心理学报 2011, Vol. 43, No.2, 203−212
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2011.00203

引入曝光因子的计算机化自适应测验选题策略*

程小扬[1,2] 丁树良[1] 严深海[2] 朱隆尹[1,2]

([1] 江西师范大学计算机信息工程学院, 南昌 330027)

([2] 赣南师范学院数学与计算机科学学院, 赣州 341000)

摘 要 在计算机化自适应测验(CAT)的研究中, 制定既高效又安全的选题策略是一个追求目标。用极大项目信息量准则(MIC)选题使得测验效率高、能力估计准确, 缺点是项目调用很不均匀, 影响考试的安全; 按 a 分层法通过控制试题曝光率以提高考试的安全性, 但该方法可能会使测验效率略有下降, 且该方法在各层内部无法实现对区分度的调整。本文针对上述两种选题策略的优缺点, 对 0-1 评分下的 CAT, 通过引入曝光因子、分阶段自动调整区分度的影响以及提高选题准确性等手段, 对 MIC 和 a-STR 进行改进, 引入了两类新的选题策略。计算机模拟实验显示, 新的选题方法效果比较理想。

关键词 曝光因子; 计算机化自适应测验; 选题策略; 3PLM

分类号 B841

1 问题的提出

基于项目反应理论(Item Response Theory, IRT)的计算机化自适应测验(Computerized Adaptive Testing, CAT)是心理测量、计算机技术和现代教育技术充分结合的产物。CAT 与传统的“千人一卷”的考试根本不同, 它的目的是为每一个考生安排一场最优的测验(Meijer & Nering, 1999)。

据 Meijer 和 Nering (1999)介绍, 国外有许多大型的测验都采用 CAT 考试形式, 在国内也有许多关于 CAT 的应用与研究。王茜娟等人(王茜娟, 丁树良, 谭渊, 2005)考虑了 CAT 选题时的内容平衡问题; 陈平等人(陈平, 丁树良, 林海菁, 周婕, 2006)和戴海琦等人(戴海琦, 陈德枝, 丁树良, 邓太萍, 2006)研究了等级反应模型下的选题策略; 刘珍等人(刘珍, 丁树良, 林海菁, 2008)对于GPCM模型进行了研究; 林海菁和丁树良(2007)甚至对具有认知诊断功能的 CAT 做过研究; 田建全等人(田建全, 苗丹民, 杨业兵, 何宁, 肖玮, 2009)将 CAT 应用到征兵入伍心理检测系统的拼图测验中, 取得了较好的效果。适用于 0-1 评分的 Logistic 模型分为单参数 Logistic 模型(1PLM), 双参数 Logistic 模型(2PLM)和三参数 Logistic 模型(3PLM)。由于 CAT 一般不允许被试放弃反应, 故在 0-1 评分情况下研究 CAT 选择 3PLM 是合适的。IRT 中(Fisher)信息量是一个重要的概念。信息量是能力的函数, 理论上信息量近似于能力估计方差的倒数, 即信息量越大, 方差越小, 测量越准确。在局部独立性假设下, 测验信息量可以表示成各个项目信息量之和。而项目信息量又与该项目区分度的平方成正比, 信息量的这些性质是下文中最大信息量选题准则(MIC)的理论基础。

在 CAT 的考试过程中, 特别在高风险(high stake)测试中, 考试安全是十分重要的。通常, CAT 通过减少项目过度曝光率来控制考试安全, 而当前 CAT 中控制项目曝光率的途径主要有两种: 一是扩大题库中优质项目的数量; 二是通过合理的选题策略使题库中的项目调用更加均匀。对于一个既定的

收稿日期: 2010−04−12

* 国家自然科学基金(编号:30860084, 60263005), 江西省教育厅青年科学基金项目(编号:GJJ10238), 全国教育考试”十一五”科研规划课题(编号:2009JKS2009), 教育部人文社科项目(编号:09YJCXLX012, 10YJCXLX049), 江西省高等学校教学改革研究课题(编号:JXJG101113)。

通讯作者: 丁树良, E-mail: ding06026@163.com

题库, 选题策略的好坏直接关系到 CAT 的质量。在 CAT 选题策略的研究中, 目前常用的选题策略是 Lord 提出的 MIC、Chang 和 Ying (1999)提出的按 a 分层法(a-STR)。MIC 能很好地综合各项目参数及能力参数, 该方法具有测验效率高、能力估计准确等优点, 缺点是区分度高的项目会被频繁使用, 而区分度低的项目较少调用(甚至不用)。在这种情况下, 考生可能对某些过度曝光的试题提前做好准备, 从而降低试题的真实难度, 进而影响能力估计的准确性和考试公平性; 对于另一些试题则使用率太低, 造成浪费。因此, 设计 CAT 选题策略时希望所有的试题都能较均匀的使用。a-STR 是针对 MIC 的缺陷而设计的, 它可以使项目曝光次数变得比较均匀。但 a-STR 只是将题库中的项目按 a 的大小分成几层, CAT 对每一个被试能力的估计是逐渐变准确的。使用 a-STR 选题策略可以在一定程度上做到在能力估计越准确的阶段, 使用区分度越高的项目; 而能力估计越不准确的阶段, 使用区分度越低的项目。由于 a-STR 选题策略分层的数目预先确定, 在同一层中对备选的项目的要求是其难度与能力估计值相匹配; 但是执行 a-STR, 区分度不能按照指定的规则跟随能力估计精度的变化而逐题做比较细微的变化, 这可能是影响该方法测验效率的一个原因。

除以上介绍的 a-STR 方法及其变式外, 国外学者提出了许多其它的方法来控管项目曝光率。1985 年 Sympson 和 Hetter 第一个提出用条件概率方法(简称 SH 法)来解决项目过度曝光问题, 其基本思想为在项目选择和最终抽取之间加设一个曝光控制参数, 用来决定项目被最终施测的机率; 为满足实际对各能力水准下控制项目曝光率的严格程度不同, Stocking 和 Lewis (1998)将 SH 法延伸, 提出条件曝光控管法(简称 SLC 法), 使得各能力水准下的曝光率得到妥善控管。SH、SLC 可有效的控管项目曝光率, 是实际常采用的控管曝光率的方法。但这些方法存在以下缺点：不能提高那些曝光率过低项目的抽取几率, 整个题库的利用率不能提高; 当施测情境与模拟曝光参数情境不同, 则无法控管试题曝光率, 如题库进行删除和添加时, 必须重新模拟计算曝光参数, 模拟曝光控制参数的计算较费时; 而且, 当模拟时的能力分布与实际被试的分布不同时, 它的控管效果大打折扣。为解决 SH 及其衍生方法的不足, var der Linden 和 Veldkamp (2004, 2007)提出了一种有效率的曝光控管方法(简称 LV 法), LV 通过线上调整试题合格机率来控管试题曝光率, 这避免了 SH 需事前迭代模拟所产生的缺失, 且对项目曝光的控制与 SH 一样有效, 但它可能因为过度曝光率的试题太多时, 存在无法选到试题的情况。所以 Chen (2004)建议采用修改的 LV (MLV)法来控管试题曝光率, MLV 与 LV 方法基本相同, 只是在需考虑内容平衡时有所不同, MLV 不会碰到选不到题的情形, 但会造成前后被试的测验重叠率较高问题(Ju, 2005); 与 MLV 控管原理基本相同, Ju (2005)提出 SH 线上(on-line)控管法(简称 SHO 法), 该方法对 SH 法进行改造, 以解决 SH 法需事先模拟计算曝光参数的过程;为避免 MLV 前后测验重叠率偏高的问题, Chen 和 Liao (2005) 建议将冻结(Freeze, 或者翻译为休眠)程序加入 SHO 中, 形成了新的曝光控管方法(简称 SHOF), 该方法能进一步提升曝光率控管效能。我国台湾学者也对 CAT 的曝光率控管方法进行了深入研究。朱怡君(2005)在硕士论文中对 CAT 测试的初探阶段提出采用 b 分层随机法选题, 精确估计阶段采用 a-邻近法(简称 a-NN), 以此来降低高曝光率项目的曝光率和提高题库中未使用项目的使用率; 吴玫玲(2006)对 SH、SHO、MLV、SHOF 四种方法在不同的情境下的控管成效进行比较, 发现 SHOF 不管在何种情境下都能取得较稳定的效果; 陈升座(2007)提出以能力分布为基础之 SHC 曝光率控管法(简称 ASHC), 此方法根据被试能力水平给予不同的曝光控制, 能力值落入较多的区间曝光控制更严格, 反之更宽松, 这样很好地改进 SHC 曝光率控管法在能力估计精准度上的过渡损耗。蔡笃松(2008)提出了在线 ASHC 控管方法(简称 ASHCOF)。最近 Cheng, Chang, Douglas 和 Guo (2009)提出平衡测验效率和曝光率的方法, 即增加非统计约束的约束加权 a 分层方法, 效果不错。

针对前所述及 a-STR 之不足, 本文建立一个新的选题策略。该选题策略有三个特点：第一, 建立一个信息量的函数(FII) 而不是直接使用信息量, 以整合 MIC 和 a-STR 的优点; 第二, 尽管按 a-STR 方法可以考虑各层之间的区分度, 但 a-STR 方法无法考虑各层内部项目的区分度(Cheng et al., 2009), 而 FII 是各个项目的区分度的函数, 这可弥补 a-STR 的这一缺陷; 第三, 加入控制曝光机制, 当某个项目在考试过程中被频繁调用时, 利用该机制使该项目在以后的测验中更难被选中, 反之当某项目较少使用时,利用该机制增加被选中的机会, 通

过这样的在线控管, 均衡整个题库项目的曝光率, 提高题库的利用率。直接把项目调用的频度作为选题策略表达式的一部分, 目的是让每一个项目的使用频度尽量与题库中所有项目的平均使用次数接近, 试图既有效降低高曝光率项目曝光, 又能提升低曝光率项目的使用机率, 从而达到曝光均匀的目的, 这不同于 SH 等方法只是控制高曝光率项目的使用, 因此它可能要比上述的其它方法在曝光率控制方面更加有效。

信息量有机地综合项目的所有参数和能力参数, 这是信息量的优点; 而 MIC 又有缺陷, 所以本文在实施 CAT 时分阶段地使用不同的信息量函数, 以扬长避短。这种分阶段地调用不同的信息量函数的方法, 使得每个被试在前期调用区分度相对较低的项目, 而在后期调用区分度较高的项目, 由于区分度是信息量函数的自变量之一, 故该方法自始至终都自动地受到区分度的调节, 而不像 a-STR 那样, 只是在层与层之间自动受区分度调节。

2　引入曝光因子的选题策略

2.1　3PLM 模型与信息量简介

伯恩鲍姆(A. Birnbaum)给出了如下的三参数 Logistic 模型(3PLM)(漆书青, 戴海琦, 丁树良, 2002):

$$P_i(\theta)=c_i+\frac{1-c_i}{1+e^{-Da_i(\theta-b_i)}} \tag{1}$$

其中θ为潜在特质, 也称为能力, D 为常数(一般取值 1.7), a_i、b_i、c_i分别表示第 i 个项目的区分度、难度和猜测度。$P_i(\theta)$表示能力为θ 的被试在第 i 个项目上正确作答的概率。3PLM 的项目信息函数(漆书青等, 2002)为:

$$I_i(\theta)=\frac{D^2a_i^2(1-c_i)}{[c_i+e^{Da_i(\theta-b_i)}][1+e^{-Da_i(\theta-b_i)}]^2} \tag{2}$$

在项目反应理论中, 由局部独立性假设, 测验信息函数为:

$$I(\theta)=\sum_{i=1}^{n}I_i(\theta)$$

其中 n 为测验项目数量, $I_i(\theta)$为能力值为θ 的被试在项目 i 上的信息函数。

2.2　新的选题策略

针对 CAT 中使用 MIC 选题会使项目的使用不均匀, 影响考试安全的缺点, 我们在 MIC 的基础上引入一个项目 j 的控制曝光因子(exposure-control factor), 记为 $ecf(j)$, 以及 $ecf(j)$ 的调节因子λ_j和区分度 a_j 的幂函数 $a(j,T,k)$, 这里 $ecf(j)=\frac{m_j}{\bar{m}}$, $a(j,T,k)=a_j^{\frac{2(T-k)}{T-1}}$, 新的选题策略如下:

$$f_j=I_j(\hat{\theta})\Big/[\lambda_j\cdot ecf(j)\cdot a(j,T,k)] \tag{3}$$

当对第 m 个考生施测时, $\hat{\theta}$表示该考生当前能力估计值, m_j表示项目 j 被前 m−1 个考生调用的次数, $\bar{m}$为前 m−1 个考生调用题库中所有项目的平均次数, 即 $\bar{m}=\sum_{j=1}^{M}m_j/M$, M 为题库项目总数。(3)式中的 T 表示分 T 个阶段选题, k(k 取值为 1, 2…T)表示当前 CAT 实施中选题所处的阶段,我们取项目i_0作为下一个施测题, 其中

$$i_0=\arg\max_{j\in R_a}f_j \tag{4}$$

这里 R_a 是尚未对该被试施测的项目集, 也可以称为该被试的剩余题库(陈平等, 2006; 刘珍等, 2008)。即对于被试α在 CAT 的作答过程中, 题库中当前尚未抽取给被试α作答的所有项目。

仔细考察(3)式, $ecf(j)$位于分母之中, 当 m_j越大, $ecf(j)$也越大, 而 f_j 则减小, 说明该项目被调用的次数越多, 则其在后面被选中几率降低; 而 m_j越小, $ecf(j)$也越小, 而 f_j 则增大, 说明该项目被调用的次数越少, 则它在后面被选中几率增加。如此达到更均匀地调用项目的目的。当然在实际施测过程中, 为避免 m_j为 0 的情况, 我们用 $m_j+\varepsilon$(ε 为足够小的正数)来代替 m_j, 用 $\bar{m}+\varepsilon$ 来代替 $\bar{m}$。称公式(3)中的λ_j为项目 j 的曝光因子参数, 它的作用是调节 $ecf(j)$对选择项目的影响。λ_j也处于分母之中, 本文中它取如下两种值:

(1) λ_j=1, 这时相当于λ_j不施加影响。

(2) λ_j的取值与 $ecf(j)$有关, 一般来讲, $ecf(j)$越大, 则λ_j取较大的值, $ecf(j)$越小, 则λ_j取较小的值, λ_j在 $ecf(j)>0.5$时取不小于 1 的值, 否则取小于 1 的值。具体取值如表 1。

表 1　λ_j的取值与 $ecf(j)$的关系

$ecf(j)$值落在的区间	≥3	[2,3)	[1.5,2)	[0.5,1.5)	[0.3,0.5)	<0.3
λ_j的取值	3	2	1.5	1	0.5	0.3

当然，根据具体的考试需要，λ_j 的取值可做相应的变化。当特别强调考试安全时，可将表 1 中各项数值的差距扩大；当要强调考试的效率时，可适当将表 1 中各项数值的差距缩小。

通过 3PLM 的信息量函数可知，信息量的大小与区分度的平方成正比。而在 CAT 的测试过程中，通常在开始阶段所估的被试能力与真实能力相差甚远，这时使用区分度大的项目，等于浪费，它还会造成早期参加考试的被试频繁地调用区分度高的项目，从而使这些区分度高的项目的 $ecf(j)$ 值变大，相应的 f_j 的值就较小，而较晚参加的被试就不容易选中这些项目，这时只能调用区分度较低但 $ecf(j)$ 值较小的项目，从而使测验长度增加，影响测验效率。受 a-STR 思想的启发，在公式(3)中加入 $a(j,T,k)$，分阶段地使用不同的函数值。在每个被试作答的前期，采用削弱区分度对信息量影响的函数值；随着 CAT 考试的进行，逐步加强区分度的影响。因此，对所选的题目分别除以 $a(j,T,k)$（即 a_j^2 至 a_j^0 之间将指数 T 等分)。如当 T=4，则选题时分四个阶段，每个阶段分别除以 a_j^2、$a_j^{\frac{4}{3}}$、$a_j^{\frac{2}{3}}$、a_j^0，用这种方法达到尽量避免被试在前期调用高区度项目的目的。为了说明这一点，设想一下，对低区分度项目，即 a_j<1 时，有 $a_j^2 < a_j^{4/3} < a_j^{2/3} < a_j^0 = 1$，于是在(3)中分母添加上 $a(j,T,k)$，对每一个被试可以达到在实施 CAT 的早期阶段区分度越小的项目越可能被使用的效果。

Chang 和 Ying (1999)的 a-STR 方法主观上想保证在能力估计较粗糙阶段使用区分度小的那些项目，而能力估计越精确，越使用区分度较大的那一批项目(记为 rough-low, accurate-high 策略，简记为 R-L，A-H 策略)；在宏观(各个分层子题库)层面上，a-STR 的确是使用了这种策略，但它无法保证在各层中间，即微观(项目)层面也采用这种选题策略(Cheng et al., 2009)；新的选题策略希望宏观和微观上都能实现能力估计越粗糙时越使用低区分度的项目；为了更好地达到这一目的，新选题策略添加控制曝光因子，使得在整个测验中，各个项目的曝光次数彼此牵制，这又相当于“中观”层面上达到项目平衡选取的目的，这等于间接地使用 R-L, A-H 策略。

定义 1　引入曝光因子的最大信息量选题策略

在 CAT 作答过程中，从剩余题库 R_a 中选取满足 $\max\limits_{j\in R_\alpha} f_j$ 的项目作为被试 α 的下一个测试项目。

为表述简洁，将 λ_j 取第一种情况(λ_j =1)的选题策略称为引入曝光因子的最大信息量法 1(简记为 Modi-MIC 1)；将 λ_j 取第二种情况(请参见表 1)的选题策略称为引入曝光因子的最大信息量法 2(简记为 Modi-MIC 2)。

定义 2　引入曝光因子的按 a 分层法

受 Cheng 等人(2009)的启发，将题库中的项目按 a 分层后，对 a-STR 也加入控制曝光因子而形成的新的选题策略，被试选择的项目 i_0 满足以下公式：

$$i_0 = \arg\min_{j\in R_{ah}} |\hat{\theta} - b_j| \cdot \frac{\lambda_j m_j}{\bar{m}}$$

这里的 R_{ah} 为被试 α 在当前层 h 尚未作答的项目集合。同样地，我们将 λ_j =1 时的选题策略称为引入曝光因子的按 a 分层法 1(简记为 Modi-a-STR 1)，将满足表 1 中所规定的 λ_j 选题策略称为引入曝光因子的按 a 分层法 2(简记为 Modi-a-STR 2)。

采用以上方法进行选题时，都要知道何时结束当前阶段或当前层，以进入下一阶段或下一层选题。本文中是根据信息量的大小来确定结束当前阶段或当前层，各阶段(层)累积测验信息量(陈平等, 2006；戴海琦等, 2006)的大小满足如下公式：

$$I_k = I_{总}(k/T)^2$$

其中 $I_{总}$ 设置为 16(即测验标准误为 0.25)，层数 T 为 4，k 表示当前所在的层数或阶段数。

3　CAT 的模拟过程

3.1　被试及题库模拟

由于被试能力真值是无法知晓的，项目参数真值也是未知的，要评价一种新的 CAT 选题策略对能力估计的影响，只能用 Monte Carlo 模拟试验。本文用 Monte Carlo 模拟方法进行试验，模拟的数据结构(陈平等, 2006)如下：

(1)产生一批项目参数，各参数的分布情况分四种情况：①区分度参数 a 服从对数正态分布，难度参数 b 服从标准正态分布，记为 $\ln a \sim N(0,1)$，$b \sim N(0,1)$；②$\ln a \sim N(0,1)$，难度参数 b 服从均匀分布，记为 $b \sim U(-3,3)$；③$a \sim U(0.2,2.5)$，$b \sim N(0,1)$；④$a \sim U(0.2,2.5)$，$b \sim U(-3,3)$。对于以上四种题库分布，猜测概率参数 c 均服从 α 为 5 和 β 为 17 的贝塔分布，记为 $c \sim Beta(5,17)$，且 $a \in [0.2,2.5]$，$b \in [-3,3]$。

(2)产生一批能力参数，能力参数 θ 服从标准

的正态分布, 记θ~$N(0,1)$。

3.2 施测过程

有了高质量的题库后, 就可安排被试参加 CAT 的测试。CAT 测试一般分为定长测试和不定长测试, 在本文中除了采用引用文献的控管曝光率的选题策略时用定长测验外, 其他情况都采用不定长测试, 即根据项目信息量的大小来结束考试。施测过程一般又可分为两个阶段: 一是对被试能力一无所知的探测阶段, 在该阶段我们采用从剩余题库中随机抽取 3 道试题给被试作答, 计算被试的初始能力, 初始能力值可用得分与失分之比的自然对数求得(若得分为 0 或得分为 3, 则分别令能力初值为−3 或+3); 当得到被试的初始能力之后, 则进入精确探测阶段, 这时依据本文所介绍的选题策略选择项目 j 作答, 确定被试在该项目上的得分, 再用贝叶斯后验期望法(EAP) (Bock & Mislevy, 1982)估计被试的能力。重复该步骤, 直至满足测验结束条件。

测试过程可以通过以下方法模拟被试得分(陈平等, 2006): 根据被试能力真值θ 和当前所选择的项目 j 的参数, 由公式(1)计算被试在第 j 个项目上的答对概率 $P_j(\theta)$, 再产生一个随机数 r($0 \leqslant r \leqslant 1$), 若 $r \leqslant P_j(\theta)$, 则该被试在第 j 个项目上得 1 分, 否则得 0 分。

3.3 评价指标

本文用能力估计的准确性、能力估计标准差、项目调用的均匀性、人均用题数、χ^2 检验统计量、测验效率、测试重叠率七个评价指标(刘珍等, 2008)来评价 CAT 的质量。

能力估计准确性

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

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, $\bar{\hat{\theta}}_i = \sum_{j=1}^{C}\hat{\theta}_{ij} / \mathrm{C}$, 表示被试 i 在 C 次模拟中得到的能力估计平均值。

能力估计标准差

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

项目调用均匀性

$$SE = \frac{1}{C}\sum_{j=1}^{C}\left(\sqrt{\sum_{i=1}^{M}\left(m_{ij} - \bar{m}_j\right)^2 \Big/ M}\right)$$

其中 M 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, $\bar{m}_j = \sum_{i=1}^{M} m_{ij} / \mathrm{M}$ 表示在第 j 次模拟中项目被调用的平均次数。

人均用题数　　$Nf = \frac{1}{C}\sum_{j=1}^{C}\left(\sum_{i=1}^{N} r_{ij} \Big/ N\right)$

其中 r_{ij} 表示被试 i 在第 j 次模拟中的作答项目数。

测验效率　　$\text{测验效率} = \sum_{i=1}^{N}\inf_i / \sum_{i=1}^{N} L_i$

其中, inf_i 为被试 i 测验的信息总量, L_i 为被试 i 的测试长度。

χ^2 检验统计量

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

其中 A_j 是第 j 题曝光率。计算 A_j 的公式为 A_j= 第 j 题被使用的次数/N

测试重叠率　　$\mathrm{Rt} = 2TO_{总} / [(N-1)\sum_{i=1}^{N} L_i]$

其中, $TO_{总}$是考生的试题重叠总数, 计算方法为: $TO_{总} = \sum_{j=1}^{M} C_{M_j}^2$。$M_j$ 是试题库中第 j 题使用次数, $C_{M_j}^2$ 为从 M_j 个元素中取出二个元素的组合数。

以上这些指标中, 测验效率是越大越好, 而其它六个指标都是越小越好。为了更直观地反映测验总体效果, 我们将这些指标进行整合, 采用的办法是统一量纲(陈平等, 2006; 戴海琦等, 2006; 刘珍等, 2008)再加权求和。具体作法是: 对于值越大越好的指标, 将该评价指标上的最大值作为分母, 把各选题策略在该指标上的值作为分子, 求两者的比值; 对值越小越好的指标, 则将评价指标上的最小值作为分子, 把各选题策略在该指标上的值作为分母, 求两者的比值。再对某选题策略的 7 个评价指标比值分别赋加权系数。加权求和值最大的, 则该选题策略在几个方面的综合效果最好; 反之则最差。用公式表述如下, 若某选题策略在指标 i 所求比值为 X_i, 加权系数为 $\mathrm{W_i}$, 则加权求和的值 Y 为

$$Y = \sum W_i X_i \qquad (5)$$

其中加权系数根据实际的测验性质及评价重点来确定。本文中我们假设试卷的七个方面同等重要, 即加权系数均设定为 1。当然, 根据考试用途和目标不同, 可分别赋不同的加权系数给各评价指标, 实际使用时, 则可以考虑选取上述 Y 值最大所对应的选题策略。

4 实验结果分析

通常, MIC 和随机选题法分别代表着最佳的效率和最好的项目曝光率控制方法(Cheng et al., 2009), 而 a-STR 能够在效率与项目曝光率两者之间实现较好地平衡。如上所述, 当前对项目曝光率控制的方法有很多, 在此, 选择 SH、MLV、SHO、a-NN、SHOF 法作为新方法在控制曝光率指标上的比较对象。由于在这些控管方法设计时, 只考虑了定长测试情况, 所以这 5 种方法实验时也采用定长测试, 长度设为 14 (通过实验, 长度 14 可达到与其它方法相近的精度), 在这些方法选题时, 常用 MIC 选题, 因此, 分别在这些方法前冠以 MIC。同理, 实验时再选用 SHO 与 a-STR 结合方法(简称 a-STR-SHO)作为控管的 a-STR 方法的比较(同样采用定长, 长度为 32)。在实验中, 题库分层数和分阶段数都设为 4, 测验总信息量取 16, 每个实验都重复 30 次。题库中项目的数量和被试的人数都为 1000。文中的模拟数据(包括项目参数、被试能力参数)由 Mathlab 7.0 程序产生, CAT 的施测程序采用 VC 6.0 编写运行。

4 个实验数据(表 2~表 5)显示, 几种选题方法的总体效果从大到小的顺序为 Modi-MIC、随机法、传统的曝光控管方法(SHOF、SH、SHO、MLV、a-NN)、MIC、Modi-a-STR、a-STR。从各项指标来看, MIC 在测验效率(人均用题数指标和测验效率指标)方面最好, 但在反映项目曝光率性能的几个指标反映最差; 随机法曝光率控制最好, 但测验效率与能力估计准确性最差。在 MIC 中加入传统的几种曝光控制方法, 则测验效率很高, 与 MIC 接近, 而项目曝光率比 MIC 好。当 a-STR 与 SHO 结合使用, 曝光率指标明显降低, 但没有 Modi-MIC 优秀, 且测验效率不高。Modi-MIC 在项目均匀性、χ^2 指标、测试重叠率等方面表现特别优秀, 它在能力估计准确性、能力估计标准差方面效果也较理想。随着曝光因子参数 λ_j 的变大, 项目均匀性会变小,

表 2 1n *a*~*N*(0,1), *b*~*N*(0,1), *c*~*Beta*(5,17)时的实验结果

选题策略	能力估计准确性	人均用题数	项目均匀性	能力估计标准差	测验效率	χ^2 指标	测试重叠率	统一量纲
MIC	0.2041	12.9	47.271	0.2441	1.333	173.2	0.1853	4.1727
Modi-MIC 1	0.2016	23.5	10.95	0.242	0.729	5.1	0.0277	4.8426
Modi-MIC 2	0.2018	25.2	9.395	0.2403	0.674	3.5	0.0277	4.9632
a-STR	0.1997	39	30.112	0.2363	0.435	23.2	0.0613	3.3327
Modi-a-STR 1	0.1984	39.3	24.206	0.2339	0.43	14.9	0.0532	3.4929
Modi-a-STR 2	0.2004	39.3	20.877	0.2373	0.429	11.1	0.0494	3.5772
随机法	0.2286	56.9	7.333	0.2589	0.231	0.9	0.0569	4.6255
MIC-SH(定长)	0.1949	14	39.742	0.2325	1.381	112.8	0.1259	4.3238
a-NN(定长)	0.199	14	41.225	0.2349	1.366	121.4	0.1345	4.2609
MIC-MLV(定长)	0.1952	14	39.715	0.2343	1.382	112.7	0.1258	4.3156
MIC-SHO(定长)	0.1946	14	40.287	0.2315	1.387	115.9	0.1291	4.3258
MIC-SHOF(定长)	0.1966	14	38.9	0.2352	1.366	107.9	0.121	4.3062
a-STR-SHO(定长)	0.1985	32	25.1	0.2359	0.549	19.6	0.0507	3.6451

表 3 1n *a*~*N*(0,1), *b*~*U*(−3,3), *c*~*Beta*(5,17)时的实验结果

选题策略	能力估计准确性	人均用题数	项目均匀性	能力估计标准差	测验效率	χ^2 指标	测试重叠率	统一量纲
MIC	0.1984	13.1	50.474	0.2397	1.308	193.94	0.2063	4.2968
Modi-MIC 1	0.1997	26.3	17.821	0.2392	0.654	12.07	0.0374	4.4589
Modi-MIC 2	0.1984	31.1	15.6	0.2378	0.549	7.82	0.038	4.4017
a-STR	0.1981	42.9	56.61	0.2364	0.398	74.73	0.1167	3.0523
Modi-a-STR 1	0.198	43.5	50.01	0.2373	0.389	57.49	0.1001	3.1126
Modi-a-STR 2	0.2009	45.3	42.805	0.2401	0.367	40.44	0.0848	3.1581
随机法	0.254	60.9	7.599	0.2937	0.154	0.95	0.061	4.5113
MIC-SH(定长)	0.2013	14	40.241	0.2441	1.245	115.67	0.1288	4.3033
a-NN(定长)	0.2014	14	44.897	0.243	1.255	143.98	0.1571	4.2413
MIC-MLV(定长)	0.2007	14	40.087	0.2427	1.246	114.78	0.1279	4.3153
MIC-SHO(定长)	0.2012	14	40.825	0.243	1.253	119.05	0.1322	4.3038
MIC-SHOF(定长)	0.2019	14	39.2	0.2448	1.23	109.7	0.1228	4.3058
a-STR-SHO(定长)	0.1947	32	41.9	0.2346	0.566	54.8	0.0859	3.4762

表 4　*a*~*U*(0.2,2.5), *b*~*N*(0,1), *c*~*Beta*(5,17)时的实验结果

选题策略	能力估计准确性	人均用题数	项目均匀性	能力估计标准差	测验效率	χ^2指标	测试重叠率	统一量纲
MIC	0.1986	12.2	40.49	0.2405	1.438	134.32	0.1457	3.6665
Modi-MIC 1	0.1991	17.6	6.914	0.2403	0.989	2.71	0.0193	5.0593
Modi-MIC 2	0.2015	18.2	6.187	0.2428	0.948	2.1	0.0193	5.2039
a-STR	0.1995	21.8	19.472	0.2394	0.794	17.36	0.0382	3.4453
Modi-a-STR 1	0.2001	22	12.865	0.2423	0.786	7.52	0.0286	3.8291
Modi-a-STR 2	0.1995	22.1	11.731	0.2406	0.78	6.22	0.0274	3.9337
随机法	0.2123	46.1	6.665	0.2458	0.326	0.96	0.0461	4.2855
MIC-SH(定长)	0.1713	14	38.964	0.2091	1.755	108.45	0.1216	3.9930
a-NN(定长)	0.1748	14	36.37	0.2125	1.678	94.49	0.1076	3.9501
MIC-MLV(定长)	0.1718	14	38.433	0.2092	1.75	105.51	0.1186	3.9936
MIC-SHO(定长)	0.1712	14	38.947	0.2091	1.756	108.35	0.1215	3.9943
MIC-SHOF(定长)	0.1707	14	37.8	0.2077	1.744	102	0.1151	4.0104
a-STR-SHO(定长)	0.1541	32	22.2	0.1874	0.918	15.5	0.0465	3.6597

表 5　*a*~*U*(0.2,2.5), *b*~*U*(–3,3), *c*~*Beta*(5,17)时的实验结果

选题策略	能力估计准确性	人均用题数	项目均匀性	能力估计标准差	测验效率	χ^2指标	测试重叠率	统一量纲
MIC	0.2	12.2	45.981	0.2409	1.42	172.78	0.1842	3.8419
Modi-MIC 1	0.1999	20.7	13.77	0.2415	0.835	9.15	0.0289	4.3649
Modi-MIC 2	0.1991	23.4	12.607	0.2409	0.735	6.78	0.0293	4.3088
a-STR	0.199	23.6	32.384	0.24	0.734	44.42	0.0671	3.2619
Modi-a-STR 1	0.1988	24.2	25.286	0.2405	0.714	26.45	0.0497	3.4662
Modi-a-STR 2	0.1995	25.4	21.527	0.2412	0.678	18.24	0.0427	3.5774
随机法	0.223	59.8	7.602	0.2625	0.211	0.97	0.0599	4.2401
MIC-SH(定长)	0.1837	14	39.884	0.2258	1.481	113.62	0.1268	3.9746
a-NN(定长)	0.1888	14	40.858	0.2305	1.435	119.24	0.1324	3.8892
MIC-MLV(定长)	0.1858	14	39.557	0.2274	1.475	111.77	0.1249	3.9603
MIC-SHO(定长)	0.1871	14	39.908	0.229	1.48	113.76	0.1269	3.9467
MIC-SHOF(定长)	0.1865	14	38.6	0.2277	1.461	106.3	0.1194	3.9625
a-STR-SHO(定长)	0.1543	32	39.2	0.1888	0.911	48.1	0.0792	3.5754

项目曝光率有所改善, 但人均用题数也随之上升, 测验效率与 MIC、SH 等方法相比在降低, 而它的测验效率与曝光率指标都要比 a-STR 好。a-STR 虽然在项目曝光率性能方面较 MIC、SH 等方法优秀, 但要比引入曝光因子的各新选题方法差, 它的总体效果也最不理想。还有, Modi-MIC 比 Modi-a-STR 在项目曝光率的改善速度更快, 从而在反映项目曝光率的几个指标都以 Modi-MIC 2 较佳。

在引入曝光因子的选题策略中, 我们引入 $\lambda_j \cdot ecf(j)$ 来控制项目的选择, 但由于题库中相应的优质项目(a 值较大的项目)有限, 而随着被试人数的增多, 是否会使这些项目调用较频繁, 从而这些项目的 $\lambda_j \cdot ecf(j)$ 值会变大, 而使较晚参加测验的被试只能选择 a 值小一些的项目导致测验长度增加, 这一问题值得重视。为了考察新采用的分阶段调用函数方法是否可以有效避免出现这种情况,分别记录了前期、中期、后期的被试的测验平均长度, 具体见表 6(题库分布为 1n *a*~*N*(0,1), *b*~*N*(0,1))。

从表 6 可知, 对于前期、中期、后期的被试的测验平均长度相差并不大, 并未引起较晚参加测验的被试的长度增加。其它三个题库下的实验结果也基本一致, 因篇幅所限, 故省略。

为更加详细地了解项目的被使用情况, 我们在实验时记录了不同选题策略下各项目曝光率, 项目曝光率的计算方法为各项目被使用的次数除以总的被试人数(Cheng et al., 2009)。然后, 我们将各项目曝光率按升序排列, 具体见表 7。

从表 7 可知, MIC 的项目曝光率极为不均衡, 它会使某些项目调用非常频繁, 而某些项目很少使用; SH、MLV、SHO、SHOF 等方法能将项目的最高曝光率控制在 0.2 左右, 但低曝光率项目的被使用率并未提高(75%的项目曝光率与 MIC 基本一致, 使用率较低), 而且, 最大的 5%的项目被使用次数超过总使用次数的 60%, 这样, 考试的安全性仍无法得到保证; a-STR 会使最大的项目曝光率减少,

表 6 各阶段被试的平均测验长度

选题策略	前 10%测验长度	中 10%测验长度	后 10%测验长度
Modi-MIC 1	25.5	24.8	23.5
Modi-MIC 2	27.2	26.2	24.7
Modi-a-STR 1	40.3	39.9	39.2
Modi-a-STR 2	40.3	40.0	39.0

注：前 10%测验长度为第 1 ~ 第 100 位被试的平均测验长度，中 10%测验长度为第 450 ~ 第 549 位被试的平均测验长度,后 10%测验长度为第 901 ~ 第 1000 位被试的平均测验长度。

表 7 各项目曝光率情况(1n *a*~*N*(0,1), *b*~*N*(0,1), *c*~*Beta*(5,17))

选题策略	min	25%	50%	75%	max	前最大 5%	前最小 5%
MIC	0.002	0.0028	0.0031	0.0035	0.6069	65.8%	0.9%
Modi-MIC 1	0.0087	0.0162	0.0201	0.0272	0.0821	12.0%	2.4%
Modi-MIC 2	0.0131	0.0179	0.0224	0.0314	0.0654	9.4%	3.1%
a-STR	0.0036	0.0185	0.0318	0.053	0.3784	20.3%	0.5%
Modi-a-STR 1	0.0064	0.0231	0.0337	0.0519	0.2514	13.2%	1.2%
Modi-a-STR 2	0.0072	0.0229	0.0342	0.0532	0.2045	11.6%	1.4%
随机法	0.0498	0.056	0.0569	0.0578	0.0607	5.3%	4.7%
MIC-SH(	0.0022	0.0029	0.0031	0.0034	0.2094	63.0%	0.9%
a-NN	0.002	0.0028	0.003	0.0033	0.3978	62.6%	0.9%
MIC-MLV	0.0021	0.0029	0.0031	0.0034	0.2017	62.9%	0.9%
MIC-SHO	0.0022	0.0029	0.0031	0.0034	0.2149	63.5%	0.9%
MIC-SHOF	0.0022	0.0029	0.0031	0.0034	0.1925	61.5%	0.9%
a-STR-SHO	0	0.0150	0.0259	0.0426	0.2064	15.8%	0.6%

注：min 为项目的最小曝光率，max 为项目的最大曝光率,25%表示从小到大第 25 个分位点的项目的曝光率，其余类推。前最大 5%表示曝光率最大 5%的项目被调用次数之和占总调用次数的百分比，前最小 5%表示曝光最小的 5%的项目被调用次数之和占总调用次数的百分比。

但最小的项目曝光率提高却不明显。Modi-MIC 2 是除随机法外，各项目曝光率最均衡的方法,它对低使用率项目的曝光率提升明显,而对高使用率项目的曝光率下降明显(接近于随机法)。

综合以上实验，引入曝光因子能大大降低项目的曝光率，特别是引入曝光因子的最大信息量法，它会使考试的安全性得到极大的提高，当然测验长度会有所增加，但比 a-STR 要好，能力估计的准确性差异不大。在以后的 CAT 测验中，Modi-MIC 2 可以成为一种备选的相当安全且高效的选题策略。

5 讨论

本文在 3PLM 下引入曝光因子的两类选题策略，一类是对 MIC 的修正，它在选题策略(函数)中除以区分度的某次幂的方式，实现了对 a 的自动分层，加入了控制曝光率的因子，使得选题更均匀，再应用信息量对项目参数和能力参数的有效综合，提高测验效率；另一类是对 a-STR 的修正，它在 a-STR 的基础上，加入控制曝光率的因子以及能力估计值与项目难度匹配方法，以期进一步提高 a-STR 的项目使用均匀性和测验效率。将新的选题策略与传统的 MIC、a-STR、SH、SHO、SHOF、MLV、a-NN、随机法比较，实验数据显示，加入曝光因子的选题策略会使项目的调用更加均匀，曝光率指标明显降低，当然测验长度有一定增加，而能力估计准确性相差不大，这为提高 CAT 的安全性提供了一条途径。当然，本文只是引入了一个曝光因子，该因子的设置是否是最科学的，其中λ_j 取怎样的值会使选题策略效果最佳，还有引入的曝光因子能否应用到多级评分模型，效果如何？另外，在试验中，随机选题策略的加权分数较高，是因为其曝光率均匀性太好(χ^2 指标值小)，而使得其综合分数飚升，所以如何确定(5)式中的权重，以反映客观事实。这些都有待于进一步研究。

参 考 文 献

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP Estimation of Ability in a Microcomputer Environment. *Applied Psychological Measurement, 6*(4), 431–444.

Cai, T. S. (2008). *The algorithm of IRT CAT system with item exposure rate*. Master's dissertation. Asia University, Taiwan.

[蔡笃松. (2008). *具试题曝光率控管之 IRT 适性测验演算法*. 硕士论文. 亚洲大学.]

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

Chen, S. Y. (2004). *Controlling Item Exposure on the Fly in Computerized Adaptive Testing*. Paper presented at the Annual Meeting of the Taiwanese Psychological Association, Taipei, Taiwan.

Chen, S. Y., & Liao, W. H. (2005). *Controlling Item Exposure and Test Overlap on the Fly in Computerized Adaptive Testing*. Paper presented at the Annual Meeting of the Psychometric Society, Tilburg, Netherlands.

Chen, P., Ding, S. L., Lin, H. J., & Zhou, J. (2006). Item selection strategies of CAT under polytomous scored item 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.]

Chen, Z. S. (2007). *Ability distribution based SHC procedure*. Master's dissertation of Graduate institute of Educational Measurement and Statistics National Taichung University, Taichung.

[陈升座. (2007). *以能力分布为基础之 SHC 曝光率控管法*. 国立台中教育大学教育测验统计研究所硕士论文.]

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

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(in Chinese). *Acta Psychologica Sinica*, *38*(5), 778–783

[戴海琦, 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]

Ju, Y. (2005). *Item exposure control in a-stratified computerized adaptive testing.Unpublished master's thesis*. National Chung-Cheng University, Chia-Yi, Taiwan.

Lin, H. J., & Ding S. L. (2007). An exploration and realization of computerized adaptive testing with cognitive diagnosis. *Acta Psychologica Sinica*, *39*(4), 747–753.

[林海菁, 丁树良. (2007). 具有认知诊断功能的计算机化自适应测验的研究与实现. *心理学报, 39*(4), 747–753]

Liu, Z., Ding, S. L., & Lin, H. J. (2008). Item selection strategies for computerized adaptive testing with the generalized partial credit model(in Chinese). *Acta Psychologica Sinica*, *40*(5), 618–625.

[刘珍, 丁树良, 林海菁. (2008). 基于 GPCM 的 CAT 选题策略比较. *心理学报, 40*(5), 618-625.]

Meijer R. R, & Nering M. L. (1999). Computerized Adaptive Testing: Overview and Introduction. *Applied Psychological Measurement, 23*(3), 187–194.

Qi, S. Q., Dai, H. Q., & Ding, S. L. (2002). *Principals of modern educational and psychological measurement (in Chinese)*. Beijing, 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*. 北京, 高等教育出版社.]

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57–75.

Sympson, J., & Hetter, R. (1985). Controlling item exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977), San Diego, CA: Navy Personnel Research and Development Center.

Tian, J. Q., Miao, D. M., Yang, Y. B., He, N., & Xiao, W. (2009). The Development of Computerized Adaptive Picture Assembling Test for Recruits in China (in Chinese). *Acta Psychologica Sinica*, *41*(2), 167–174.

[田建全, 苗丹民, 杨业兵, 何宁, 肖玮. (2009). 应征公民计算机自适应化拼图测验的编制. *心理学报*, *41*(2), 167–174.]

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*, 273–291.

van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistic*s, *32*, 398–418.

Wang, X. J., Ding, S. L., & Tan, Y. (2005). Research on c-Stratified CAT with Variable Length (in Chinese). *Journal Of Jiang Xi Normal University: Natural Science*, *29*(3), 227–230.

[王茜娟, 丁树良, 谭渊. (2005). 按 c-分层不定长 CAT 的研究[J]. *江西师范大学学报*:*自然科学版*, *29*(3), 227–230.]

Wu, M. L. (2006). *Investigating Item Exposure Control on the Fly in Computerized Adaptive Testing*. Master's dissertation. National Chung Cheng University.

[吴玫玲. (2006). *电脑适测验在线曝光率控管之研究*. 硕士论文. 国立中正大学心理研究所.]

Zhu, Y. J. (2005). *Item Exposure Control in A-Stratified Computerized Adaptive Testing*. Master's dissertation. National Chung Cheng University, Chiayi.

[朱怡君. (2005). *a-分层电脑适性测验之曝光率控管*. 硕士论文. 国立中正大学心理研究所.]

New Item Selection Criteria of Computerized Adaptive Testing with Exposure-Control Factor

CHENG Xiao-Yang[1,2]; DING Shu-Liang[1]; YAN Shen-Hai[2]; ZHU Long-Yin[1,2]

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

([2] *College of Mathematics and Computer Science*, *Gannan Normal University*, *Ganzhou* 341000, *China*)

Abstract

As far as Computerized Adaptive Testing (CAT) is concerned, the issue of item selection strategy has received more attention because of its vital role. It is well known that there are two typical selection strategies called Maximum Information Criterion (MIC) and a-Stratification (a-STR). However, both of the two strategies have their advantages together with their downsides. On the one hand, MIC method can obtain high efficiency and accurate estimation of ability; on the other hand, its uneven item selection may lead to the insecurity of examination. Meanwhile, though a-STR can improve the test security by controlling the item exposure rate, it may result in the inefficiency of the test and failure in adjusting the discrimination within the layers. As a result, the development of both effective and safe item selection strategies has always been a goal to pursue in studies on CAT.

According to the previous studies, the test security can be enhanced and the item pool utilization rate can be increased by balancing the item exposure rate. Therefore, in 0-1 scored CAT, two new item selection strategies are proposed in this paper to improve the MIC and a-STR methods by introducing exposure factor, adjusting automatically the discrimination by stage and increasing the accuracy of item selection. One of the new item selection strategies has three prominent characteristics: First, a function of item information (FII) rather than the item information function is set up to combine the advantages of both MIC and a-STR. Second, the effect of the discrimination on different stages in CAT is taken into account and a function of item discrimination is used in the FII to make up for the defect of a-STR for not being able to control the item discrimination in the internal layer. Third, mechanism of online control exposure is adopted. While some specific items in a certain examination process are more frequently exposured than others, with the help of the mechanism, they will turn out to become less likely to be selected in the future tests. At the same time, some specific items in a certain examination process are less frequently exposured than others, with the help of the mechanism, they will turn out to become more likely to be selected in the future tests. Thanks to the mechanism of online control exposure, the whole exposure rate of all the items in the item pool is evened and the utilization rate of the item pool is increased.

In order to fill the gap between the exposure rate of each item and mean exposure rate of all the items in the item pool, this paper attempts to treat the item exposure rate directly as a part of the selection strategies expression. And it also tries to equalize the exposure rate by decreasing the exposure rate of the items with high ones and increasing the uses of the items with low ones. The approach differs from the approaches which only control the items with high exposure rate, e.g. SH. The results of Monte Carlo simulations show that compared with other approaches, the approach proposed in this paper is more effective in terms of exposure control and more ideal in the performance of other indexes.

Key words exposure-control factor; computerized adaptive testing; item selection strategy; 3PLM