心理学报 2012, Vol. 44, No.3, 400−412
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2012.00400

多级评分计算机化自适应测验动态综合选题策略*

罗 芬 丁树良 王晓庆

(江西师范大学计算机信息工程学院, 南昌 330022)

摘 要 多级评分可以提供更多关于被试的信息, 是计算机化自适应测验的一个发展方向, 选题策略是计算机化自适应测验的研究重点。对于多级评分的等级反应模型, 本文拟用区间估计的思想改进近期提出的几种选题策略, 并且将两级评分 *b-STR* 和 *a-STR* 推广到多级评分以改进最大信息量选题策略。*Monte Carlo* 模拟实验表明在达到或接近原有选题策略测验精度的基础上, 本文提出的几种新选题策略有的能够有效降低测验长度, 有的可以极大降低项目曝光率。

关键词 等级反应模型; 计算机化自适应测验; 选题策略; 区间估计; 多级评分 *b-STR*

分类号 B841

1 研究背景

计算机化自适应测验(Computerized Adaptive Testing, *CAT*)针对不同考生的特质水平因人施考, 可以更客观、更准确地测出考生的真实能力水平(漆书青, 戴海琦, 丁树良, 2002)。然而有人在 *Science* 发表文章(Quellmalz & Pellegrino, 2009), 主张对于高风险(high stake)测验, 还是需要慎用 *CAT* 这种形式。既然 *CAT* 有不少优点, 为什么在高风险测验中必须慎用呢? 原因在于 *CAT* 的安全性可能会受到威胁, 特别是对于频繁使用的质量比较好的项目, 常常会因为被考生“分享”相应的信息而失密。因此目前对 *CAT* 的最重要的组成部分—选题策略的研究, 往往关注提高能力(或者潜在特质, 甚至是潜在特质向量)估计精度以及加强考试安全性(Barrada, Olea, Ponsoda, & Abad, 2010)这两方面。加强 *CAT* 安全性的一个重要手段是使项目曝光度更均匀, 以充分利用题库中所有项目, 降低题库建设的成本(程小扬, 丁树良, 颜深海, 朱隆尹, 2011)。

传统的以最大 *Fisher* 信息量(*MFI*, 以下将 *Fisher* 信息量简称为信息量)为代表的选题策略, 是在被试能力估计基础上从题库中该被试还未作答的项目里(下文称这些项目为该被试的剩余题库)选择具有最大信息量的项目。由于项目信息量与区分度平方成正比, 因此高区分度的项目更容易被选择, 导致其曝光率过高, 从而降低测验的安全性(Chang & Ying,1996, 1999); 同时, 题库项目的应用不均衡也将导致低区分度的项目不能被充分利用, 造成低区分度项目的巨大浪费。

针对 *CAT* 中传统的选题策略的不足之处, 很多学者致力于研究新的选题策略, 希望新的选题策略既不降低测验效率, 又能解决传统选题策略所遇到的种种问题。按区分度 *a* 分层选题策略(*a*-stratified, *a-STR*) (Chang & Ying, 1996, 1999)和综合区分度 *a* 和难度 *b* 分层的选题策略(*a*-stratified with *b* blocking, *b-STR*) (Chang, Qian, & Ying, 2001)部分地解决了传统选题策略特别是两级评分 *CAT* 选题策略所面临的难题。传统 *MFI* 选题策略, 从与当前能力估计值相匹配的项目中, 选用高区分度项目。因此是一种“降 *a*”方法。而 Chang 和 Ying (1999)提出的 *a-STR* 方法是一种“升 *a*”方法(Cheng, Chang, Douglas, & Guo, 2009)。所谓“升 *a*”方法, 是指对每个被试进行 *CAT* 时, 随着测验进程深入, 所选项目的区分度更大, 即越到能力估计的精确阶段越使用

收稿日期: 2010−10−21

* 国家自然科学基金(30860084, 60263005, 31160203,31100756), 国家教育科学规划项目(CCA110109), 教育部人文社科项目(09JJCXLX012, 10YJCXLX049), 江西省教育厅科技计划项目(GJJ11385). 全国教育考试科研规划课题(2009JKS2009)。

通讯作者: 丁树良, E-mail：ding06026@163.com

高质量的题目。然而 *a-STR* 只是部分实现这一目标, 在不同的层之间, *a* 随层数的变大而加大, 但在同一层内所选的多个项目, *a-STR* 就不能保证同一层内这些项目也能逐题按“升 *a*”方法选题(Cheng et al., 2009), 我们考虑, 为克服 *a-STR* 的这一不足, 是否可以设计一个 *a* 的函数, 使该函数自动按 *a* 分层且在层内区分度还能随着测验精度的提高而提高。如果在该函数中综合信息函数、等级难度/步骤参数以及能力估计的“距离函数”, 可否改进 *MFI*? Chang, Qian 和 Ying (2001)提出 *b-STR* 策略 。其核心思想是强制 *a-STR* 每一层中项目难度 *b* 的分布与整个题库中 *b* 的分布相似, 以保证不同被试的能力可以更好地与项目难度匹配, 他们认为更重要的是要让难度值 *b* 与能力值 θ 相匹配。问题是多级评分项目中有多个 *b*, *b-STR* 又该如何设计?

多级评分反应模型是当前测验发展的新方向之一(Meijer & Nering, 1999)。美国著名测验机构 *ETS* 也正在大力推进对等级计分题的研究与应用, 反对盲目单纯推崇选择题。多级评分模式的特点是每个项目有多个等级难度(步骤参数) (Dodd, De Ayala, & Koch, 1995), 在给定要求之下, 如何寻找到合适的项目提供给被试是多级评分模型下 *CAT* 研究的重点。李铭勇等人(李铭勇, 张敏强, 简小珠, 2010)报道, “在国际上, 有关等级反应模型的计算机自适应测验安全问题的研究相对较少, …, 罕有国外研究者对此问题作出研究”; 相反近年来国内对于有关计算机自适应测验中多级评分模型的选题策略的研究还比较活跃(陈平, 丁树良, 林海菁, 周婕, 2006; 戴海崎, 陈德枝, 丁树良, 邓太萍, 2006; 刘珍, 丁树良, 林海菁, 2008; 罗照盛, 欧阳雪莲, 漆书青, 戴海琦, 丁树良, 2008)。

国外近年来有一些基于 *Samejima* (1969)等级反应模型(*GRM*)以及基于分部评分模型(*PCM*)的多级评分 *CAT* 的研究结果。比如, 基于健康相关的生活质量(*HRQOL*)问卷的数据拟合 *GRM* (Choi & Swartz, 2009)。基于 *GRM*, 由于可以使患者减少作答的负担和及时报告评分结果, 患者自陈结果报告(*patient-reported outcomes, PROs*)使用 *CAT* 形式。Choi 和 Swartz (2009)比较了几种选题策略, 这些选题策略包括 *MFI*、最大似然加权信息(*MLWI*)、最大后验加权信息(*MPWI*)、最大期望信息(*MEI*)、最小期望后验方差(*MEPV*)和最大期望后验加权信息(*MEPWI*), 他们还选择了最差的随机选题策略和最好的 *MFI* 选题策略的选题结果作为对照。因为患者自陈结果一般不是高风险的, 所以该文章只是比较已有选题策略在能力估计上的表现, 以考察不同研究者所获得的相互冲突的观察结果, 而没有考虑内容平衡和曝光度控制。他们得出以下几点值得注意的结果: 贝叶斯方法在多级评分 *CAT* 中并没有像在 0-1 评分中那样比其他方法表现更好; 复杂的选题策略表现并不比简单的更好; 区分度的影响比项目难度影响更大。这些结果启发我们在设计基于 *GRM* 的选题策略方案时, 可以不使用贝叶斯方法, 也不必考虑太复杂的方案, 同时必须注重区分度的影响。

传统的 *MFI* 是依据当前能力估计值挑选合适的项目, 然而当前能力估计值并不一定是被试的真实能力值(van der Linden, 1998), Penfield (2006)注意到这个事实, 并结合贝叶斯方法提出最大后验加权信息(*MPWI*)和最大期望信息(*MEI*)两种选题策略。但是这些方法必须考虑被试群组能力水平的分布和题库结构的分布, 而且 Penfield 给出的方法没有涉及到如何提高 *CAT* 的安全性, 也没有涉及到区分度对 *CAT* 的影响等重要问题(因为他讨论的是所有项目区分度相等的分部评分模型 *PCM*)。鉴于本文主要目的是试图寻找既能保持能力估计准确性又能提高测验安全性的 *CAT* 选题策略, 故下文没有考虑 *MPWI* 和 *MEI*。

国内不少研究者希望将 0-1 评分 *CAT* 选题策略的最新成果应用到多级评分 *CAT*, 这里首先要解决多级评分项目有多个难度/阈值参数的问题, 一般通过降维(reduction of dimensionality)的方法, 将多级记分模型中的多个难度等级或多个步骤参数用一个综合指标来概括。通常有以下两种具体方式: 其一, 使用某个等级难度/步骤参数来表示所有等级难度/步骤参数; 其二, 使用平均数/中位数来表示所有/部分等级难度或步骤参数(陈平 等, 2006; 戴海崎 等, 2006; 刘珍 等, 2008)。然而罗照盛等人(2008)发现这种综合指标有一定的缺陷, 即项目参数降维法损失了不少信息, 进而罗照盛等人(2008)提出了一种邻近等级匹配法, 这一方法考虑了多个等级难度或步骤参数, 但是仍有一些问题有待解决: 如何表示邻近等级难度? 取多少个邻近等级? 如果不取邻近等级难度, 而综合项目所有参数, 效果是否更好? 以什么为工具/载体对所有等级难度/步骤参数进行综合? 如何利用这个综合指标对项目进行挑选? 而且, 以上并没有像 Penfield (2006)那样考虑到 *CAT* 是对能力进行序贯估计(*sequential*

estimate), 即能力估计精度是随作答反应项目数的增加而提高的; 如何将这种能力估计精度的动态变化反映到选题策略之中? 陈平等人(2006)使用的“影子题库”是降低项目曝光率的好办法, 但是估计精度是不断变化的, “影子题库”中所含题量固定是否合适? 对这些问题的反思, 又启发我们进一步考虑可否设计一种选题策略, 使之成为能力估计精度的函数, 同时又兼有影子题库的优点? 另外, 多级评分中每个项目有多个等级难度, 能否将 *b-STR* 的思想进行改造运用于多级评分模式?

对于 *GRM*, 戴海琦等人(2006)比较了五种选题策略: 难度等级平均数与能力匹配策略、难度等级中位数与能力匹配策略、信息量最大选题策略以及两种不同划分的 *a* 分层选题策略。陈平等人(2006)提出去两端平均法和等级难度匹配法两种选题策略, 即选择与当前能力值最接近的难度等级所属的项目, 并增设了“影子题库”控制题目的曝光率(即选择与当前能力匹配的 5 个项目, 再随机选一个项目提供给被试)。*CAT* 测验通常采用三个评价指标, 即测验效率、测验长度、题库的平均曝光率。我们通过模拟研究发现以上各方法在测验效率上表现较好, 但对于项目的曝光率和测验长度这两个指标则还不能令人满意。

统计中有时可以用区间估计的长度来表达估计的精度, 而相应的区间内的点与“真值”的距离大小的概率(如果可以计算的话)可以由区间的长度来限制。如果区间估计中只有若干个点满足实际要求, 那么这里区间估计中的“点的集合”可以想象为“影子题库”, 而区间的长短不仅可以对应估计的精度, 而且可能对应集合中点的多少(如果区间中的点是有限的); 同时, 区间估计一般是样本的函数, 它是动态变化的, 可否利用这一特点使得选题策略也具有动态性是值得注意的。项目反应理论(*IRT*)中的信息量可以用来刻画能力估计的精度, 且它又综合了所有项目参数, 所以我们考虑新的选题策略能否借用“测验”信息量的估计值 $\sum_{j=1}^{m_\alpha} I_j(\hat{\theta}_\alpha)$ 来表达, 其中 $\hat{\theta}_\alpha$ 是被试 α 能力估计值, m_α 是 α 当前所做的项目数, 它是一个可以变化的数。

对于基于 *GRM* 的 *CAT*, 本文针对上述一些近期所发表的选题策略的不足加以改进, 另外结合国外文献(Choi & Swartz, 2009)提出的观点, 提出一些新的选题策略。这些新的选题策略依据两类想法, 一类欲利用区间估计的思想扩大选题的范围, 而不失测验的效率; 另一类借助于 *a-STR* 和 *b-STR* 思想, 采用自动动态分层方法, 通过少量降低测验效率来换取项目曝光率和测验长度方面的较大改进。为了验证新选题策略的优劣, 我们以传统的选题策略(*MFI* 及随机选题)作为对照, 并且将几种近期发表的选题策略以及对它们进行改进后的选题策略共 9 种, 在同等条件下做模拟实验, 进行比较。

2 基于 *GRM* 模型的 9 种选题策略

我们将上述提到的选题策略分成两类, 第一类包括测验效率最高的 *MFI* 和曝光率最低的随机选题策略, 用以和其他策略相比较; 第二类包括陈平等人(2006)及戴海琦等人(2006)的已有研究成果, 以及我们在其基础上结合区间估计思想导出的新选题策略; 还包括我们在 *MFI* 与动态 *a-STR* 和动态 *b-STR* 相结合基础上导出的另一种新选题策略。

为了方便描述, 用 Mid_b_j 表示项目 *j* 难度的中位数, Avg_b_j 表示项目 *j* 难度的平均值, $Avg1\text{-}b_j$ 表示剔除项目 j 的最大和最小难度后的平均值。由于 *CAT* 中每一个被试所做的项目是不相同的, 下文中, 某个被试的剩余题库是指题库中该被试尚未反应的项目的集合。

2.1 传统的选题策略

最大信息量选题法(*MFI*): 在当前能力估计值下, 从剩余题库中挑选信息量最大的项目。这种选题策略的测验效率最高, 可用作为精度和效率的参照标准。

随机选题(*RAN*): 从剩余题库随机选择项目使得在题库中所有项目的曝光率近似相等, 该方法的优点是项目使用均匀性最好, 可用作项目使用均匀性的参照标准。

2.2 戴海琦等及陈平等选题策略的研究成果

在 *GRM* 中, 每个项目有多个难度等级, 各个等级区分度相同。用每个项目等级难度的中位数作为整个项目的难度值, 若 f_j 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:

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

$$Avg\_b_j = \frac{1}{f_j}\sum_{t=1}^{f_j} b_{jt}$$

$Avg1\_b_j = \frac{1}{f_j-2}\sum_{t=2}^{f_j-1} b_{jt}$, 其中 b_{jt} 是指第 *j* 个项目除去最大难度值和最小难度值后剩余的难度

值(由于 *GRM* 中项目难度从小到大排列, 所以求和可以从 2 开始, 到 *fj*-1 结束)。

戴海琦等(2006)及陈平等人(2006)选题策略实质是难度降维选题法(以下统称难度降维选题法, *RDSI*), 这些新的选题策略包括:

中位数匹配法: 从剩余题库挑选满足极小化 $\left|Mid\_b_j-\theta\right|$ 的项目(简记为 *mid*)

平均难度匹配法: 从剩余题库挑选满足极小化 $\left|Avg\_b_j-\theta\right|$ 的项目(简记为 *Avg*)

去两端后平均难度匹配法: 从剩余题库挑选满足极小化 $\left|Avg1\_b_j-\theta\right|$ 的项目(简记为 *Avg*1), 由于在陈平等(2006)的文章中指出, 在各个指标上 *Avg*1 方法的实验结果均好于 *Avg* 方法的实验结果, 故本文不讨论 *Avg* 方法。

等级难度匹配法: 从剩余题库挑选极小化 $\left|b_{jt}-\theta\right|$ 的项目, 用题库中每个项目的每个等级难度去匹配当前的能力估计值, 挑选等级难度与当前能力估计值最接近的项目。

2.3　区间估计选题法和动态综合选题法

Chang 等人(2001)设计的 *b-STR* 选题策略只适合两级评分的情形。上文提到的 *b-STR* 的基本思想对新的选题策略的制定有启发作用, *b-STR* 的目标是使不同被试的能力与项目难度相匹配, 这一点十分重要。我们考虑 *b-STR* 的想法可否扩展到有多个难度的 *GRM-CAT*? 如何使得能力值与项目难度向量相匹配? 注意到 *b-STR* 中, 分层数及分块数是固定的, 这种固定方式和 *CAT* 的不定长序贯估计方式多少有点不协调, 所以这里提出动态化 *b-STR*, 即自动地按 *b* 分层, 使得对于每个被试的每一次能力估计值都从剩余题库中选出若干个比较适合被试作答的项目, 而所谓的"比较适合被试作答的项目"的挑选原则是根据统计中区间估计的思想提出的。这是一种新的选题准则, 它包括以下两种选题方法: 区间估计选题法和动态综合选题法。

2.3.1　区间估计选题法　我们先借用降维的思想, 对项目难度向量进行降维, 以解决 *GRM* 中"难度向量与能力相匹配"的难题, 其次我们希望在 *CAT* 初始阶段(即能力估计比较粗糙阶段)在比较宽的范围内选取项目; 而 *CAT* 越接近结束, 能力估计越准确, 越应该在比较窄的范围内选项目。由于测验的信息量随着测验长度的增加而增加, 这也表示对能力测量精度的提高, 于是我们选取测验的信息量 $I(\hat{\theta}_\alpha)$ 平方根的倒数(记为 ε)来度量测量的精度, 即: $\varepsilon=1/\sqrt{I(\hat{\theta}_\alpha)}$, 其中 $I(\hat{\theta}_\alpha)=\sum_{j=1}^{m_\alpha}I_j(\hat{\theta}_\alpha)$。用 B_j 表示项目 j 的难度向量, 而 $h(B_j)$为 B_j 的一个函数, 它是一个 1 维的量。令

$$S_h=\{j:\left|h(B_j)-\theta\right|<\varepsilon,\ j\text{ 是剩余题库中一个项目}\} \tag{1}$$

S_h 是一个项目集合, 它是剩余题库的一个子集, 且用 $\left|h(B_j)-\theta\right|<\varepsilon$ 表示 S_h 中项目的难度向量与能力 θ 的接近程度。显然, 对于固定的 θ, ε 越大, $h(B_j)$可能变化的范围越大; 而 ε 越小, $h(B_j)$可能变化的范围越小。由 ε 的定义, 知对能力估计越准确($I(\theta)$越大), 越要求 $h(B_j)$接近 θ。这就"动态"地实现了 *b-STR* 的目的, 即越到能力 θ 估计精确阶段, 越要求所选项目中的难度"向量"与 θ 相匹配。由于项目的信息量是对项目参数 a_j、B_j 及能力 θ 的一个综合, 故这个新的选题策略(1)表面上是对 B_j 进行了降维($h(B_j)$), 但实质上还进行了综合考虑。由于(1)中括号内的绝对值不等式和数理统计中区间估计的形式很相像, 且一般来说满足(1)的项目一般不止一个, 故这个选题策略又和"影子题库"的做法相似。

令 $h(B_j)$=Mid_b_j, 则对应着中位数选题策略的改进, 记为 *mid*-ε, 即

$$S_{Mid}=\{j:\left|Mid\_b_j-\theta\right|<\varepsilon\}$$

令 $h(B_j)$=*Avg*1_b_j, 则对应着删除 B_j 中最大及最小分量后的平均值选题策略的改进, 记为 *Avg*1-ε, 即

$$S_{Avg1}=\{j:\left|Avg1\_b_j-\theta\right|<\varepsilon\}$$

令 $h(B_j)=b_{jt}$, b_{jt} 为 B_j 的一个分量, 则

$S_b=\{(j,t):\left|b_{jt}-\theta\right|<\varepsilon\}$ 是等级难度匹配法的改进, 记为 b_{jt}-ε。

这里, *mid*-ε, *Avg*1-ε, b_{jt}-ε 三种方法合称为区间估计选题法(记为 *IESI*)。

2.3.2　动态综合选题法

刻画项目难度向量与能力接近的另一种形式是

$$\prod_{t=1}^{f_j}\left|b_{jt}-\theta\right| \tag{2}$$

由于不同项目的等级数 f_j 可能不同, 为了使得不同项目之间这种接近程度可以比较, 可以将上式求几何平均, 即考察指标

$$\sqrt[f_j]{\prod_{t=1}^{f_j}\left|b_{jt}-\theta\right|} \tag{3}$$

对每一个被试欲使项目区分度小的项目在实施 *CAT* 的较早阶段使用, 而区分度较大的项目在较晚阶段使用, 故考察另一个函数

$$a_j^{m(k)} \tag{4}$$

这里 k=1, 2, …, K。K 为欲考察的阶段数, 相当于 *a-STR* 中的分层数, 比如 K=4。本文中取 m (1)=2,

m (2)=3/2, m (3)=1/2, m (4)=0。这个函数与信息量相结合：

$$a_j^{m(k)} / I(\hat{\theta}_\alpha) \quad (5)$$

再与上述难度的几何平均相结合，成为

$$a_j^{m(k)} \sqrt[f_j]{\prod_{t=1}^{f_j} \left| b_{jt} - \theta \right|} \Big/ I(\hat{\theta}_\alpha) \quad (6)$$

选取项目 j，使上式达到最小，则当 *CAT* 的实施早期阶段，$I(\hat{\theta}_\alpha)$ 较小，而 m (k)值较大，比如 k=1, m (1)=2，如果项目难度与能力 θ 同样接近条件下，则由于(5)的调节作用，就不至于象 *MFI* 那样，会选择区分度大的项目，而是可能“降低”区分度的作用，但在实施 *CAT* 的后阶段，比如 k=4, m (k)=0,则(6)中只有信息量和 b 起作用，表示要取项目难度参数最“接近”能力且信息量最大的项目。我们设定当被试累积信息量达到 M/30(其中 M 为 *CAT* 中预定信息量)时，m (k)=2；被试累积信息量达到 5*M/30 时，m (k)=3/2；被试累积信息量达到 14*M/30 时，m (k)=1/2；被试累积信息量大于 14*M/30 时，m (k)=0。当然 m (k)值的变化还是一个经验取值方法(戴海琦等, 2006)。当被试累积信息量大于 14*M/30, m (k)=0，这个阶段实际上与 *MFI* 的原理很接近了。

由于(6)中区分度的选取原则是动态的，故这一选题策略称为动态综合选题法(记为 *DC*)。构造这个选题策略的目的是为了改进 *MFI*，即在保持能力估计精度的基础上降低项目曝光率。

3 实验步骤

实验采用 *Monte Carlo* 模拟方法，比较在同等条件下各种选题策略的表现。

3.1 模拟产生被试能力和题库

用 N (p, q)表示平均值为 p，方差为 q 的正态分布，U (p, q)表示在[p, q]区间上的均匀分布。

我们模拟了 4 个服从不同分布的题库(陈平等, 2006)，每个题库含有 1000 个项目，每个项目的难度等级数从{3, 4, 5, 6}中随机选取。题库参数(区分度参数 a，难度参数 b)的分布如下：

第一种题库 $b \sim N(0,1), \ln a \sim N(0,1)$；第二种题库 $b \sim U(-3,3), \ln a \sim N(0,1)$；第三种题库 $b \sim N(0,1)$, $a \sim U(0.4, 2.5)$；第四种题库 $b \sim U(-3,3)$, $a \sim U(0.4, 2.5)$；并且在各种分布条件下，限定 a 的取值范围为[0.4, 2.5]。

我们分别模拟了 1000 个能力服从 N (0, 1)的被试群体(称为群体 1)和 19 个不同能力点，每个能力点有 1000 人的被试群体(称为群体 2)，这些被试群体分别参与对应不同题库的 *CAT* 的测试过程。

结束条件：测验的信息量达到预定值 M (设 M=16)或达到最大测验长度(30 题)。

3.2 能力的估计方法

设能力的先验分布为标准正态分布，采用贝叶斯期望后验估计方法估计能力，$\hat{\theta}_\alpha = \frac{\sum_{k=1}^{q} X_k L(X_k) A(X_k)}{\sum_{k=1}^{q} L(X_k) A(X_k)}$，其中 q 为节点数(q=2[$\sqrt{m}$], m 为测验长度(漆书青等人, 2002)，对多级评分项目，$q = 2\left[\sqrt{\sum_{j=1}^{m} f_j}\right]$，且 $q \leqslant 31$)。式中[x]为对 x 取整函数。

$$X_k = -3 + 6 * \frac{i-1}{q-1} (i=0, 1, \ldots q\text{-}1),$$

$$A(X_k) = \frac{8}{q} * e^{-X_k^2} / \sqrt{2\pi},$$

$$L(X_k) = \prod_{j=1}^{m} \prod_{t=0}^{f_j} P_{\alpha jt}^{u_{\alpha jt}}$$

其中 f_j 是第 j 个项目的满分值，$P_{\alpha jt}$ 是第 α 个被试在第 j 个项目恰得 t 分的概率，$u_{\alpha jt}$=1 是第 α 个被试在第 j 个项目得 t 分，否则 $u_{\alpha jt}$=0。

3.3 CAT 的施测过程

CAT 的施测过程，大体分为两个阶段：一是试验性探查阶段，对两级评分项目从题库中随机抽取若干个项目施测，直到被试既有答对又有答错时探查阶段停止；而对于多级评分项目，要使得被试既不全部得 0 分，也不全部得满分；二是精确估计真值阶段，运用某种选题策略从题库中挑选合适的项目提供给被试。在测试过程中，每获得一个项目的反应要重新计算被试的能力，并依据新的能力估计值计算累积信息量。

对于每个题库结构，我们采用同一批被试共计 1000 人参加测试，每个被试分别采用 9 种选题策略记录实测过程。为了保证模拟实验结果可比较，同一被试在试验性探查阶段所选择的项目都是一致的，然后对同一被试分别运用 9 种选题策略施测考察五个评价指标的变化。某个被试在指定的选题策略的模拟测试过程请参见漆书青等人(2002)一书。所有 1000 个被试经过这样的测试以后，分别计算 9 种选题策略下如下一节中所示的五个评价指标值。

4 模拟结果分析

4.1 五个评价指标

$$ABS = \sum_{i=1}^{N} \left(\left| \hat{\theta}_i - \theta_i \right| \right) / N$$

$$RMSE=\sqrt{\sum_{i=1}^{N}(\hat{\theta}_i-\theta_i)^2/N}$$

$$Bias=\sum_{i=1}^{N}(\hat{\theta}_i-\theta_i)/N$$

$$Nf=\sum_{i=1}^{N}n_i/N$$

$$\chi^2=\sum_{j=1}^{L}(r_j-\overline{r})^2/\overline{r},\ \ \overline{r}=\sum_{j=1}^{L}r_j/L$$

其中, N 表示被试人数, L 表示题库容量, $\theta_i,\hat{\theta}_i$ 分别表示被试 i 的能力真值与估计值, n_i 是被试 i 在某次测试中的作答项目数, r_j 表示在一次所有的被试群体参加的测验中项目 j 的使用次数。

本文中 *ABS* 是平均绝对偏差, *Bias* 是平均偏差, *Bias* 越接近零, 则表明该方法越接近无偏; *ABS* 和 *RMSE* 值越小, 说明估计的精度越高。*Nf* 用来评估测验效率, 值越小, 说明测验效率越高。χ^2 统计量 (Chang et al., 1996; 1999)用来评估题库项目的曝光率, 值越小, 说明曝光率越均匀, *CAT* 的安全性越好。

4.2　结果分析

将被试群体 1 和群体 2 分别参加模拟的 *CAT*, 比较 9 种选题策略在上述五个指标上的表现。为了便于描述, 沿用 2.2 和 2.3 节的定义, *RDSI* 是指 *mid* 策略、*Avg*1 策略、等级难度匹配法; *IESI* 是指 *mid*-ε 策略、*Avg*1-ε 策略、b_{jt}-ε 策略, 并且记 *SixS*=*RDSI* ∪ *IESI*,即集合 *SixS* 是这两个集合的并集合。

4.2.1　9 种不同选题策略在 *ABS*、*RMSE* 和 *Bias* 上的表现　表 1、表 2 为四种不同题库结构下, 群体 1 分别运用上述 9 种选题策略, 满足 *CAT* 终止规则时, 能力估计在 *ABS*、*RMSE* 和 *Bias* 这三个指标上的表现。从表 1 中我们可以看出, 在这两个指标上:

每种相同的题库结构下, *RDSI* 所代表的选题策略的表现基本相当;

RDSI 和 *IESI* 相对应的选题策略的表现都变化不大;

DC 策略的表现比集合 *SixS* 中六种选题策略稍差, 但比 *MFI* 策略好;

RDSI 在 $b\sim N(0,1)$题库时表现好于 $b\sim U(-3,3)$题库;

DC 策略在不同的题库结构下这两个指标的值变化不大, 但均比 *MFI* 要好。

从表 2 中我们可以看出, 各个方法基本上是无偏估计(平均值接近 0)。

表 1　能力服从 *N* (0, 1), 9 种不同选题策略的 *ABS* 和 *RMSE*

选题策略	$b\sim N(0,1)$, $lna\sim N(0,1)$		$b\sim U(-3,3)$, $lna\sim N(0,1)$		$b\sim N(0,1)$, $a\sim U(0.4,2.5)$		$b\sim U(-3,3)$, $a\sim U(0.4,2.5)$	
	ABS	*RMSE*	*ABS*	*RMSE*	*ABS*	*RMSE*	*ABS*	*RMSE*
mid	0.164	0.206	0.173	0.217	0.166	0.209	0.181	0.228
mid-ε	0.162	0.204	0.180	0.225	0.166	0.210	0.173	0.222
*Avg*1	0.163	0.210	0.182	0.229	0.167	0.211	0.175	0.219
*Avg*1-ε	0.162	0.206	0.178	0.222	0.168	0.212	0.179	0.226
等级难度匹配法	0.165	0.207	0.170	0.216	0.164	0.208	0.180	0.227
b_{jt}-ε	0.173	0.217	0.175	0.222	0.171	0.216	0.180	0.226
DC	0.174	0.227	0.177	0.224	0.174	0.222	0.175	0.222
MFI	0.183	0.230	0.179	0.232	0.187	0.234	0.189	0.240
RAN	0.171	0.219	0.178	0.225	0.174	0.223	0.181	0.225

表 2　能力服从 *N* (0, 1), 9 种不同选题策略的 *Bias*

选题策略	$b\sim N(0,1)$, $lna\sim N(0,1)$	$b\sim U(-3,3)$, $lna\sim N(0,1)$	$b\sim N(0,1)$, $a\sim U(0.4,2.5)$	$b\sim U(-3,3)$, $a\sim U(0.4,2.5)$
mid	0.002	0.006	0.002	−0.009
mid-ε	0.000	0.007	0.001	−0.002
*Avg*1	−0.005	0.006	0.011	0.004
*Avg*1-ε	0.008	0.004	0.000	0.009
等级难度匹配法	0.005	−0.003	0.006	−0.006
b_{jt}-ε	0.005	0.002	−0.003	−0.004
DC	0.011	0.008	0.008	0.015
MFI	−0.006	0.001	0.011	−0.001
RAN	0.004	0.008	0.018	−0.002

图 1 和图 2 是四种不同题库结构下，群体 2 分别运用上述 9 种选题策略，满足 *CAT* 终止规则时，能力估计的 *ABS* 和 *RMSE*。从图 1 和图 2 可以看出，结论与表 1 的结论基本一致。从图 1 和图 2 中，我们还可以看到，当题库结构满足 $b \sim N(0, 1)$ 时，9 种选题策略在 19 个能力点上这两个指标值呈抛物线状，即两端的能力点的指标值要大于中间的能力点；当题库结构满足 $b \sim U(-3, 3)$ 时，*SixS* 中所有的选题策略，随能力点值的升高，指标值呈下降的趋势，*DC* 的变化趋势和 *MFI* 基本相同，两端的能力点的指标值要高于中间的能力点，*DC* 在 19 个能力点上 *ABS* 值和 *RMSE* 的值基本都好于 *MFI*。

4.2.2 9 种不同选题策略在 *Nf* 上的表现

在四种不同题库结构下，群体 1 分别运用上述 9 种选题策略，满足 *CAT* 终止规则时，表 3 为平均的测验长度(*Nf*)，图 3 为在每个测验长度上累计人数占群体人数的百分比。

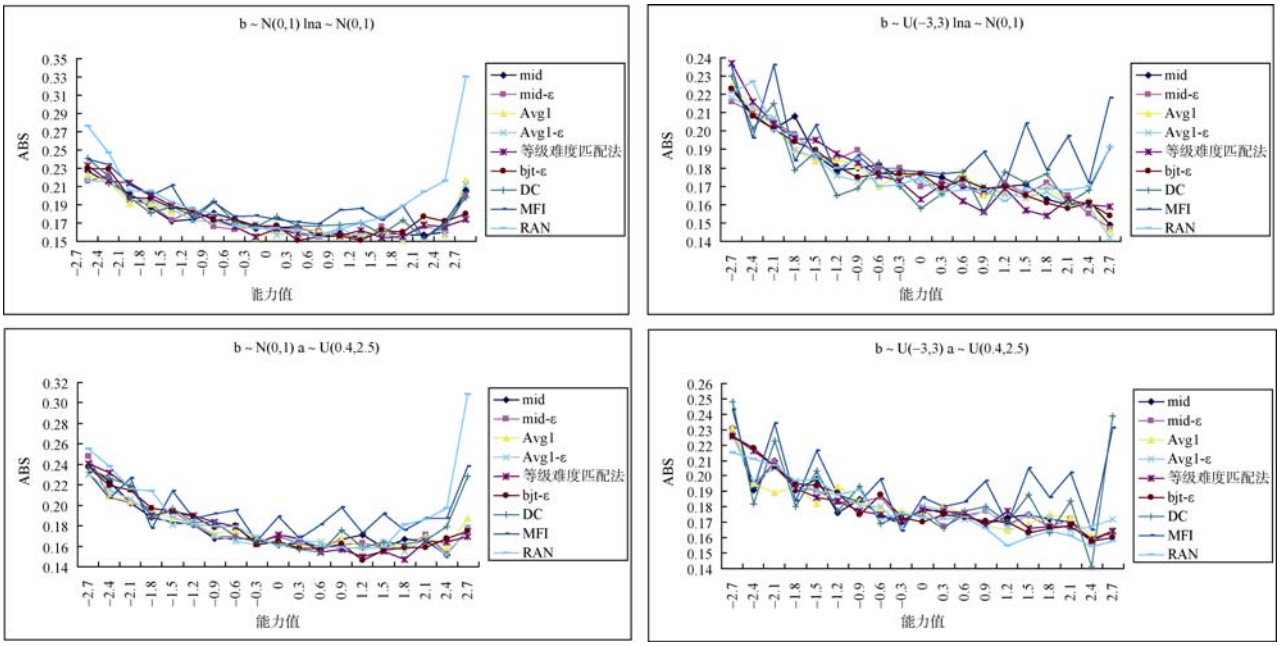

图 1 各种选题策略在 19 个能力点上 *ABS* 值

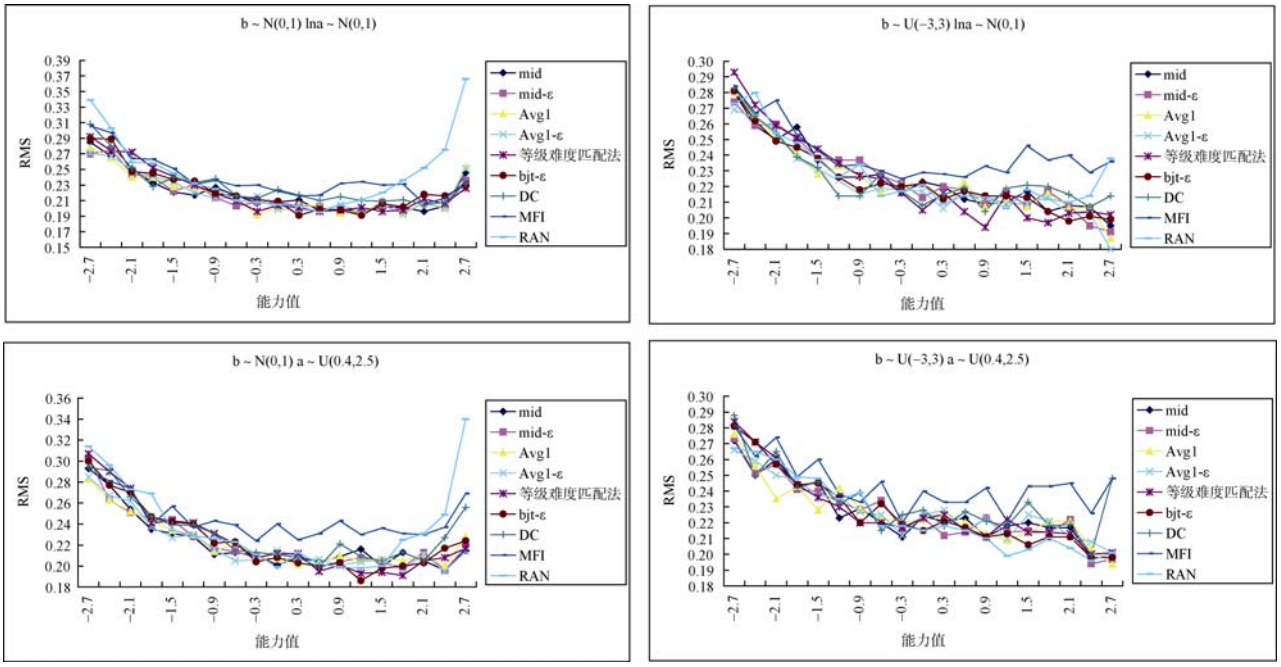

图 2 各种选题策略在 19 个能力点上 *RMSE* 值

表 3　能力服从 $N(0, 1)$, 9 种不同选题策略的 Nf

选题策略	$b \sim N(0, 1)$, $lna \sim N(0, 1)$	$b \sim U(-3, 3)$, $lna \sim N(0, 1)$	$b \sim N(0, 1)$, $a \sim U(0.4, 2.5)$	$b \sim U(-3, 3)$, $a \sim U(0.4, 2.5)$
mid	18.637	18.720	12.687	12.432
mid-ε	19.534	18.991	13.008	12.782
*Avg*1	20.346	19.054	13.190	12.995
*Avg*1-ε	20.072	19.644	13.160	13.216
等级难度匹配法	21.357	18.478	11.794	12.237
b_{jt}-ε	20.003	18.563	13.327	12.664
DC	9.608	9.443	8.743	8.433
MFI	7.034	6.735	6.403	6.397
RAN	22.782	23.103	16.214	17.271

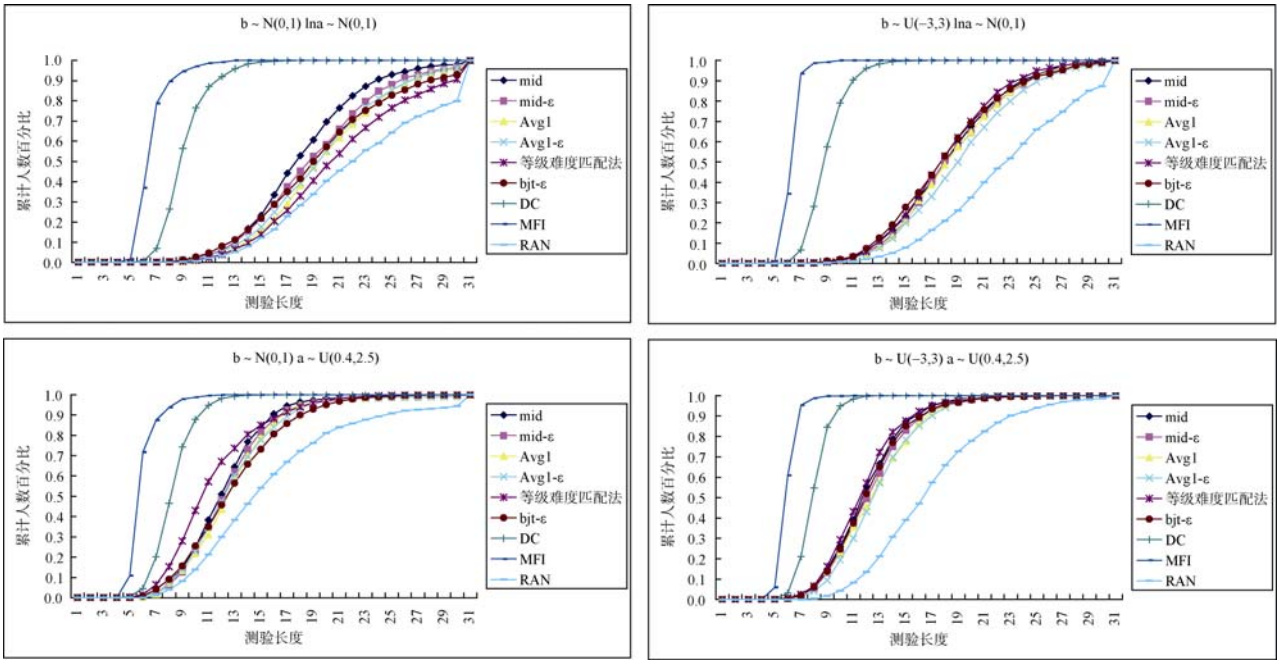

图 3　能力服从 $N(0, 1)$, 满足终止规则, 被试人数的累计百分比

从表 3 可以看出, *RDSI* 的策略中, 除第三种题库结构, *mid* 策略的 *Nf* 值要稍大于等级难度匹配法, 其他题库结构, *mid* 策略较 *RDSI* 中的其他策略, *Nf* 值均要小些或相当。*RDSI* 与 *IESI* 对应的策略相比, 除第三种题库结构, 等级难度匹配法比 b_{jt}-ε 策略 *Nf* 值要小些, 其他题库结构, 各个策略及改进策略的 *Nf* 值基本相当。*DC* 策略与集合 *SixS* 中的策略相比, *Nf* 值要小很多, 稍大于 *MFI* 策略, 且在不同的题库结构下, 测验长度的变化不大。*RDSI* 和 *IESI* 在满足 $a \sim U(0.4, 2.5)$的题库结构时 *Nf* 值小于满足 $lna \sim N(0, 1)$的题库结构。

从图 3 中, 我们同样可以看出在四种题库结构下, *DC* 较集合 *SixS* 在平均测验长度上均有较大的优势, 稍逊于 *MFI* 策略。集合 *SixS* 在满足终止规则时, *Nf* 值的表现差别不大。*RAN* 策略的表现最差。

图 4 为在第一种题库结构下, 群体 2 分别运用上述 9 种选题策略, 满足 *CAT* 终止规则时, 在每个测验长度上累计人数占群体人数的百分比。从图 4 中, 我们同样可以看出, *DC* 较集合 *SixS* 在平均测验长度上均有较大的优势, 稍逊于 *MFI* 策略。集合 *SixS* 在 $\theta = -1.8$、$\theta = -0.9$、$\theta = 0$ 这三种情况下, 平均测验长度的表现基本一致。在 $\theta = 0.9$, 等级难度匹配法要逊于集合 *SixS* 中的其他五种选题策略。在 $\theta = 1.8$, 等级难度匹配法、b_{jt}-ε 策略表现基本一致, 均逊于集合 *SixS* 中其他四种策略。而在其他题库结构下的实验结果我们同样发现 *DC* 稍逊于 *MFI* 策略, 但是较集合 *SixS* 在平均测验长度上均有较大的优势, 而集合 *SixS* 中所有策略满足终止规则时, 在每个测验长度上的表现差别不大。

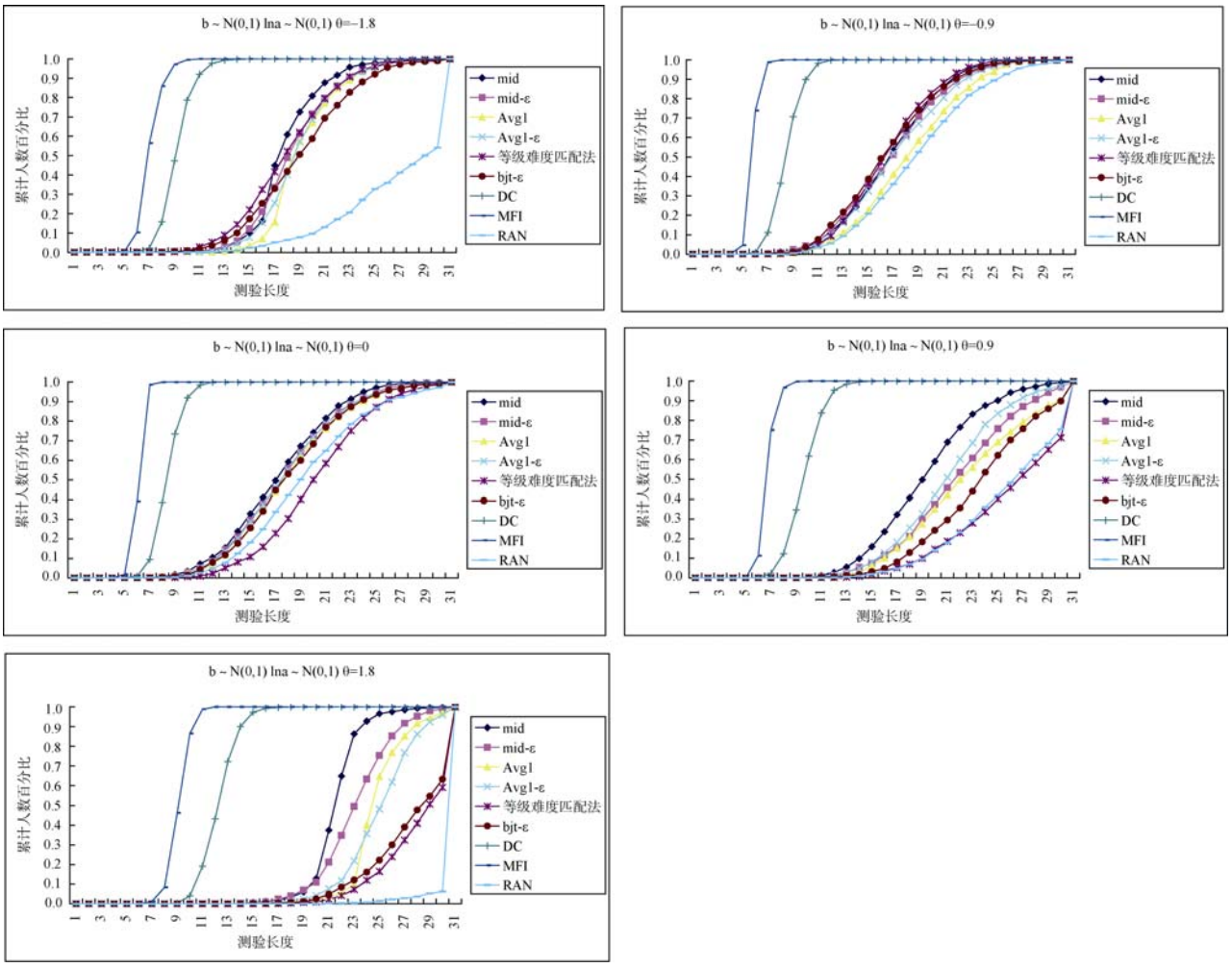

图 4 第一种题库结构, 5 个能力点各个测验长度的累计人数百分比

从图 4 还可以看出, 在能力为–1.8 和 1.8 时, *RAN* 选题策略比在其他能力点上需要更多的项目才能达到终止条件。

图 3, 图 4 同样表明 *RDSI* 和 *IESI* 在满足 $a \sim U$ (0.4, 2.5)的题库结构时 *Nf* 值小于满足 $lna \sim N$ (0,1) 的题库结构, 而 *DC* 策略在不同的题库结构下, 测验长度的变化不大。

4.2.3 9 种不同选题策略在χ^2上的表现

表 4 为四种不同题库结构下, 群体 1 分别运用上述 9 种选题策略, 满足 *CAT* 终止规则时χ^2值。从表 4 中我们可以看出:

DC 策略逊于 *mid* 策略和 *Avg*1 策略, 但好于 *MFI* 策略;

RDSI 在 $b \sim U$ (–3, 3), $a \sim U$ (0.4, 2.5)题库的曝光率较其他题库结构要好;

RDSI 和 *IESI* 相对应的选题策略相比, *RDSI* 中的选题策略逊于 *IESI* 中的选题策略, 特别是等级难度匹配法与 b_{jt}-ε 策略相比, b_{jt}-ε 策略表现很出色; *IESI* 中的选题策略中, b_{jt}-ε 策略与 *RAN* 的χ^2值很接近, 并且在不同题库结构下χ^2 值差异不大; 在难度服从均匀分布时, *Avg*1-ε 表现相当好, 此时比 b_{jt}-ε 还好。

图 5 为四种不同题库结构下, 群体 1 分别运用上述 9 种选题策略, 满足 *CAT* 终止规则时, 将题库中各个项目调用次数占总人数的百分比从低到高排序, 从 0%到 100%, 以 10%为步长递增, 以下称这些百分点为曝光点, 各个相邻曝光点区间累积曲线示意图。显然曝光率越均匀, 各曝光点的连线越接近一条直线, 否则就变成折线。

从图 5 中可以看出, 在 $b \sim N$ (0, 1), $a \sim U$ (0.4, 2.5)题库结构下, *MFI* 曝光率前 50%的累计项目个数不到题库容量的 40%, 其他 8 种选题策略前 50%的累计项目个数都达到 50% ~ 60%, 曝光率为 50%以后的累计项目个数, 各种选题策略的差别不大; 而其他题库结构, *MFI* 策略前 30%曝光率的累计项

目数不到题库容量的 20%, 其他选题策略前 30%曝光率的累计项目数均接近题库容量的 30% ~ 40%, 曝光率为 30%以后的累计项目个数, 各种选题策略的差别不大。

表 4　能力服从 N (0, 1), 9 种不同选题策略在 χ^2 上的表现

	$b \sim N$ (0, 1), $lna \sim N$ (0, 1)	$b \sim U$ (−3, 3), $lna \sim N$ (0, 1)	$b \sim N$ (0, 1), $a \sim U$ (0.4, 2.5)	$b \sim U$ (−3, 3), $a \sim U$ (0.4, 2.5)
mid	27.743	12.018	17.701	6.967
mid-ε	11.822	1.906	7.006	1.425
*Avg*1	29.284	8.290	16.598	5.592
*Avg*1-ε	18.573	1.467	10.910	1.296
等级难度匹配法	72.468	55.418	35.535	30.905
b_{jt}-ε	1.754	2.953	1.403	1.970
DC	55.697	75.072	35.734	43.076
MFI	167.999	203.041	133.976	116.374
RAN	1.037	1.027	0.989	0.956

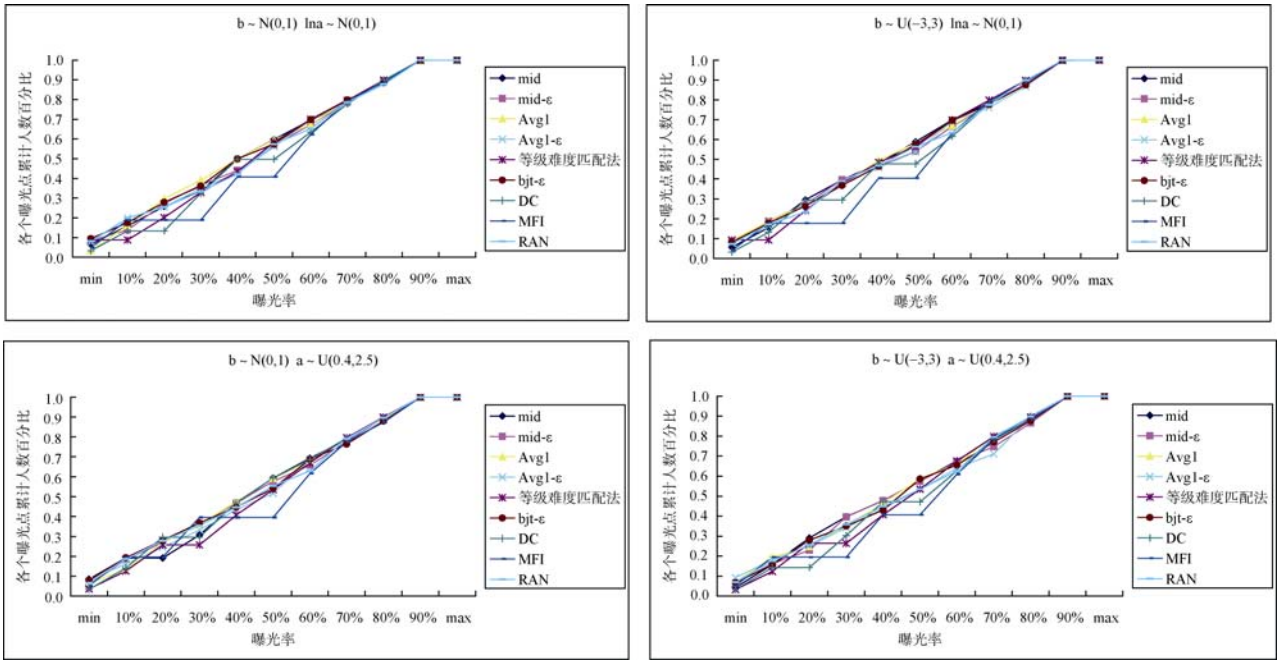

图 5　各个曝光点上累计项目个数百分比

5　结论与讨论

从 *Monte Carlo* 模拟实验结果中(包括剖面图)我们有如下发现:

由于能力估计方法是固定的, 所以所有选题策略对应的能力估计都是无偏的。

(1)*IESI* 与 *RDSI* 相比, *ABS* 和 *RMSE* 的表现基本相当, 测验的平均长度 *Nf* 也基本相当, 但 χ^2 的表现有较大的改进;

SixS 中的六种选题策略, 在低能力点时(比如能力值为−2.7, −2.4 时), *ABS* 和 *RMSE* 较大。在 $b \sim U$ (−3, 3)的题库下, 随着能力值的提高, *ABS* 和 *RMSE* 的值在逐步减少; 在 $b \sim N$ (0, 1)题库下, 随着能力值的提高, *ABS* 和 *RMSE* 的值在逐步减少, 而增加到接近高能力点时(比如能力为 2.4, 2.7 时), 这两个指标的值又增大了, 但还是小于低能力点的值; 而在 $b \sim U$ (−3, 3)题库下, 同一能力点上测验长度 *Nf* 的表现相差不大, 在 N (0, 1)题库下, 越接近于高能力点, 不同选题策略的表现有些差异, *mid* 策略最好, b_{jt}-ε 策略最差。

(2)*DC* 策略在不同的题库结构, 对能力估计, 平均测验长度的影响不大, *DC* 策略的 χ^2 在第三种

题库结构下表现最好, 在第二种题库结构下的表现稍差;

DC 在平均测验长度(*Nf*)比 *MFI* 要多用两个项目, *ABS*、*RMSE*、χ^2 均比 *MFI* 好, 特别 χ^2 的值降到不足 *MFI* 的一半, 即极大地降低了题库中项目的曝光率;

DC 策略在不同能力点上能力估计的 *ABS* 和 *RMSE* 与 *MFI* 的变化趋势相同, 在低能力点和高能力点的值要大于其他的能力点, 但基本都小于 *MFI*。

另外对于不同的题库还有如下发现:

(1)能力估计的 *ABS*、*RMSE* 值与题库中 *b* 的分布有关, 比如 *SixS* 中六种选题策略当 *b* 服从标准正态分布要好于 *b* 服从均匀分布; 曝光率也与题库中 *b* 的分布有关, *mid*、*mid*-ε、*Avg*1、*Avg*1-ε 这四种选题策略的实验表明, *b* 服从均匀分布要好于 *b* 服从标准正态分布;

(2)测验长度与题库中 *a* 的分布有关, *SixS* 所代表的 6 种选题测略的实验结果表明, *a* 服从均匀分布要好于 *lna* 服从标准正态分布;

针对降维的选题策略(*RDSI*), 我们给出的 *IESI* 降低了 χ^2 的值。这是因为采用了区间估计选题法, 通过已测验项目累计的信息量动态调整选题范围以及建立影子题库, 它不仅有使用影子题库的优点, 而且影子题库的"大小"还由测验精度所控制, 因此尽管增加了候选的项目, 但是测验精度基本保持。

针对 *MFI* 进行分析, 它的优点是可以综合项目的所有参数和被试能力参数, 其效果是测验短而精度高; 缺点是导致高区分度项目的过早使用或者频繁使用, 降低题库使用率, 危害考试的安全。我们给出 *DC* 选题策略, 它使用信息量并且引入因子 $\sqrt[f_j]{\prod_{t=1}^{f_j}\left|b_{jt}-\theta\right|}$, 使得它不仅能够用数学式表达多个等级难度与能力估计值接近的要求, 而且具有 *MFI* 的优点; 同时 *DC* 不断调剂区分度的作用, 使得 *DC* 选题策略在保持 *MFI* 的一些优点的同时, 还能降低曝光率, 降低 χ^2 的值, 即避免 *MFI* 的一些缺点。

迄今为止没有人讨论过等级评分模型的 *b*-*STR* 是什么样子, 但是针对 0-1 评分模型设计的 *b*-*STR* 的思想对于构造等级评分模型的选题策略是可以借用的, 本文中就借用了这种思想构造等级评分模型的 *b*-*STR* 选题策略; 然而这种借用是否有道理? *a*-*STR* 是静态的, 动态的 *a*-*STR* 是什么样子? 本文构造了一种动态 *a*-*STR*, 它和静态 *a*-*STR* 相比, 又有什么优点? 是否还可以构造其他形式的动态 *a*-*STR*? 这些似乎都是可以进一步讨论的问题。

由于测验的精度和安全性相互制约, 如何比较 *CAT* 中选题策略的优劣难以说清楚。对定长 *CAT*, Barrada 等 (2010)提出一个方案: 在限定测量精度条件下比较测验安全性, 或者在限定测验安全性条件下比较测验的准确性。然而, 这个方案和结果都不能直接推广到不定长 *CAT* 场合(比如对不定长 *CAT* 不清楚如何解释重叠率(*overlap rates*))。他们似乎是针对 0-1 评分的情况, 所以我们并没有采用他们的比较方案。程小扬等人(2011)对 0-1 评分 *CAT* 引入了一种新的选题策略, 这种选题策略其实可以用于多级评分 *CAT*。程小扬等人的新选题策略和本文引入方法的比较, 值得研究。

参 考 文 献

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement, 34*(6), 438–452.

Chang, H. H., Qian, J. H., & Ying, Z. L. (2001). A-stratified multistage CAT with b-blocking. *Applied Psychological Measurement, 25*, 333–341.

Chang, H. H., & Ying, Z. L. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213–229.

Chang, H. H., & Ying, Z. L. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211–222.

Chen, P., Ding, S. L., Lin, H. J., & Zhou, J. (2006). Item selection strategies of computerized adaptive testing based on graded response model. *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.]

Cheng, X. Y., Ding, S. L., Yan, S. H., & Zhu, L. Y. (2011). New item selection criteria of computerized adaptive testing with exposure-control factor. *Acta Psychologica Sinica, 43*(2), 203–212.

[程小扬, 丁树良, 严深海, 朱隆尹. (2011). 引入曝光因子的计算机化自适应测验选题策略. *心理学报, 43*(2), 203–212.]

Cheng, Y., Chang, H. H., Douglas, J., & Guo, F. M. (2009). Constraint-weighted a-stratification for computerized adaptive testing with nonstatistical constraints. *Educational and Psychological Measurement, 69*(1), 35–49.

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33*(6), 419–440.

Dai, H. Q., Chen, D. Z., Ding, S. L., & Deng, T. P. (2006). The comparison among item selection strategies of CAT with multiple-choice items. *Acta Psychologica Sinca, 38*(5), 778–783.

[戴海崎, 陈德枝, 丁树良, 邓太萍. (2006). 多级评分题计算机自适应测验选题策略比较. *心理学报, 38*(5),

778–783.]
Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*(1), 5–22.
Li, M. Y., Zhang, M. Q., & Zhu, J. X. (2010). Test security methods in computerized adaptive testing. *Advances in Psychological Science, 18*(8), 1339–1348.
[李铭勇, 张敏强, 简小珠. (2010). 计算机自适应测验中测验安全控制方法评述. *心理科学进展, 18*(8), 1339–1348.]
Liu, Z., Ding, S. L., & Lin, H. J. (2008). Item selection strategies for computerized adaptive testing with the generalized partial credit model. *Acta Psychologica Sinica, 40*(5), 618–625.
[刘珍, 丁树良, 林海菁. (2008). 基于 GPCM 的计算机自适应测验选题策略比较. *心理学报, 40*(5), 618–625.]
Luo, Z. S., Ouyang, X. L., Qi, S. Q., Dai, H. Q., & Ding, S. L. (2008). IRT information function of polytomously scored items under the graded response model. *Acta Psychologica Sinca, 40*(11), 1212–1220.
[罗照盛, 欧阳雪莲, 漆书青, 戴海琦, 丁树良. (2008). 项目反应理论等级反应模型项目信息量. *心理学报, 40*(11), 1212–1220.]
Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*(3), 187–194.
Penfield, R. D. (2006). Applying bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education, 19*(1), 1–20.
Qi, S. Q., Dai, H. Q., & Ding, S. L. (2002). *Principles of modern educational and psychological measurement* (PP. 141–142, 252–256). Beijing, China: Higher Education Press.
[漆书青, 戴海琦, 丁树良. (2002). *现代教育与心理测量学原理* (PP. 141–142, 252–256). 北京, 中国: 高等教育出版社.]
Quellmalz, E. S., & Pellegrino, J. W. (2009). Perspective: Technology and testing. *Science, 323*(2), 75–79.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs, 34*(17).
van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*(2), 201–216.

Dynamic and Comprehensive Item Selection Strategies for Computerized Adaptive Testing Based on Graded Response Model

LUO Fen; DING Shu-Liang; WANG Xiao-Qing

(*School of Computer and Information Engineering*, *Jiangxi Normal University*, *Nanchang* 330022, *China*)

Abstract

Item selection strategy (ISS) is a core component in *Computerized Adaptive Testing* (CAT). Polytomous items can provide more information about examinee compared with dichotomous items, and adopting polytomously scored items in test is a research direction of CAT. As we know, the most widely used ISS is the *maximum Fisher information* (MFI) criterion, which raises concerns about cost-efficiency of the pool utilization and poses security risks for CAT programs. Chang & Ying (1999) and Chang, Qian, & Ying (2001) proposed two alternative item selection procedures, the *a*-stratified method (*a*-STR) and the *a*-stratified with *b* blocking method (*b*-STR) based on dichotomous model, with the goal to remedy the problems of item overexposure and item underexposure produced by MFI. However, the technology of *a*-STR and *b*-STR is static because the items are stratified according to the given information at the beginning of test. Based on *graded response model* (GRM), a technique of the reduction dimensionality of difficulty (or step) parameters was employed[0] to construct some ISSs recently. The limitation of this dimension reduction technique is that it loses a lot of information. Thus, in order to improve *MFI*, two new item selection methods are proposed based on GRM: (1) modify the technique of the reduction dimensionality of difficulty (or step) parameters by integrating the interval estimation; (2) dynamic *a*-STR and dynamic *b*-STR methods are implemented in the testing process. On one hand, these new ISSs can avoid and remedy the limitations of MFI and make good use of the advantages of the *Fisher information function* (FIF); FIF compresses all item parameters and ability parameters, so it is a comprehensive tool for all parameters in nature. On the other hand, the new ISSs employ the property that FIF could represent the inverse of the variance of the ability estimation, let ε be the square root of the reciprocal of

the Fisher information, d be the absolute deviation between the estimate ability and the function of the parameters of an item, which may be chosen and could be changed during the course of CAT, the inequality of $d<\varepsilon$ has the form of interval estimation, and its utility could be imaged as a more flexible shadow item pool.

A simulation study based on GRM was conducted. Four item pools of different structures were simulated, and 1000 examinees was generated and their abilities were randomly drawn from the standard normal distribution *N* (0,1). Each pool consists of 1000 polytomous items and the maximum score of each item was randomly selected from set {3, 4, 5, 6}. In this paper, we assume the prior distribution of ability is standard normal and the Bayesian *expected a posteriori* (EAP) is employed to estimate the ability parameter. The CAT test stopped when the accumulative information satisfies the pre-determined value *M* (*M*=16) or reaches the pre-assigned test length 30.

The results of the simulation study show that the new item selection methods required shorter test lengths and lower average exposure rates than the other methods, while maintaining the accuracy of ability estimation. More specifically, the new ISSs which applied the idea of the interval estimate were better than the correspondent ISS in terms of the Chi-square value. And the same effect appeared when comparing the dynamic *a*-STR and dynamic *b*-STR ISS with *MFI*. Some important results are also found by comparing different structure of item pool. The accuracy of ability estimation and item exposure rate were related to the distribution of the difficult parameters *b*, that is, the accuracy of ability estimation obtained from the condition in which *b* was sampled from *N* (0,1) was better than that when *b* was sampled from uniform distribution. The conclusion of item exposure rate is on the contrary. Also, the test length was related to the distribution of the discrimination parameter *a*, the test length required by the condition in which *a* was sampled from uniform distribution was shorter than that when the logarithm of *a* was sampled from *N* (0,1). In a word, in terms of controlling and balancing the item exposure, the new ISSs may gain an advantage over the former correspondent ISS.

Key words *Graded Response Model* (GRM); *Computerized Adaptive Testing* (CAT); Item selection; Interval estimation; *b*-STR based on polytomously scored model