,每个项目只有一个难度等级,难度值与被试能力值最接近的项目认为是最适合被试作答的项目。而在 GRM 中,每个项目有多个难度等级,于是有学者提出用难度等级平均数 $\bar{b}j$ 或者难度等级中位数 $bj(med)$[4]作为整个项目的难度值。若 fj 表示第 j 个项目的难度等级数,b_{jt} 表示第 j 个项目的第 t 个难度等级值,则:

$$\bar{b}_j=\frac{1}{f_j}\sum_{t=1}^{f_j}b_{jt}$$

$$b_j(med)=\begin{cases}\{b_{jt},\ t=1,2,\cdots,\ f_j\}\text{的中间一个,当}f_j\text{为奇数}\\ \{b_{jt},t=1,2,\cdots,\ f_j\}\text{的中间两个的算术平均,当}f_j\text{为偶数}\end{cases}$$

$\bar{b}_j$ 和 $b_j(med)$ 这两个量刻画了难度等级的"中心位置"。在此基础上,有人提出了基于 GRM 的两种 CAT 选题策略:难度等级平均数与能力匹配策略和难度等级中位数与能力匹配策略[4]。

定义 1 难度等级平均数与能力匹配策略(以下简称为平均数法),即从剩余题库中调用难度等级平均数与被试当前能力值最接近的项目。

定义 2 难度等级中位数与能力匹配策略(以下简称为中位数法),即从剩余题库中调用难度等级中位数与被试当前能力值最接近的项目。

另外,本文又提出了两种选题策略:去两端平均法和等级难度值与能力匹配策略。

定义 3 去两端平均法

假设项目 j 有 $f_j(f_j\geq 3)$ 个等级,则去两端平均法为:从剩余题库中调用使下式值最小的项目 j,

对于 $j\in L\alpha$,

$\sum_{t=2}^{f_j-1}|b_{jt}-\hat{\theta}|/(f_j-2)$

其中 $\hat{\theta}$ 为被试能力估计值。去两端平均法考虑到在两端的难度等级可能会出现极端值,所以没有使用最低难度等级信息和最高难度等级信息。事实上,如果项目有三个难度等级,不考虑最高和最低的难度等级,去两端平均法就变成了中位数法。

对 0 -1 评分模型选题策略的思想进行拓展,就有以下的等级难度值与能力匹配策略。

定义 4 等级难度值与能力匹配策略(以下简称等级难度匹配法)

等级难度匹配法为:从剩余题库中调用使下式值最小的项目 j,

$$|b_{jt}-\hat{\theta}| \quad (j\in L\alpha;\ t=1,2,\cdots,f_j)$$

该选题策略将被试当前能力值与剩余题库中所有项目的所有难度等级值进行比较,调用与被试当前能力值最接近的难度等级所在的项目。如果采用 0 -1 评分,即 $f_j=1$ 时,这一选题策略便是难度与当前能力估计值相匹配的策略。

基于项目调用均匀性的考虑,可在选题策略中增设"影子题库",也即根据被试能力估计值从剩余题库中抽取一批最适合的项目(称为影子题库),然后从中随机调用一个让被试作答。本文在对这四种

选题策略进行比较研究的同时,也考虑了“影子题库”对项目调用均匀性的影响。为了便于比较,将随机选题策略作为参照。随机选题策略中没有“影子题库”的概念,所以对随机选题策略,我们没有增设“影子题库”的环节。

3　Monte Carlo 模拟试验

在对 GRM 下 CAT 选题策略进行探讨的基础上,编制了相应的计算机模拟程序,用常用的评价 CAT 的四个指标[4~6]对各选题策略进行比较,并且探讨了“影子题库”对项目调用均匀性的影响、不同的项目参数分布对 CAT 评价指标的影响及能力估计方法的优劣。

3.1　Monte Carlo 模拟试验

记 $N(x,y)$ 表示期望为 x、方差为 y 的正态分布,$U(a,b)$ 表示在[a,b]上的均匀分布。

采用 Monte Carlo 方法模拟被试能力真值和项目参数。

3.1.1　被试和题库　首先模拟生成一批被试(1000 人),被试能力真值服从标准正态分布($\theta \sim N(0,1)$)。然后根据不同的项目参数(包括难度参数 b 和区分度参数 a)分布情形,模拟生成四个题库。第一个题库中,$b \sim N(0,1)$, $lna \sim N(0,1)$;第二个题库中,$b \sim U(-3,3)$,$lna \sim N(0,1)$;第三个题库中,$b \sim N(0,1)$, $a \sim U(0.1,2.3)$;第四个题库中,$b \sim U(-3,3)$,$a \sim U(0.1,2.3)$。这 4 个题库的相同之处在于:项目数均为 1000 题,且每个项目均有 6 个难度等级;b 介于 -3 至 3 之间, a 介于 0.1 至 2.3 之间。

3.1.2　能力估计方法　采用贝叶斯后验期望法(EAP)和条件极大似然估计法(MLE)估计被试能力。在实现极大似然估计法时,又分别用 Fisher - Score 迭代方法(记为 F - S)和牛顿 - 拉夫逊迭代方法(记为 N - R)求解似然方程[1]。

3.1.3　影子题库　对上述四种选题策略既考虑增设“影子题库”的情况,也考虑不使用“影子题库”的情况。

3.1.4　试验设计　对各选题策略进行比较的同时,为了探讨 4 种项目参数分布对 CAT 评价指标的影响、3 种能力估计方法的优劣及增设“影子题库”是否对项目调用均匀性有影响,本研究采用 4 × 3 × 2 交叉设计,共 24 个试验。每次试验对 1000 个被试均模拟执行 30 次 CAT 全过程(详见 3.2.1),每次试验都对四种选题策略进行比较。

3.2　CAT 模拟

本研究通过编制计算机程序模拟实现 CAT 的全过程,主要有以下一些模拟步骤:

3.2.1　CAT 全过程模拟　整个测试过程分为探测性阶段和正式测试阶段。在探测性阶段,由于对被试的能力水平一无所知,只能从题库或剩余题库中随机抽取不同的项目给被试进行模拟作答(详见 3.2.2),直到作答题数不少于 3 且作答总分既不为 0 分也不为满分时,结束探测性阶段。根据被试的作答得分向量,估计能力初值,进入正式测试阶段。在正式测试阶段,针对不同的选题策略,从剩余题库中调用与被试当前能力值匹配的项目(或从“影子题库”中随机抽题)。被试作答完后,估计出新的能力值,并且计算测验信息量,直到测验信息量达到目标信息量(即可允许的测量误差),结束整个测试。模拟过程中,能力估计迭代精度定为 0.01,“影子题库”大小定为 5,目标信息量定为 25。

3.2.2　被试作答模拟　根据被试能力真值 $\theta\alpha$ 和当前项目 j 的参数,分别计算被试 α 在第 j 个项目上得分不低于各分值的概率,即 $P^*_{aj1}, P^*_{aj2}, \cdots, P^*_{aj,fj}$(见 1 式)。再生成落在区间(0,1)上的随机数 r,若 $r \in (P^*_{aj,t+1}, P^*_{ajt})$ $(t=0,1,2,\cdots,f_j)$,则可以确定被试 α 在第 j 个项目上得分为 t[4]。

3.3　评价指标

选题策略的优劣直接关系到 CAT 的质量,当其它条件固定仅改变选题策略时,对 CAT 的评价实际上就是对选题策略的评价,故本文用常用的评价 CAT 的四个指标对选题策略进行评价。

3.3.1　能力估计准确性　能力估计准确性通常用返真性(*Recovery*)来衡量,计算方法如下:

$$Recovery = \frac{1}{C}\sum_{j=1}^{C}\sum_{i=1}^{N}|\theta_i - \hat{\theta}_{ij}|/N$$

其中,N 表示被试人数,C 表示模拟重复的次数。θ_i 为被试 i 的能力真值,$\hat{\theta}_{ij}$ 代表被试 i 在第 j 次模拟得到的能力估计值。因此,*Recovery* 指标反映了能力真值与能力估计值的绝对偏差的平均。*Recovery* 越小,能力估计准确性越高。

3.3.2　选题策略的稳定性　用能力估计标准差(*Se*)作为衡量选题策略是否稳定的指标。*Se* 的计算公式如下:

$$Se = \frac{1}{N}\sum_{i=1}^{N}\left(\sqrt{\sum_{j=1}^{C}(\hat{\theta}_{ij} - \bar{\hat{\theta}}_i)^2/C}\right)$$

其中 $\bar{\hat{\theta}}_i = \frac{1}{C}\sum_{j=1}^{C}\hat{\theta}_{ij}$,表示被试 i 在 C 次模拟中得到的能力估计平均值。*Se* 指标反映了被试能力估

计值的离散程度。*Se* 值越小，离散程度就越小，选题策略就越稳定。

3.3.3 项目调用均匀性 项目调用均匀性是度量CAT安全性的一个重要方面，通过项目调用次数标准差衡量，计算公式为：

$$\text{项目调用次数标准差} = \frac{1}{C}\sum_{j=1}^{C}\left(\sqrt{\sum_{i=1}^{L}(m_{ij}-\bar{m}_j)^2/L}\right)$$

其中 L 为题库中项目总数，m_{ij} 表示项目 i 在第 j 次模拟中被调用的次数，$\bar{m}_j = \frac{1}{L}\sum_{i=1}^{L} m_{ij}$ 表示在第 j 次模拟中项目被调用的平均次数。项目调用次数标准差越小，说明项目被调用的均匀性越好，测验的安全性就越高。

3.3.4 测验效率 评价测验效率的高低，采用人均用题数指标，公式为：

$$\text{人均用题数} = \sum_{j=1}^{C}\left(\sum_{i=1}^{N} p_{ij}/N\right)/C$$

其中 p_{ij} 表示被试 i 在第 j 次模拟中的作答项目数。人均用题数越少，说明测验效率越高。

用上述四个指标对选题策略进行评价难以给人一个整体印象，为此我们采用统一量纲加权求和的综合评价方法[4]。本文认为上述四个指标都同等重要，所以赋加权系数均为1。上述四个指标值都是越小越好，因此将各选题策略在某评价指标上的最小值作为分子，分别除以各选题策略在该评价指标上的值，即得到在该评价指标上统一量纲的结果。加权求和值越大的选题策略，综合评价越高。

4 试验结果与分析

研究结果见附表1，限于篇幅，我们没有将随机选题策略的试验结果列入附表1中。附表1数据A/B/C中的A、B和C分别是CAT模拟过程中采用EAP方法、F－S方法和N－R方法估计能力得到的结果。另外，由于F－S方法和N－R方法在估计能力时经常出现迭代不收敛的情况，因此把人均迭代失败次数作为参考指标列入研究结果（见附表1中I5）。

首先对采用EAP估计得到的结果进行分析，我们可以发现，各选题策略中增设“影子题库”能够在大致保持（有时甚至提高）能力估计精度的前提下，降低项目调用次数标准差，提高项目调用均匀性。四种选题策略的能力估计准确性都很高，而且相差不大，这是因为在不定长CAT中，测验终止的标准都是达到指定的测量精度，所以能力估计精度差异不大。在能力估计标准差指标上，各选题策略存在差别，但差别不大，说明各选题策略稳定性接近。另外，等级难度匹配法的项目调用均匀性最好，而且效果非常明显。各选题策略的人均用题数也存在差别。通过对上述四个评价指标统一量纲加权求和（限于篇幅，只给出 $b \sim N(0,1)$，$lna \sim N(0,1)$ 时的结果见附表2）后发现，等级难度匹配法的综合表现都最好。我们还发现当区分度服从均匀分布时，各选题策略的人均用题数明显少于区分度服从对数标准正态分布时。这主要是因为区分度均匀分布时，测验过程中调用高区分度项目的概率较区分度服从对数标准正态分布时大，而信息量又与区分度的平方成正比，所以被试作答较少的项目就能达到目标信息量。

采用F－S能力估计方法的研究结果与采用EAP能力估计方法的研究结果基本一致，但也存在特殊情况。如当难度服从标准正态分布、区分度服从对数标准正态分布时，在平均数法中增设“影子题库”并未提高项目调用均匀性。这说明除极个别特殊情况，在选题策略中增设“影子题库”还是可以明显提高项目调用的均匀性。采用N－R能力估计方法的研究结果则与EAP能力估计方法的研究结果基本一致。

类似于附表2，我们可以得到所有24次试验的统一量纲加权求和结果（限于篇幅，我们就不一一列出）。结果显示各选题策略的综合排名由高至低依次是：等级难度匹配法，中位数法，平均数法，去两端平均法。

比较三种能力估计方法的估计精度（也即比较返真性指标中数据A/B/C的A、B和C值）可以发现，EAP方法稍微优于F－S方法（当 $b \sim N(0,1)$，$a \sim U(0.1,2,3)$ 而且增设“影子题库”时，平均数法和去两端平均法的F－S估计更精确，这说明F－S方法偶尔优于EAP方法），而F－S方法又明显优于N－R方法。这主要是因为F－S方法和N－R方法在能力估计过程中存在迭代不收敛情况，导致能力估计精度下降，其中采用N－R能力估计方法产生的人均迭代失败次数又远远多于F－S能力估计方法。文献[2]中提到：Chen，Hou，Fitzpatrick & Dodd（1995）在基于Andrich评分量表模型（ARSM）的CAT中比较了EAP方法和MLE方法，也认为EAP方法在能力估计返真性方面要优于MLE方法[2]；另外，在基于0－1评分模型的CAT系统中，Reckase（1981）& Urry（1977）推荐使用项目难度参数服从均匀分布的题库，因为与难度参数服从标准正态分

布的题库相比,估计能力时会产生更少的不收敛次数[2]。附表 1 中绝大部分数据也可以印证这一点。

我们对上述的 24 个试验进行了多次重复,都得到了类似的结论。另外,我们对项目难度等级数为五等级的情况也做了大量模拟试验,也发现类似的结果。

5　讨论

在研究过程中,我们发现采用 F - S 方法和 N - R 方法估计能力时,大量迭代不收敛,这严重影响能力估计精确性和选题策略的稳定性。本研究中的模拟程序只是将能力迭代值控制在[- 3,3]之间,并且规定每次能力估计时迭代循环次数不超过 20 次。因此,如何更加有效地处理能力估计过程中迭代发散的问题,是一个亟待解决的课题。另外,张华华博士提出的 a - stratified 方法受到越来越多的关注[5]。研究表明,a - stratified 方法在不损失能力估计精度的情况下,能够提高项目调用的均匀性。本文研究的四种选题策略并没有采用 a - stratified 方法,这将作为后继研究的一个内容。还有,本文研究了项目难度等级为五、六等级的情况,对于题库中包括不同难度等级数项目的情况,我们也将作进一步的探讨。而且,采用真实考试数据对等级反应模型下的 CAT 选题策略进行比较研究,也将是我们下一步的研究方向。

6　结论

通过以上模拟研究,可以得出以下结论:

(1)各选题策略的综合评价由高至低依次是:等级难度匹配法,中位数法,平均数法,去两端平均法。

(2)增设"影子题库"可以降低各选题策略的题目调用次数标准差,又以等级难度匹配法降低的幅度最大。而且在提高项目被调用均匀性的同时,能力估计精度并没有明显下降,有时甚至提高。

(3)等级难度匹配法的项目调用均匀性最好,而且效果非常明显。

(4)与区分度服从对数标准正态分布相比,区分度服从均匀分布时各选题策略的人均用题数明显减少。

(5)Fisher - Score 能力估计方法产生的人均迭代失败次数远远少于 N - R 方法;而且在同一种能力估计方法中,难度均匀分布比难度正态分布时产生的人均迭代失败次数要少。

(6)能力真值服从标准正态分布时, EAP 方法估计能力效果最好,Fisher - Score 方法次好,N - R 方法最差。当能力真值分布不确定时,可以采用 Fisher - Score 方法对能力进行估计。

参　考　文　献

1　Qi S Q, Dai H Q, Ding S L. Principles of modern educational and psychological measurement (in Chinese). Beijing: Higher Education Press, 2002
(漆书青, 戴海琦, 丁树良. 现代教育与心理测量学原理. 北京: 高等教育出版社, 2002)

2　Barbara G D, De Ayala R J, William R K. Computerized adaptive testing with polytomous items. Applied Psychological Measurement, 1995, 19 (1): 5 ~ 22

3　Barbara G D, William R K, De Ayala R J. Operational characteristics of adaptive testing procedures using the graded response model. Applied Psychological Measurement, 1989, 13 (2): 129 ~ 143

4　Chen D Z. Comparison study of item selection strategies of computerized adaptive testing with the Samejima graded response model (in Chinese). Master's dissertation of Jiangxi Normal University, 2004
(陈德枝. Samejima 等级反应模型下 CAT 选题策略的比较研究. 江西师范大学硕士学位论文, 2004)

5　Wen J B, Hou J T. The application of a - stratified method to the unfixed - length CAT (in Chinese). The fifth Symposium on Psychological and Educational Testing in Chinese Communities, 2001
(文剑冰, 侯杰泰. A - stratified 方法在不定长度 CAT 中的应用. 第五届华人社会心理与教育测量学术研讨会, 2001)

6　Yi Q, Chang H H. A - stratified CAT design with content - blocking. British journal of mathematical and statistical psychology, 2003, 56 (2): 359 ~ 378

附表 1 4×3×2 交叉设计试验结果

项目参数分布	是否使用影子题库	评价指标	平均数法	中位数法	去两端平均法	等级难度匹配法
难度服从标准正态分布,区分度服从对数标准正态分布	否	I1	0.169 / 0.184 / 0.425	0.170 / 0.178 / 0.319	0.168 / 0.181 / 0.327	0.169 / 0.177 / 0.219
		I2	0.207 / 0.239 / 0.824	0.208 / 0.220 / 0.614	0.205 / 0.237 / 0.626	0.207 / 0.218 / 0.410
		I3	57.975/63.140/70.791	51.233/56.991/62.624	71.193/72.447/75.243	17.257/16.999/17.547
		I4	29.756/30.772/37.611	31.971/32.758/40.883	30.724/31.731/39.148	29.323/29.842/29.338
		I5	/ 1.146 / 11.022	/ 0.592 / 11.229	/ 0.967 / 10.846	/ 0.142 / 1.076
	是	I1	0.168 / 0.186 / 0.350	0.170 / 0.180 / 0.314	0.171 / 0.188 / 0.301	0.169 / 0.178 / 0.222
		I2	0.205 / 0.249 / 0.702	0.208 / 0.231 / 0.612	0.209 / 0.262 / 0.586	0.207 / 0.218 / 0.419
		I3	57.320/63.355/67.923	50.493/55.963/61.095	67.716/69.351/71.731	8.831 / 9.044 / 9.789
		I4	29.739/30.839/35.255	31.996/32.898/40.405	30.789/32.001/37.321	29.296/29.729/29.194
		I5	/ 1.181 / 7.985	/ 0.710 / 10.531	/ 1.179 / 8.491	/ 0.101 / 1.068
难度服从[-3,3]上的均匀分布,区分度服从对数标准正态分布	否	I1	0.164 / 0.168 / 0.219	0.164 / 0.166 / 0.252	0.164 / 0.167 / 0.244	0.165 / 0.168 / 0.218
		I2	0.198 / 0.205 / 0.385	0.199 / 0.204 / 0.479	0.199 / 0.205 / 0.456	0.200 / 0.205 / 0.382
		I3	45.126/51.273/52.667	39.017/44.251/46.449	87.586/84.671/83.459	18.189/17.900/17.737
		I4	30.836/30.738/30.796	30.734/30.770/30.752	31.737/31.525/31.541	30.317/30.420/30.513
		I5	/ 0.152 / 0.633	/ 0.093 / 0.976	/ 0.081 / 0.886	/ 0.133 / 0.704
	是	I1	0.166 / 0.166 / 0.235	0.164 / 0.168 / 0.232	0.165 / 0.166 / 0.233	0.166 / 0.168 / 0.219
		I2	0.199 / 0.203 / 0.447	0.198 / 0.205 / 0.421	0.200 / 0.204 / 0.431	0.201 / 0.206 / 0.389
		I3	41.201/46.957/48.275	36.813/42.332/44.533	83.396/80.976/79.753	9.593 / 9.337 / 9.330
		I4	30.636/30.552/30.591	30.849/30.821/30.867	31.991/31.857/31.840	30.063/30.114/30.144
		I5	/ 0.137 / 0.860	/ 0.087 / 0.796	/ 0.087 / 0.787	/ 0.117 / 0.709
难度服从标准正态分布,区分度服从[0.1,2.3]上的均匀分布	否	I1	0.177 / 0.179 / 0.387	0.176 / 0.203 / 0.388	0.176 / 0.180 / 0.358	0.176 / 0.179 / 0.290
		I2	0.217 / 0.220 / 0.769	0.215 / 0.299 / 0.768	0.216 / 0.221 / 0.711	0.216 / 0.219 / 0.605
		I3	35.970/39.047/45.732	27.869/31.402/39.179	49.342/49.628/53.284	10.331/ 9.997/ 11.330
		I4	18.072/18.231/21.621	17.755/18.413/23.062	18.490/18.677/22.792	18.114/18.178/17.963
		I5	/ 0.231 / 5.425	/ 0.817 / 7.153	/ 0.094 / 5.892	/ 0.031 / 1.120
	是	I1	0.178 / 0.177 / 0.378	0.173 / 0.196 / 0.360	0.179 / 0.178 / 0.370	0.174 / 0.179 / 0.278
		I2	0.217 / 0.218 / 0.752	0.212 / 0.290 / 0.723	0.219 / 0.218 / 0.733	0.213 / 0.220 / 0.580
		I3	32.866/36.505/43.208	26.209/29.844/37.529	44.082/44.350/49.290	5.382 / 5.421 / 7.720
		I4	18.057/18.187/21.406	17.900/18.655/22.735	18.410/18.549/22.912	18.121/18.271/18.087
		I5	/ 0.219 / 5.241	/ 0.784 / 6.483	/ 0.113 / 6.251	/ 0.034 / 1.147
难度服从[-3,3]上的均匀分布,区分度服从[0.1,2.3]上的均匀分布	否	I1	0.168 / 0.174 / 0.255	0.170 / 0.170 / 0.259	0.167 / 0.171 / 0.256	0.168 / 0.174 / 0.254
		I2	0.205 / 0.212 / 0.496	0.205 / 0.209 / 0.502	0.202 / 0.210 / 0.498	0.205 / 0.213 / 0.492
		I3	35.760/39.329/40.369	31.784/34.759/36.421	58.735/58.417/57.954	12.048/11.781/11.917
		I4	19.152/19.047/19.092	19.285/19.248/19.321	17.563/17.806/17.936	18.158/18.253/18.377
		I5	/ 0.077 / 0.517	/ 0.038 / 0.597	/ 0.038 / 0.586	/ 0.057 / 0.587
	是	I1	0.173 / 0.173 / 0.264	0.169 / 0.172 / 0.255	0.171 / 0.173 / 0.253	0.169 / 0.173 / 0.252
		I2	0.209 / 0.212 / 0.524	0.206 / 0.210 / 0.493	0.209 / 0.212 / 0.488	0.206 / 0.213 / 0.481
		I3	32.269/36.059/37.235	29.889/32.810/34.554	53.243/53.032/52.840	6.646 / 6.449 / 6.836
		I4	18.978/18.873/18.908	19.402/19.362/19.436	17.606/17.866/17.986	18.160/18.244/18.298
		I5	/ 0.077 / 0.550	/ 0.041 / 0.578	/ 0.043 / 0.558	/ 0.053 / 0.554

为描述方便,将返真性记为 I1,能力估计标准差记为 I2,项目调用次数标准差记为 I3,人均用题数记为 I4,人均迭代失败次数记为 I5。

表中数据 A/B/C 表示,A 是由 EAP 估计得到,B 是由 F-S 估计得到,C 是由 N-R 估计得到。EAP 方法不需迭代,所以 I5 相应数据为空。

附表 2　EAP 估计、$b \sim N(0,1)$, $lna \sim N(0,1)$ 时 各选题策略的统一量纲加权求和结果

是否使用影子题库	评价指标	平均数法	中位数法	去两端平均法	等级难度匹配法
否	返真性	0.994083	0.988235	1	0.994083
	能力估计标准差	0.990338	0.985577	1	0.990338
	项目调用次数标准差	0.297663	0.336834	0.242397	1
	人均用题数	0.985448	0.917175	0.9544	1
	加权求和结果	3.267532	3.227821	3.196798	3.984421
是	返真性	1	0.988235	0.982456	0.994083
	能力估计标准差	1	0.985577	0.980861	0.990338
	项目调用次数标准差	0.154065	0.174896	0.130412	1
	人均用题数	0.985104	0.915614	0.951509	1
	加权求和结果	3.139169	3.064322	3.045238	3.984421

Item Selection Strategies of Computerized Adaptive Testing based on Graded Response Model

Chen Ping[1], Ding Shuliang[1], Lin Haijing[1,2], Zhou Jie[1]

([1]*Computer Information Engineering College, Jiangxi Normal University, Nanchang* 330027, *China*)

([2]*Jiangxi GongMao Vocational College, Nanchang* 330100, *China*)

Abstract

Item selection strategy (ISS) is an important component of Computerized Adaptive Testing (CAT). Its performance directly affects the security, efficiency and precision of the test. Thus, ISS becomes one of the central issues in CATs based on the Graded Response Model (GRM). It is well known that the goal of IIS is to administer the next unused item remaining in the item bank that best fits the examinee's current ability estimate. In dichotomous IRT models, every item has only one difficulty parameter and the item whose difficulty matches the examinee's current ability estimate is considered to be the best fitting item. However, in GRM, each item has more than two ordered categories and has no single value to represent the item difficulty. Consequently, some researchers have used to employ the average or the median difficulty value across categories as the difficulty estimate for the item. Using the average value and the median value in effect introduced two corresponding ISSs.

In this study, we used computer simulation compare four ISSs based on GRM. We also discussed the effect of "shadow pool" on the uniformity of pool usage as well as the influence of different item parameter distributions and different ability estimation methods on the evaluation criteria of CAT. In the simulation process, Monte Carlo method was adopted to simulate the entire CAT process; 1000 examinees drawn from standard normal distribution and four 1000 – sized item pools of different item parameter distributions were also simulated. The assumption of the simulation is that a polytomous item is comprised of six ordered categories. In addition, ability estimates were derived using two methods. They were expected a posteriori Bayesian (EAP) and maximum likelihood estimation (MLE). In MLE, the Newton – Raphson iteration method and the Fisher Score iteration method were employed, respectively, to solve the likelihood equation. Moreover, the CAT process was simulated with each examinee 30 times to eliminate random error. The IISs were evaluated by four indices usually used in CAT from four aspects – – the accuracy of ability estimation, the stability of IIS, the usage of item pool, and the test efficiency. Simulation results showed adequate evaluation of the ISS that matched the estimate of an examinee's current trait level with the difficulty values across categories. Setting "shadow pool" in ISS was able to improve the uniformity of pool utilization. Finally, different distributions of the item parameter and different ability estimation methods affected the evaluation indices of CAT.

Key words　graded response model, computerized adaptive testing, item selection strategy, shadow pool.