

认知诊断 CAT 中项目曝光控制方法的比较

毛秀珍^{1,2} 辛 涛²

(¹ 四川师范大学教育科学学院, 成都 610066) (² 北京师范大学发展心理研究所, 北京 100875)

摘 要 项目曝光率关系到题库建设和测验安全, 是计算机化自适应测验(*Computerized Adaptive Testing*, CAT)需要考虑的重要问题。在认知诊断 CAT 情形下, 首先基于传统 CAT 中 α -分层方法的思想提出按项目信息量对题库分层的分层多阶段(*Stratified Multistage*, SM)选题方法; 然后将 SM 方法与项目合格(*Item Eligibility*, IE)方法相结合得到 SMIE 方法。在此基础上, 开展模拟研究比较 SM、IE、SMIE、最大修正优先指标 (*Maximum Modified Priority Index*, MMPI)方法、限制阈值(*Restrictive Threshold*, RT)方法和限制进度(*Restrictive Progressive*, RPG)方法的选题表现。总体上, 它们的测量精度从高到低依次为 IE、SM、SMIE、RT、RPG 和 MMPI 方法; 项目曝光分布均匀性的优劣次序为 MMPI、RPG、SMIE、RT、SM 和 IE 方法; SMIE 和 RT 方法能较好地平衡测量精度和项目曝光均匀性要求。

关键词 认知诊断计算机化自适应测验; 选题方法; 测量精度; 项目曝光率

分类号 B841

1 引言

计算机化自适应测验(*Computerized Adaptive Testing*, CAT)根据被试潜特质水平自适应地选择测验项目能极大地提高测量精度。特别地, 认知诊断计算机化自适应测验(*Cognitive Diagnostic Computerized Adaptive Testing*, CD-CAT)以认知诊断理论为基础, 通过被试作答较少项目反馈其掌握了哪些知识没有掌握哪些知识, 从而能为教师、家长和学生提供丰富的教学指导信息。同传统 CAT 一样, 它包括题库、初始项目的选择、选题策略、特质估计方法和测验终止规则五个基本组成部分。其中, 选题策略直接影响项目调用均匀性, 关系到测验安全和题库建设。一方面, 曝光率越大, 项目使用的次数越多。于是, 考生之间越可能分享试题信息从而威胁测验安全并降低反应真实性。尽管大多数 CD-CAT 是诊断性测验, 测验风险低、测验安全不是大问题。但是当认知诊断的结果与某些教学评价指标挂钩时, CD-CAT 可能变成高风险测验, 测验安全就将变得非常重要。另一方面, 题库建设和维护是一项费时费力的基础性工作。认知诊断题

库不仅要包含大量质量合格的项目, 而且随着时间的推移, 题库中的一些项目会因为存在缺陷、过时或过度曝光等原因需要用新题去替换或进行增补(陈平, 2011; 唐小娟, 丁树良, 俞宗火, 2012)。如果项目曝光率低表明项目没有得到充分利用。当大部分项目都曝光过低甚至没有使用时, 将极大地浪费题库建设中花费的人力和财力。由此可见, 降低过度曝光项目的使用率和提高曝光不足项目的使用率对 CD-CAT 具有重要意义。一般地, 测量精度和项目曝光控制往往相互制约。因此, 好的选题策略应平衡测量精度和项目曝光均匀性要求。

到目前为止, 传统 CAT 中至少有以下 5 类项目曝光控制方法(Georgiadou, Triantafillou, & Economides, 2007): (1)随机化方法, 例如随机程序法(*Randomesque Procedure*, RP) (Kingsbury & Zara, 1989)和进度程序方法(*Progressive Procedure*, PG) (Revuelta & Ponsoda, 1998); (2)分层选题策略, 例如 α -分层选题策略(Chang & Ying, 1999); (3)曝光率参数控制方法, 例如 Sympon-Hetter 方法(Sympon & Hetter, 1985)和项目合格方法(*Item Eligibility*, IE) (van der Linden & Veldkamp, 2004); (4)结合多种方

法的选题策略, 例如 Leung, Chang 和 Hau (2003)将 α -分层方法和约束 CAT 方法相结合选题以及(5)多阶段适应性测验设计方法(Luecht & Nungester, 1998)。此外, 程小杨、丁树良、严深海和朱隆尹(2011)提出引入曝光因子的项目选择方法也结合了多种方法选题。

其中, α -分层方法和 IE 方法是两种典型的控制项目曝光率的选题方法。前者按项目区分度参数从小到大的顺序将题库分为 K 层并将测验分为 K 个阶段。当测验进行到第 k ($k=1, 2, \dots, K$) 阶段时, 从第 k 层中选择难度 b 与能力估计值 $\hat{\theta}$ 最接近的 n_k 个项目, 并使得各阶段项目个数之和等于测验长度。 α -分层方法简单易行, 能提高低区分度项目的使用率和题库使用率, 但不能显著降低过度曝光项目的使用率, 项目曝光率分布范围广(Chang & Ying, 1999)。后者对每个被试将生成一个由所有对其有资格施测的项目构成的影子题库, 然后从该影子题库中选择项目(van der Linden & Veldkamp, 2004, 2007)。研究表明, 除个别项目外, IE 方法使所有项目的曝光率都小于或等于预先设定的最大值。虽然它能较好地控制过度曝光项目的使用率, 却不能提高曝光不足项目的使用率。

针对 CD-CAT 中使用常用的选题策略会导致题库使用率低、项目曝光不均匀(陈平, 李珍, 辛涛, 2011)的问题, 陈平(2011)借鉴 α -分层方法和最大优先指标方法(Cheng & Chang, 2009)的思想提出按确定性输入, 噪声‘与’门模型(the deterministic inputs, noisy “and” gate model, DINA) (Junker & Sijstma, 2001)的项目猜测和失误参数之和分层的选题方法(记为 SHE (Stra))和基于 Shannon 熵(SHE)信息量的最大修正优先指标方法(记为 SHE (MMPI))。其中, SHE (MMPI)方法利用最大优先指标方法的思想, 将最大项目曝光率作为唯一约束条件并结合 Shannon 熵信息量, 得到第 i 个项目的修正优先指标(the Modified Priority Index, MPI)为(陈平, 2011):

$$MPI_i = \frac{r - n_i / N}{r} \cdot (C - SHE_i) \quad (1)$$

其中 r 是允许的最大项目曝光率, n_i 是第 i 个项目被调用的次数, N 是参加测验的总人数。为了统一方向, 常数 C 必须大于所有项目的 SHE 信息量, 例如令 C 等于 100。该方法能大幅度降低过度曝光项目的使用率并提高曝光不足项目的使用率、题库利用率高且测验重叠率低, 项目曝光率近似均匀分布, 但同时也降低了测量精度。

另外, Wang, Chang 和 Huebner (2011)基于 PG 和 RP 方法提出限制进度 (Restrictive Progressive, RPG)方法和限制阈值(Restrictive Threshold, RT)方法。他们运用融合模型(the Reparameterized Unified Model, RUM) (Hartz, 2002)并假设测验考查 4 个独立属性的条件下基于后验加权 Kullback-Leibler 信息量(Posterior Weighted Kullback-Leibler, PWKL)进行模拟研究。结果表明, RT 方法在测量准确性方面一致优于 RPG 方法, 其曝光均匀性比 RPG 方法的结果稍差。总体上, RT 和 RPG 方法能提高曝光不足项目的使用率并控制最大项目曝光率, 而且测量精度较高。但测验考查的属性个数较少、实验条件简单, 所以有必要考察它们在其它条件下的选题结果。

基于上述 CD-CAT 相关研究的不足, 本研究拟借鉴已有方法的思想探索并比较多种项目曝光控制方法的表现。具体而言, 首先基于传统 CAT 中 α -分层方法的思想, 提出按项目信息量对题库分层的分层多阶段(Stratified Multistage, SM)选题方法; 其次将 SM 方法与 IE 方法相结合得到 SMIE 方法; 再次将基于 SHE 信息量的 MMPI 方法应用到 Kullback-Leibler (KL)信息量、PWKL 和结合后验和距离加权的混合 KL (Hybrid Kullback-Leibler, HKL)信息量的情形。在此基础上, 比较 SM、IE、SMIE、MMPI、RT 与 RPG 方法的选题表现。

2 选用的认知诊断模型和选题方法

2.1 DINA 模型

认知诊断模型刻画了项目特征、被试特质水平和作答反应之间的关系, 是认知诊断理论的基础之一。迄今为止, 研究者提出了大量认知诊断模型。其中, DINA 模型假设测验考查的属性之间不具有补偿关系, 并用失误和猜测参数表示作答过程中的随机因素。失误参数表示当被试掌握项目 i 考查的所有属性时错误作答项目的概率, 即 $s_i = P(X_{ij} = 0 | \prod_{k=1}^K \alpha_{jk}^{q_{ik}} = 1)$; 猜测参数表示当被试至少有一个项目 i 考查的属性没有掌握时正确作答该项目的概率, 即 $g_i = P(X_{ij} = 1 | \prod_{k=1}^K \alpha_{jk}^{q_{ik}} = 0)$ 。其项目反应函数为:

$$P(X_{ij} = 1 | \alpha_j, s_i, g_i) = (1 - s_i)^{\eta_{ij}} \cdot g_i^{(1 - \eta_{ij})} \quad (2)$$

(2)式中 $\eta_{ij} = \prod_{k=1}^K \alpha_{jk}^{q_{ik}}$ 。 q_{ik} ($i=1, \dots, N; k=1, 2, \dots, K$) 为 Q 矩阵的元素, N 表示测验包含的项目个数, K 表示测验考查的属性个数。 q_{ik} 只取 1 或 0, 如果项目 i 考查属性 k , 则 $q_{ik} = 1$, 否则 $q_{ik} = 0$ 。 $\alpha_j = (\alpha_{j1},$

$\alpha_{j2}, \dots, \alpha_{jK}$) 为被试 j 的知识状态或称为属性掌握模式。如果他(她)掌握属性 i , 则 α_{ji} 等于 1, 否则 α_{ji} 等于 0。于是, 当被试掌握项目 i 考查的所有属性时 $\eta_{ji} = 1$, 其正确作答项目的概率为 $1 - s_i$; 如果被试 j 至少有一个项目 i 考查的属性没有掌握, 则其正确作答项目 i 的概率为 g_i 。

DINA 模型的参数较少, 模型简单容易解释, 而且 de la Torre (2009) 详细介绍了估计 DINA 模型参数的方法。因而它得到广泛的研究应用(de la Torre & Douglas, 2004, 2008)。特别地, DINA 模型还应用于 CD-CAT 的相关研究(陈平, 2011; Chen, Xin, Wang, & Chang, 2012; Cheng, 2009, 2010; 汪文义, 2009)。基于 DINA 模型的这些优势, 本文选用 DINA 模型模拟项目参数和被试反应。

2.2 六种控制项目曝光率的选题方法

2.2.1 SM 方法

Henson (2004) 定义认知诊断测验中项目 i 的区分度参数(Cognitive Diagnostic Discrimination Index, CDI)为:

$$CDI_i = \frac{1}{\sum_{u \neq v} d^2(\alpha_u, \alpha_v)^{-1}} \sum_{u \neq v} [d^2(\alpha_u, \alpha_v)^{-1} \times KL_i(\alpha_u, \alpha_v)] \quad (3)$$

其中, $d(\alpha_u, \alpha_v) = \sqrt{\sum_{k=1}^K (\alpha_{uk} - \alpha_{vk})^2}$ 表示知识状态 α_u 与 α_v 的欧氏距离。由(3)式可知, 诊断性测验中项目 KL 信息量与 CDI 指标不成正比。这与项目反应理论中费舍信息量和区分度参数 a 的关系不同。于是, 若参照 a -分层方法的思想, 用 CDI 代替 a 参数对题库分层选题, 不会取得与 a -分层方法一样的效果。但是分层的思想仍可以应用到 CD-CAT。因为通过限制项目选择的范围可以控制项目曝光率。另外, 项目信息量可以采用 KL、PWKL、HKL 或 SHE 信息量指标计算, 而项目信息量的大小是影响测量精度的主要因素之一。因此, 若在选择每个项目之前, 依据知识状态估计值处项目信息量的大小对题库分层, 既可以控制可选项目的范围, 又能选取较好的项目。特别地, KL、PWKL 和 HKL 的计算方法分别为:

$$KL_i(\hat{\alpha}^{t-1}) = \sum_{c=1}^{2^K} \sum_{x=0}^1 \log[P(X_i = x | \hat{\alpha}^{t-1}) / P(X_i = x | \alpha_c)] P(X_i = x | \hat{\alpha}^{t-1}) \quad (4)$$

$$PWKL_i(\hat{\alpha}^{t-1}) = \sum_{c=1}^{2^K} \sum_{x=0}^1 \log[P(X_i = x | \hat{\alpha}^{t-1}) / P(X_i = x | \alpha_c)] P(X_i = x | \hat{\alpha}^{t-1}) \pi_{i,t}(\alpha_c) \quad (5)$$

$$HKL_i(\hat{\alpha}^{t-1}) = \sum_{c=1}^{2^K} \sum_{x=0}^1 \log[P(X_i = x | \hat{\alpha}^{t-1}) / P(X_i = x | \alpha_c)] P(X_i = x | \hat{\alpha}^{t-1}) \pi_{i,t}(\alpha_c) (1/d(\hat{\alpha}^{t-1}, \alpha_c)) \quad (6)$$

$\pi_{i,t}(\alpha_c) = \pi_{i,0}(\alpha_c) L(u^{t-1} | \alpha_c)$ 表示被试作答完前 $t-1$ 个项目后知识状态 α_c 的后验概率。其中, $\pi_{i,0}(\alpha_c)$ 为知识状态 α_c 的先验概率。令 u_k 代表项目 k 上的作答反应, 那么, 似然函数 $L(u^{t-1} | \alpha_c) = \prod_{k=1}^{t-1} [P_k(\alpha_c)]^{u_k} [1 - P_k(\alpha_c)]^{1-u_k}$ 。另外, SHE 的计算方法为:

$$SHE_i(\pi_{j,t}) = \sum_{q=0}^1 (H(\pi_{j,t} | x_j^{t-1}, x_{j,i} = q) \cdot P(x_{ji} = q | x_j^{t-1})) \quad (7)$$

式 (7) 中 $H(\pi_{j,t} | x_j^{t-1}, x_{j,i} = q) = -\sum_{c=1}^{2^K} (\pi_{j,t}(\alpha_c) \cdot \log(\pi_{j,t}(\alpha_c)))$ 是 $\pi_{j,t}(\alpha_c) (c=1, 2, \dots, 2^K)$ 的香农熵, 根据全概率公式可知 $P(x_{ji} = q | x_j^{t-1}) = \sum_{c=1}^{2^K} (P(x_{ji} = q | \alpha_c) \cdot \pi_{j,t-1}(\alpha_c))$ 。特别地, 若按 KL、PWKL 或 HKL 选题, 应选择具有最大信息量的项目; 若采用 SHE 信息量, 则应选择使 SHE 最小的项目(Cheng, 2009)。

一般地, CD-CAT 中运用 SM 方法选题包括以下几个步骤。第一, 产生初始知识状态。第二, 因为 PWKL、HKL 和 SHE 信息量要计算知识状态的后验概率, 所以在选择第一个项目时首先计算各个项目在初始知识状态处的 KL 信息量, 并按其值从小到大的顺序将题库分为 K 层, 然后将测验分为 K 个阶段。通常根据测验长度和题库容量来确定 K 的值(Chang & Ying, 1999)。最后在每一层具有最大信息量的 5 个项目中随机选择第一题。第三, 施测第一题, 并估计知识状态。第四, 此后固定一种信息量指标(如 KL、PWKL、HKL 或 SHE)计算剩余题库(对某个被试而言, 所有未施测项目的集合)中各个项目在知识状态估计值处的信息量, 并将剩余题库按信息量(若是 SHE 信息量, 则依其相反数)从小到大的顺序分为 K 层, 使得将已施测项目放入对应层之后, 各层项目个数的比率保持不变。当测验进行到第 $k (k=1, 2, \dots, K)$ 阶段, 在第 k 层具有最大信息量的 5 个项目中随机选择下一个项目。第五, 施测并估计知识状态。第六, 判断是否满足测验终止规则。若满足就结束测验, 否则回到第四步。

SM 方法与 a -分层方法的不同之处在于: (一)

前者在选择每个项目之前都要对题库分层, 而 a -分层方法只在测验之前进行一次分层; (二)前者在每次选题之前分层的结果可能不同, 而 a -分层方法则对题库进行固定分层, 所有被试都采用相同的项目层。另外, SM 方法不同于按项目所包含属性模式分层的选题方法(尚志勇, 丁树良, 2011)。因为可达矩阵 R (Tatsuoka, 2009)的每列代表一类考查相同属性的项目。尚志勇和丁树良(2011)称其为一种模式, 并在测验之前按项目所属的模式对题库分层, 每次在“最佳”项目类中选题。而 SM 方法是按项目信息量的大小对题库分层。

2.2.2 IE 方法

CD-CAT 中, IE 方法的思想与 van der Linden 和 Veldkamp (2004, 2007)中 IE 方法的思想完全一致。具体而言, 该方法首先定义 E_i 和 A_i 分别表示项目 i 是合格的和施测项目 i 的事件。易知事件 E_i 与 A_i 满足 $A_i \subseteq E_i$, 并且 $P(A_i) = P(A_i | E_i) \cdot P(E_i)$ 。然后在被试 $j+1(j=0,1,2,\dots,N-1)$ 参加测验之前计算所有项目对其合格的概率。假设最大项目曝光率为 $r^{\max}$, 那么 $P(A_i) \leq r^{\max}$, 即

$$P(E_i) \leq \frac{r^{\max}}{P(A_i | E_i)} = \frac{r^{\max} \cdot P(E_i)}{P(A_i)} \quad (8)$$

式(8)是可重复的, 即项目 i 对被试 $j+1$ 合格的概率为 $P^{j+1}(E_i) \leq \frac{r^{\max} \cdot P^{1,\dots,j}(E_i)}{P^{1,\dots,j}(A_i)}$ 。它等价于

$$P^{j+1}(E_i) = \min\left\{\frac{r^{\max} \cdot P^{1,\dots,j}(E_i)}{P^{1,\dots,j}(A_i)}, 1\right\} \quad (9)$$

其中,

$$P^{1,\dots,j}(E_i) = \frac{\text{对前 } j \text{ 个被试而言, 项目 } i \text{ 合格的频数}}{j} = \frac{\varphi_j}{j} \quad (10)$$

$$P^{1,\dots,j}(A_i) = \frac{\text{对前 } j \text{ 个被试而言, 施测项目 } i \text{ 的频数}}{j} = \frac{\alpha_j}{j} \quad (11)$$

为保证式(9)有意义, 每当(11)式中 $\alpha_j = 0$ 都有 $P^{j+1}(E_i) = 1$ 。第三, 将 $P^{j+1}(E_i)$ 与 $(0, 1)$ 区间的随机数 r_i 比较。若 $P^{j+1}(E_i)$ 大于 r_i , 则项目 i 有资格施测给被试 $j+1$ 。于是, 所有有资格施测给被试 $j+1$ 的项目就构成一个影子题库。最后, 从该影子题库中选择使 KL、PWKL 或 HKL 信息量最大(或使 SHE 最小)的项目施测给被试 $j+1$ 。

2.2.3 SMIE 方法

该方法在每个被试参加测验之前, 运用 IE 方法生成由所有有资格施测给该被试的项目组成影子题库, 然后运用 SM 方法选题。即在选择每个项目之前, 对影子题库中所有未施测的项目按知识状态估计值处信息量(若是 SHE 信息量, 则依其相反数)从小到大排序并分为 K 层。然后将测验分为 K 个阶段, 每个项目从对应项目层里信息量最大的 5 个项目中随机选择, 直到测验结束。

2.2.4 最大修正优先指标(Maximum Modified Priority Index, MMPI)方法

根据陈平(2011)提出的 SHE (MMPI)方法, 容易得到项目 i 基于 KL、PWKL 或 HKL 信息量的修正优先指标(Modified Priority Index, MPI)为:

$$MPI_i = \frac{r - n_i / N}{r} \cdot Index_i \quad (12)$$

其中, r 表示允许的最大项目曝光率, n_i 为第 i 个项目被调用的次数, N 为被试总数, $Index_i$ 表示项目 i 的 KL、PWKL 或 HKL 信息量。于是, 该方法将选择剩余题库中 MPI 值最大的项目。

2.2.5 RT 方法

根据 Wang 等(2011)可知, RT 方法在每个被试参加测验之前将曝光率大于预先设定的最大值的那些项目从题库中去掉后形成一个影子题库, 然后按如下两步选择第 $t(t=1,2,\dots,L)$ 个项目。第一, 将剩余题库中信息量落在最大(最小)信息量的一个较小区间内的项目放在一起构成候选项目集。例如, 当按 KL、PWKL 或 HKL 选题时, 候选项目集由信息量落在区间 $[\max(Index) - \delta, \max(Index)]$ 的项目构成; 当按 SHE 选题时, 则将信息量落在区间 $[\min(Index), \min(Index) + \delta]$ 的项目放在一起构成候选项目集。其中 $\delta = [\max(Index) - \min(Index)] * (1 - t/L)^\beta$, L 为测验长度。 β 的值越大, δ 越小, 测量越准确, 项目曝光均匀性越差。因此, β 是平衡项目曝光分布和测量精度的权重, 其值可根据测验要求灵活设置, 本文令 β 等于 0.5。第二, 在候选项目集中随机选取一个项目。

2.2.6 RPG 方法

当采用 KL、PWKL 或 HKL 信息量, 按 RPG 方法选择第 t 题可表示为(Wang et al., 2011):

$$i_t \equiv \arg \max\{(1 - \exp(-r))[(1 - t/L)R_i + Index_i \times \beta t / L], \quad i \in S_t\} \quad (13)$$

对 SHE 信息量, 则按下式选题:

$$i_t \equiv \arg \max \{ (1 - \exp_i / r) [(1 - t / L) R_i + (C - SHE_i) \times \beta t / L], \quad j \in S_t \} \quad (14)$$

L 为测验长度, $\exp_i$ 与 r 分别表示项目 i 的曝光率和期望项目曝光率, C 的含义与 SHE (MMPI) 方法中 C 的定义相同, S_t 为剩余题库。令 H^* 等于剩余题库中所有项目信息量的最大值, 那么 $R_i \sim U(0, H^*)$, β 是平衡项目曝光控制和测量准确性的权重, 它使得项目选择随着测验要求变得更加灵活, 本文取 $\beta = 0.5$ 。

3 方法

本文采用 MATLAB (R2010a) 为工具编写程序, 进行模拟实验。

3.1 题库的模拟

首先, Stocking (1994) 指出, 一个好的题库包含的项目数量至少是测验长度的 12 倍。Chang 和 Zhang (2002) 还建议采用更大的比例。鉴于本文的测验长度为 20, 为保证题库有足够多项目, 本研究模拟 300 个试题, 一共考查 6 个属性, 并假设属性之间相互独立。

其次, 本研究采用 Cheng (2009, 2010) 的方法确定项目特征。即每个题目至少考查一个属性, 并且考查每个属性的概率为 0.2。例如, 确定项目 i 是否考查属性 k 的方法为: 产生一个 $(0, 1)$ 的随机数 r , 如果 r 大于 0.2, 那么项目 i 不考查属性 k , 反之项目 i 考查属性 k 。最后可得项目 Q 矩阵。

再次, 确定项目参数。本研究采用 DINA 模型计算被试正确作答项目的概率。根据以往研究, 确定失误和猜测参数均服从区间 $(0.05, 0.25)$ 的均匀分布 (Cheng, 2010; de la Torre & Douglas, 2004)。于是, 产生 300 组服从该分布的随机数作为项目参数。项目参数描述统计的结果见表 1。

表 1 项目参数的描述统计结果

	最小	最大	平均值	标准差
失误参数	0.050	0.250	0.146	0.061
猜测参数	0.050	0.250	0.147	0.057

3.2 模拟被试的知识状态和作答反应

假设每位考生掌握每个属性的概率为 0.5, 则知识状态服从均匀分布。首先, 模拟 2000 个被试的属性掌握模式作为真值。然后, 运用 DINA 模型计算被试 i 正确作答项目 j 的概率 P_{ij} , 并产生 $(0, 1)$ 区间的随机数 p_{ij} 。如果 P_{ij} 大于 p_{ij} , 则被试 i 在项目 j

上的反应为 1, 否则为 0。

根据 Q 矩阵和被试知识状态的真值, 表 2 列出了考查各个属性的项目个数和掌握各个属性的被试人数。因为每个项目至少考查一个属性, 所以实际考查每个属性的项目个数会稍大于期望值 60。表 3 统计了考查可能属性个数的项目个数和相应的被试人数。

表 2 考查(掌握)各个属性的项目个数(被试人数)

属性	1	2	3	4	5	6
项目个数	82	66	77	85	80	74
被试人数	993	986	999	994	994	1001

表 3 考查(掌握)各种可能属性个数的项目个数(被试人数)

属性个数	0	1	2	3	4	5	6
项目个数	0	174	93	28	5	0	0
被试人数	25	175	488	659	447	185	21

3.3 知识状态的估计方法

随机选择一种知识状态作为被试的初始知识状态, 在被试作答完每个项目之后运用极大似然 (Maximum Likelihood Estimation, MLE) 方法估计考生的知识状态。

3.4 项目选择方法和测验终止规则

本研究采用 SM、IE、SMIE、MMPI、RT 和 RPG 六种方法选题。每一种选题方法分别根据 KL、PWKL、HKL 和 SHE 计算项目信息量。最大项目曝光率固定为 0.20, 测验长度为 20, 一共进行 24 次模拟实验。

3.5 评价指标

采用单个属性和属性掌握模式的判准率表示测量准确性。令 l_i 表示第 i 个属性被正确判断为掌握或没有掌握的被试人数, k 表示属性掌握模式得到正确判断的被试人数, N 表示被试总人数。于是, 属性 i 的判准率等于 l_i / N , 属性掌握模式的判准率等于 k / N 。

项目曝光率即项目的使用频率。本文选用项目曝光率的四分位数、未使用的项目个数、过度曝光的项目个数(即曝光率大于 0.20 的项目个数)、 χ^2 统计量和测验重叠率评价各选题方法项目曝光率的结果。其中, $\chi^2 = \sum_{i=1}^N [(ER_i - \overline{ER_i})^2 / \overline{ER_i}]$ 表示项目观察曝光率和期望曝光率之间的差异 (Chang & Ying, 1999)。项目 i 的期望曝光率 $\overline{ER_i}$ 等于测验长度 L 除以题库容量 M 。 χ^2 越小, 总体上项目观察曝光率与期望曝光率之间的差异越小。测验重叠率

定义为随机选择的两个被试之间期望重叠的项目个数与测验长度之比。假设有 N 个被试参加长度为 L 的测验, 它可以通过计算所有被试组的共同项目数之和除以 $LN(N-1)/2$ 得到。为简化计算并避免比较成对被试施测共同项目的个数, 运用公式(15) (Chen, Ankenmann, & Spray, 2003)计算测验重叠率,

$$\hat{T} = \frac{L}{M} S_{er}^2 + \frac{M}{L} \tag{15}$$

其中, M 为题库容量, S_{er}^2 表示项目曝光率的方差。测验重叠率越小, 项目曝光控制越好。

4 结果

4.1 单个属性和属性掌握模式的判准率

由表 4 易知, 当固定选题方法时: (1)按不同信息量指标选题对单个属性的判准率相差不大(KL 除外); (2)按 KL 信息量选题的模式判准率最低, 比其它信息量指标的判准率低 10%以上; (3)采用 PWKL 和 HKL 指标选题的模式判准率无明显差异, 它们的模式判准率均稍低于 SHE 信息量的模式判准率

(SM 方法除外)。

当固定信息量指标时, 不同方法之间无论对单个属性还是属性掌握模式的判准率都具有较大差异。任何两种方法对单个属性的判准率之差小于它们的模式判准率之差。具体而言: 第一, IE 方法的模式判准率一致高于 SM 方法的判准率。第二, 比较 SMIE、RT 和 SM 方法的结果知, 总体上它们对单个属性的判准率没有明显差异; SM 方法的模式判准率稍优于 SMIE 与 RT 方法的结果, RT 方法的模式判准率最低(HKL 和 SHE 信息量除外)。第三, RPG 和 MMPI 方法的判准率几乎没有差异(KL 除外)。它们对单个属性和属性掌握模式的判准率都远远低于 RT 方法的结果。例如, MMPI 和 RPG 方法中按 HKL 信息量选题的模式判准率为 0.81, 而 RT 方法中对应的模式判准率为 0.89。总体上, 这六种方法的测量精度从高到低依次为: IE、SM、SMIE、RT、RPG 和 MMPI 方法。

得到这样的结果并不奇怪。首先 IE 方法在有资格施测的项目库中选择信息量最大的项目, 可选

表 4 各种选题方法的判准率结果

选题方法	信息量指标	属性						模式判准率
		1	2	3	4	5	6	
IE	KL	0.97	0.95	0.95	0.97	0.97	0.94	0.79
	PWKL	0.98	0.99	0.99	0.99	0.99	0.99	0.94
	HKL	0.98	0.98	0.98	0.98	0.98	0.98	0.92
	SHE	1.00	0.99	0.99	0.99	0.99	0.99	0.96
SM	KL	0.95	0.92	0.96	0.97	0.93	0.92	0.72
	PWKL	0.98	0.98	0.99	0.98	0.99	0.98	0.91
	HKL	0.98	0.98	0.99	0.98	0.98	0.98	0.91
	SHE	0.98	0.98	0.98	0.98	0.98	0.97	0.90
SMIE	KL	0.95	0.94	0.94	0.93	0.96	0.93	0.72
	PWKL	0.98	0.97	0.98	0.98	0.98	0.97	0.89
	HKL	0.98	0.97	0.98	0.98	0.98	0.98	0.89
	SHE	0.98	0.98	0.98	0.98	0.98	0.98	0.90
RT	KL	0.94	0.95	0.96	0.93	0.95	0.94	0.70
	PWKL	0.98	0.97	0.98	0.98	0.98	0.98	0.88
	HKL	0.98	0.97	0.98	0.98	0.97	0.99	0.89
	SHE	0.98	0.98	0.98	0.98	0.99	0.98	0.90
MMPI	KL	0.93	0.90	0.90	0.91	0.92	0.90	0.60
	PWKL	0.95	0.92	0.96	0.96	0.95	0.94	0.80
	HKL	0.96	0.92	0.96	0.96	0.95	0.95	0.81
	SHE	0.98	0.94	0.97	0.97	0.97	0.95	0.83
RPG	KL	0.95	0.92	0.94	0.94	0.93	0.92	0.64
	PWKL	0.95	0.96	0.96	0.95	0.96	0.97	0.81
	HKL	0.96	0.95	0.97	0.95	0.96	0.97	0.81
	SHE	0.96	0.94	0.97	0.98	0.97	0.97	0.82

项目的个数较多, 所以其测量精度较高。其次, SM 方法通过分层控制每个项目选择的范围, 同时随着测验的进行, 将在质量越高的项目层中选择测验项目。因此, 它在控制项目曝光率的同时必将在一定程度上降低测量精度。第三, SMIE 方法结合运用 SM 和 IE 方法, 因而将进一步降低测量精度。第四, RT 方法每次都在项目信息量落在最大(小)信息量的一个较小区间内的项目中随机选择, 也将具有较高的测量精度。第五, MMPI 方法和 RPG 方法的结果与陈平(2011)的结果一致表明将最大项目曝光率约束与项目信息量结合构造项目选择指标, 能控制最大项目曝光率, 但也将一定程度上降低测量精度。

4.2 项目曝光率结果

各种选题方法项目曝光率的结果见表 5。当固定选题方法时, 按不同指标选题的项目曝光分布相似。当固定信息量指标时, 不同方法项目曝光率的

结果存在明显差异。这表现在: 第一, MMPI 方法中项目曝光率指数的四分位数接近期望项目曝光率 0.067, 最小值和最大值之差不大于 0.03, 测验重叠率低于 7%, 题库利用率达 100%, χ^2 值很小。因而, 项目曝光率近似均匀分布。第二, RPG 方法的项目曝光率主要集中于[0.05, 0.12]之间, 结果与 Wang 等(2011)基本一致。其项目曝光分布均匀性略差于 MMPI 方法的结果。第三, SMIE 方法能提高曝光不足项目的使用率并控制最大项目曝光率。特别地, 除 KL 外, 题库利用率达 100%, 测验重叠率小于 10%, 最大项目曝光率低于 0.22, 项目曝光率分布比较均匀。但其测验重叠率和 χ^2 值明显大于 RPG 方法的结果。因此, SMIE 方法的项目曝光分布均匀性不及 RPG 方法的结果。第四, SM 方法能提高曝光过低项目的使用率和题库利用率。尽管它与 SMIE 方法中过度曝光项目的个数差不多, 但它们的最大项目曝光率分别低于 0.65 和 0.22。这意味

表 5 各种选题方法项目曝光率的结果

选题方法	信息量指标	NU	曝光率					OE	OR	χ^2
			最小	Q ₁	Q ₂	Q ₃	最大			
IE	KL	171	0.000	0.000	0.000	0.191	0.226	54	0.189	36.664
	PWKL	123	0.000	0.000	0.005	0.188	0.225	48	0.183	34.586
	HKL	123	0.000	0.000	0.004	0.181	0.230	41	0.184	34.780
	SHE	126	0.000	0.000	0.004	0.177	0.223	45	0.180	33.59
SM	KL	28	0.000	0.021	0.052	0.086	0.649	9	0.145	23.540
	PWKL	0	0.003	0.042	0.055	0.071	0.563	7	0.129	18.540
	HKL	0	0.006	0.042	0.054	0.071	0.574	7	0.128	18.360
	SHE	0	0.0395	0.040	0.053	0.070	0.569	7	0.132	19.584
SMIE	KL	19	0.000	0.021	0.055	0.092	0.212	11	0.116	14.716
	PWKL	0	0.005	0.042	0.059	0.074	0.211	4	0.091	7.486
	HKL	0	0.008	0.043	0.056	0.074	0.210	5	0.091	7.476
	SHE	0	0.014	0.043	0.055	0.0694	0.210	11	0.093	7.906
RT	KL	0	0.000	0.032	0.052	0.085	0.223	1	0.104	16.304
	PWKL	0	0.002	0.003	0.044	0.077	0.222	3	0.116	14.763
	HKL	0	0.003	0.028	0.045	0.078	0.222	3	0.116	14.641
	SHE	0	0.010	0.042	0.57	0.081	0.222	2	0.090	6.962
MMPI	KL	0	0.064	0.065	0.066	0.067	0.086	0	0.067	0.039
	PWKL	0	0.064	0.065	0.066	0.067	0.083	0	0.067	0.041
	HKL	0	0.064	0.065	0.066	0.067	0.082	0	0.067	0.032
	SHE	0	0.065	0.065	0.065	0.067	0.073	0	0.067	0.009
RPG	KL	0	0.056	0.059	0.062	0.071	0.119	0	0.069	0.588
	PWKL	0	0.056	0.059	0.063	0.071	0.107	0	0.068	0.487
	HKL	0	0.056	0.059	0.063	0.071	0.106	0	0.068	0.492
	SHE	0	0.065	0.065	0.065	0.066	0.22	1	0.069	0.576

注: NU 表示没有使用的项目个数; OE 表示过度曝光的项目个数; OR 表示测验重叠率。

着 SM 方法仅增加少数几个项目的曝光率提高 SMIE 方法的判准率, 使得测验重叠率和 χ^2 统计量大于 SMIE 方法的结果, 从而降低了项目曝光均匀性。第五, 对 RT 方法而言, 结果与 Wang 等(2011)一致表明它能在一定程度上提高大部分项目的使用率并降低最大项目曝光率, 综合各项指标可知其曝光均匀性稍次于 SMIE 方法的结果(SHE 指标除外)。第六, SM 和 IE 方法在控制项目曝光率方面具有不同特点。首先, 当按相同指标选题时, SM 方法中未使用的项目个数和过度曝光的项目个数都一致小于 IE 方法的结果; 其次, SM 方法的题库利用率高达 90%, IE 方法的题库利用率约 50%; 再次, IE 方法的最大项目曝光率低于 0.23, 显著低于 SM 方法的最大项目曝光率。所以, SM 方法能提高大部分项目曝光率并适当降低最大项目曝光率, IE 方法只是稍稍降低原本曝光过高项目的使用率而没有显著提高项目曝光均匀性。于是, IE 方法的项目曝光均匀性不及 SM 方法。总之, 各方法项目曝光均匀性的优劣依次为: MMPI、RPG、SMIE、RT、SM 以及 IE 方法。

以 HKL 信息量选题为例, 图 1 为各方法按 HKL 信息量指标选题时项目曝光率分布曲线图。结合表 5 可知: (1) MMPI 与 RPG 方法的项目曝光率分布相近, 它们的项目曝光率分别集中在 [0.06, 0.09] 与 [0.05, 0.12] 之间; (2) SM 方法可以提高曝光过低项目的使用率, 但不能显著降低最大项目曝光率, 分布范围广且分散; (3) IE 方法能极大地降低最大项目曝光率, 但不能提高曝光过低项目的使用率, 题库利用率低; (4) SMIE 方法结合了 SM 与 IE 方法的优点, 曝光率分布比 SM 和 IE 方法更均匀, 但不及 MMPI 方法的表现; (5) RT 方法能较好地控制最大项目曝光率并明显提高曝光不足项目的使用率。对

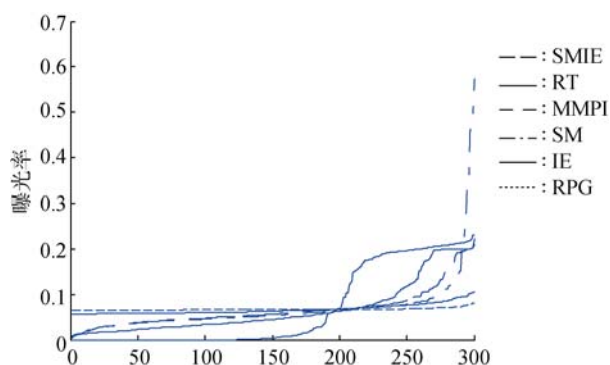


图 1 六种选题方法按 HKL 信息量选题时项目曝光率分布图

PWKL 和 HKL 而言, 其项目曝光均匀性介于 SMIE 和 SM 方法之间, 而对 KL 和 SHE 而言, RT 方法和 SMIE 方法的项目曝光率没有明显差异。

5 结论和讨论

CD-CAT 是认知诊断和自适应测验相结合的产物。它根据被试的特质水平适应性选题不仅实现了测验的“量体裁衣”, 还能反馈考生在各个属性水平上掌握与否的诊断信息, 便于教师因材施教进行补救教学。CAT 中测量精度是保证推论可靠性的重要条件, 要提高测量精度就需要施测最优项目, 这往往导致项目曝光不均匀。而建设和维护认知诊断题库需要花费大量成本, 所以研究者期望每个项目都得到充分利用、更好地发挥题库的作用。于是, 提高项目曝光均匀性和题库利用率就变得格外重要。

针对 CD-CAT 中项目曝光控制问题, 本文首先提出按项目信息量分层的选题方法, 然后比较六种项目曝光控制方法的选题表现。结果表明这些选题方法各有优势与不足。若强调测验安全可选用 MMPI 或 RPG 方法, 强调测量准确性可选用 IE 或 SM 方法。相比之下, SMIE 和 RT 方法能最好地平衡项目曝光控制和测量准确性要求。尽管本研究获得一些有意义的结果, 但是研究设计还有待改进。例如, 本文假设属性相互独立。事实上, 属性之间往往存在一定相关和具有层级关系。Leighton, Gierl 和 Hunka (2004) 指出属性之间主要存在线性型、收敛型、发散型和无结构型四种结构。因此, 有必要考察这些方法在不同属性相关和层级结构下的选题表现。

本文运用的项目曝光控制方法不局限于 DINA 模型, 因而可以应用到其它认知诊断模型, 比如规则空间方法和融合模型, 等等。尽管本研究表明 RT、SM 和 SMIE 方法具有很多优点, 但结论仍局限于模拟研究。因而, 建构认知诊断题库开展实证研究以检验上述选题方法的表现是下一步研究的重点。另外, 对以提供诊断信息为目的的测验而言, 则要求在提高测量精度的同时尽量减少测验项目。因而, 探索变长 CD-CAT 的选题方法同样具有重要意义。此外, 根据标准对测验结果进行分类的测验在国内占有重要地位, 考察这些方法在分类测验中的表现或探索分类测验的选题策略同样是今后研究的一个方向。最后, 由于参加 CAT 的考生作答不全相同的项目, 如何保证测验效度以提高测验结果的可靠性和可比性也是 CAT 选题需要考虑的重要问题。要提高测验效度就要求测验满足一些内容约

束,例如正确答案选项的分布均匀,不能同时包含具有提示信息的项目以及考查各个属性的项目个数要足够多等等。因而,探索满足内容约束的选题方法也是未来研究的重点。

参 考 文 献

- Chang, H. H., & Ying, Z. L. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211–222.
- Chang, H. H., & Zhang, J. M. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika*, 67, 387–398.
- Chen, P. (2011). *Item replenishing in cognitive diagnostic computerized adaptive testing –based on DINA model*. Unpublished doctoral dissertation, Beijing Normal University.
- [陈平. (2011). 认知诊断计算机化自适应测验的项目增补——以 DINA 模型为例. 博士学位论文, 北京师范大学.]
- Chen, P., Li, Z., & Xin, T. (2011). A note on the uniformity of item bank usage in cognitive diagnostic computerized adaptive testing. *Studies of Psychology and Behavior*, 9(2), 125–132.
- [陈平, 李珍, 辛涛. (2011). 认知诊断计算机化自适应测验的题库使用均匀性初探. *心理与行为研究*, 9(2), 125–132.]
- Chen, P., Xin, T., Wang, C., & Chang, H. H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, 77, 201–222.
- Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, 40, 129–145.
- 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.
- [程小杨, 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, 203–212.]
- Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74, 619–632.
- Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement*, 70, 902–913.
- Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 62, 369–383.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34, 115–130.
- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333–353.
- de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, 73, 595–624.
- Georgiadou, E. G., Triantafyllou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment*, 5, 4–38.
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.
- Henson, R. A. (2004). *Test discrimination and test construction for cognitive diagnosis*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258–272.
- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2, 359–375.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoaka's rule-space approach. *Journal of Educational Measurement*, 41, 205–237.
- Leung, C. K., Chang, H. H., & Hau, K. T. (2003). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. *Educational and Psychological Measurement*, 63, 257–270.
- Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, 35, 229–249.
- Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, 35, 311–327.
- Shang, Z. Y. & Ding, S. L. (2011). The exploration of item selection strategy of computerized adaptive testing for cognitive diagnosis. *Journal of Jiangxi Normal University (Natural Science)*, 35(4), 418–421.
- [尚志勇, 丁树良. (2011). 认知诊断自适应测验选题策略探新. *江西师范大学学报(自然科学版)*, 35(4), 418–421.]
- Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Rep. No. 94–95). Princeton, NJ: Educational Testing Service.
- Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.
- Tang, X. J., Ding, S. L., & Yu, Z. H. (2012). Application of computerized adaptive testing in cognitive diagnosis. *Advances in Psychological Science*, 20(4), 616–626.
- [唐小娟, 丁树良, 俞宗火. (2012). 计算机化自适应测验在认知诊断中的应用. *心理科学进展*, 20(4), 616–626.]
- Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Taylor & Francis Group.
- 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 Statistics*, 32, 398–418.
- Wang, C., Chang, H. H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, 48, 255–273.
- Wang, W. Y. (2009). *Research on item selection strategies of computerized adaptive testing—based on GRM and DINA*. Unpublished master's dissertation, Jiangxi Normal University.
- [汪文义. (2009). 计算机化自适应测验选题策略研究—以 GRM 和 DINA 模型为例. 硕士毕业论文, 江西师范大学.]

A Comparison of Item Selection Methods for Controlling Exposure Rate in Cognitive Diagnostic Computerized Adaptive Testing

MAO Xiuzhen^{1,2}; XIN Tao²

(¹ College of Education, Sichuan Normal University, Chengdu, 610066, China)

(² Institute of Developmental Psychology, Beijing Normal University, Beijing, 100875, China)

Abstract

Item exposure rate is the utilization frequency of an item. When the exposure rate is high, examinees will likely share item content. If there are too many over-exposed items, test security and hence the validity of the assessment will certainly be compromised. Furthermore, with a lot of under-exposed items having low or zero item-exposure rates, the manpower and financial resources spent on item construction will be wasted and the item pool construction will become more challenging. Item exposure control is, therefore, an important issue in computerized adaptive testing (CAT).

Cognitive diagnostic CAT (CD-CAT) combines and makes use of the strengths of cognitive diagnosis theory and CAT. The system will be able to provide information on the knowledge competence of the examinees by administering fewer items than traditional assessment. Based on the a -stratified method and the item eligibility method in regular CAT, the present study proposed and compared the performance of six techniques, namely, (a) the item eligibility (IE) method, (b) the stratified multistage (SM) approach, (c) the stratified multistage-item eligibility (SMIE) method, (d) the restrictive threshold (RT) method, (e) the maximum modified priority index (MMPI) method, and (f) the restrictive progress (RPG) method.

With noting it that the SM approach is similar to the a -stratified method in item selection steps. The SM approach, however, different with the a -stratified method firstly in that it stratifies the remaining item pool based on the values of item information at the estimated attributed mastery pattern while the a -stratified method is based on the values of item discrimination parameter a . Secondly, in the SM method, the remainder item bank are stratified into a number of levels before the selection of each item, whereas in the a -stratified method, the item pool is stratified only once before the test and all the examinees have the same item strata. The SMIE method combines the SM and the IE method.

MATLAB (R2010a) was used in the simulation experiments to write the CD-CAT code and the deterministic inputs, noisy "and" gate (DINA) model was applied in this study.

Results showed that: (a) the SM method used in CD-CAT produced widely distributed item exposure by increasing the exposure rates of most items and fully utilizing the item pool but without greatly diminishing the maximum exposure rate and measurement accuracy; (b) other than a few items, the exposure rates of the IE method were lower than the setting maximum exposure rate, but most items still had extremely low exposure rates and hence resulting in a narrow distribution of item exposure and the highest measurement precision; (c) SMIE and RT methods behaved similarly in that not only could they increase the utilization frequency of the under-exposed items but they could also decrease the maximum exposure rate to a certain extent; (d) the MMPI and the RPG methods performed similarly with almost evenly distributed item exposure but at the great sacrifice of the measurement precision.

As a whole, the performances of different methods in the order of their measurement accuracy are IE, SM, SMIE, RT, RPG and MMPI. The order in terms of their performances in exposure control is: MMPI, RPG, SMIE, RT, SM and IE. All in all, the SMIE and RT methods are able to balance measurement accuracy and item exposure well.

Key words cognitive diagnostic computerized adaptive testing; measurement accuracy; item exposure control; item selection method