

认知诊断 CAT 中具有非统计约束选题方法的比较*

毛秀珍¹ 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 项目曝光控制和内容约束关系到测验安全、测验的信度和效度, 是计算机化自适应测验 (Computerized Adaptive Testing, CAT) 中两类重要的非统计约束条件。本文在认知诊断 CAT 中针对内容约束和项目曝光控制要求, 运用 5 种方法选择测验项目。它们分别是: (1) Monte Carlo 方法与项目合格方法相结合, 记为 MC-IE; (2) Monte Carlo 方法与最大优先指标方法相结合, 记为 MC-MPI; (3) Monte Carlo 方法与限制阈值方法相结合, 记为 MC-RT; (4) Monte Carlo 方法与限制进度指标方法相结合, 记为 MC-RPG 以及 (5) Monte Carlo 方法与最大后验概率方法相结合, 记为 MC-PP。然后通过在线性、收敛、发散、无结构和独立五种属性结构下构建题库并运用重参化融融统和模型模拟被试反应比较它们的选题表现。研究发现, (1) 相同选题方法在不同属性结构下项目曝光率的分布类似, 测量精度按线性、收敛、发散、无结构和独立结构的顺序依次降低; (2) 相同属性结构下, 不同方法的测量精度高低依次为 MC-PP、MC-IE、MC-RT、MC-MPI 和 MC-RPG 方法; 项目曝光均匀性优劣依次为 MC-RPG、MC-MPI、MC-RT、MC-IE 和 MC-PP 方法。统一量纲值表明, MC-RPG 方法的综合表现最好, MC-MPI 方法的表现次之。

关键词 认知诊断理论; 计算机化自适应测验; 测量精度; 项目曝光率; 内容约束。

分类号 B841

1 引言

计算机化自适应测验 (Computerized Adaptive Testing, CAT) 的核心是基于被试在已作答项目上的反应估计其能力水平, 并根据选题策略连续地从剩余题库中选择最适合被试作答的项目施测被试, 直到满足终止规则。它与固定内容的测验相比在缩短测验长度的同时实现对被试能力水平的高效测量 (Chang & Ying, 1999)。

认知诊断评估根据被试的作答反应推测其是否掌握测验所测属性 (如任务、子任务、认知过程和技能等等), 从而提供更详细的诊断信息。从 2001 年美国法案 “No Child Left Behind” (NCLB) 到 2011 “Race to the top” 早期学习挑战经费的设立, 再到我国《国家中长期教育改革和发展规划纲要 (2010-2020) 年》都特别强调测验的认知诊断功能。认知诊断成为近期的研究热点, CAT 是一种新兴

的、有前途的测验形式。将二者结合在一起的 “认知诊断计算机化自适应测验 (Cognitive Diagnostic Computerized Adaptive Testing, CD-CAT)” 不仅能对学生在测验所测属性上的掌握水平进行分类, 而且提高了诊断评估的准确性和效率 (Cheng, 2009)。例如 Tatsuoka 和 Tatsuoka (1997) 采用 CD-CAT 施测分数加法测验证明它在测验结束能即时且较准确的判断考生是否掌握测验考查的每个属性 (例如 “从分数中分离出整数部分”、“通分”、“化简”), 估计考生的知识结构, 并能有效地指导补救教学。因此, CD-CAT 在教育测量领域得到越来越多的关注 (Chen, Xin, Wang, & Chang, 2012; Cheng, 2009; McGlohen & Chang, 2008; Xu, Chang, & Douglas, 2003; 林海菁, 丁树良, 2007; 涂冬波, 2009, 等等)。

同传统 CAT 一样, CD-CAT 包括题库、初始项目的选择、选题策略、特质估计方法和终止规则几个组成部分。其中选题策略决定每个被试作答的测

收稿日期: 2014-01-05

* 四川省教育厅自然科学基金 (13ZB0155)、国家自然科学基金青年项目 (31400897) 资助。

通讯作者: 辛涛, E-mail: xintao@bnu.edu.cn

验项目,它是影响测量准确性和有效性的重要因素之一。正如 Cheng 和 Chang (2009)以及 van der Linden (2000)指出 CAT 选题不仅要考虑统计优化(*statistical optimization*)问题,还应满足一些非统计约束(*nonstatistical constraints*)条件。统计优化指如何选题以提高测量精度。非统计约束主要包括项目曝光控制和内容约束。同传统 CAT 一样,非统计约束也是 CD-CAT 实践中需要考虑的重要问题。首先,项目曝光率即项目的使用频率。如果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。反之,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。其次,内容约束即要求测验的构成满足一些条件。CD-CAT 通常要求所考查的各个内容领域具有恰当的比例,项目呈现的顺序合理,正确答案选项的分布平衡、项目的长度适当,等等。如果违反这些约束将影响被试作答,从而降低测量准确性和有效性。例如,若测验的标准答案正好都是 A 或 C,显然可能导致学生猜测或者怀疑自己的答案并使其感到焦虑;又如,如果先施测的项目为其后选择的项目提供了作答线索必将提高被试正确作答项目的概率。由此可见,项目曝光均匀性和内容约束是影响测验安全、测验信、效度的重要因素。

一般地,统计优化和非统计约束之间相互制约,满足一些非统计约束往往会降低测量精度。因此,好的选题策略应在保证测验效度的前提下尽量提高测量精度和项目曝光均匀性。目前,CD-CAT 的

选题策略主要集中于提高测量精度(Cheng, 2009; McGlohen & Chang, 2008; Xu et al., 2003; Wang, 2013)、控制项目曝光率(Wang, Chang, & Huebner, 2011; 毛秀珍, 辛涛, 2013)和满足内容约束(Cheng, 2010; Mao & Xin, 2013)。另外, Hsu, Wang 和 Chen (2013)探讨了变长 CD-CAT 的终止规则和项目曝光控制问题。到目前为止,未曾有研究同时考虑内容约束和项目曝光控制。因此,本文借鉴 CD-CAT 的相关研究在同时具有内容约束和项目曝光控制条件下比较多种选题方法的表现。

2 认知诊断模型

认知诊断理论包括 3 个基本概念。首先,属性表示测验考查的知识、技能或策略,常用于描述项目和被试特征。根据属性之间的先决条件可将属性结构划分为图 1 所示的独立属性结构和 4 种层级结构:线性型、收敛、发散和无结构(Leighton, Gierl, & Hunka, 2004)。其次,知识状态(*Knowledge States*, KS)指的是根据属性层级关系确定的所有符合逻辑的属性掌握模式;项目属性模式(*Attribute Pattern of Item*, API)指根据属性层级关系确定的所有符合逻辑的项目考核模式。运用规则空间方法(Tatsuoka, 2009)可分析不同属性结构下所有可能的 KS 和 API。例如,当测验考查 2 个独立属性时,所有 KS 为(0, 0)、(0, 1)、(1, 0)和(1, 1),所有 API 为(0, 1)、(1, 0)和(1, 1)。一般地,测验考查 K 个独立属性时,一共有 2^K 类 KS,有 $2^K - 1$ 类 API。当属性间存在层级关系时,KS 的总数远远小于 2^K 。

第三,矩阵 $Q_{K \times N}$ 表示项目与属性的关系, K 和

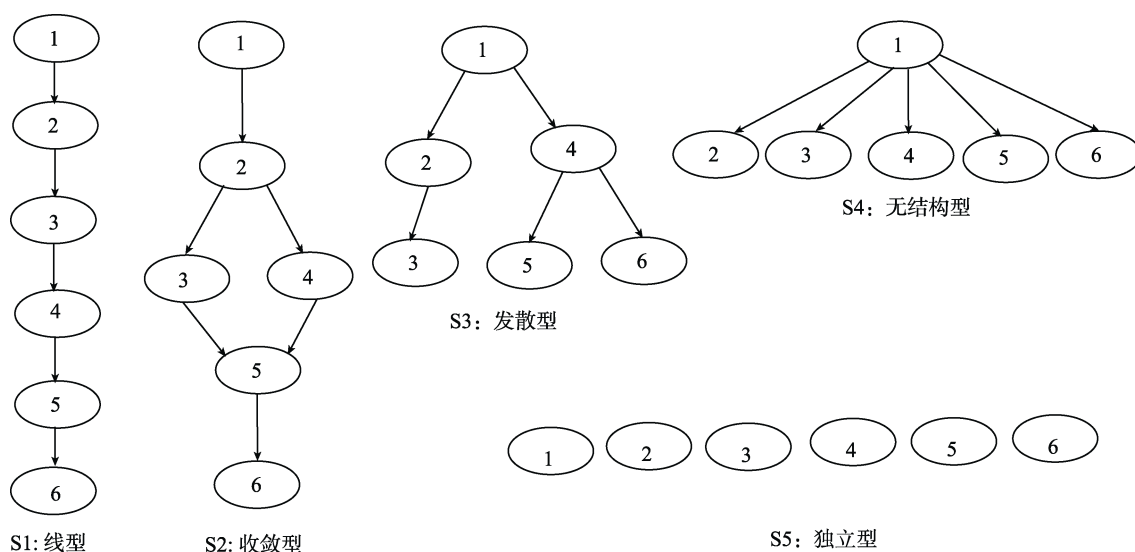


图 1 六个属性为例的 5 种属性结构

N 分别代表属性与项目的个数。其元素 $q_{ki} (k=1, \dots, K; i=1, \dots, N)$ 只取 1 或 0, 如果项目 i 考查属性 k , 则 $q_{ki}=1$, 反之 $q_{ki}=0$ 。再次, 被试 j 的 KS 通常表示为 K 维 0-1 向量 $\alpha_j = (\alpha_{j1}, \alpha_{j2}, \dots, \alpha_{jK})$ 。如果他(她)掌握属性 k , 则 α_{jk} 为 1, 否则为 0。

大部分认知诊断模型通过数理公式刻画项目特征、被试特质水平和作答反应的关系。例如, 重参化统合模型(*The Reparameterized Unified Model*, RUM) (Hartz, 2002)表示为

$$P\left(\prod_{k=1}^K Y_{ijk} = 1 \mid \alpha_j, \theta_j, \pi_i^*, r_{ik}^*, c_i\right) = \pi_i^* \cdot \prod_{k=1}^K r_{ik}^{*(1-\alpha_{jk})q_{ik}} \cdot P_{c_i}(\theta_j) \quad (1)$$

若令 π_{ik} 和 r_{ik} 分别表示被试掌握和未掌握属性 k 时在项目 i 上正确运用该属性的概率, 即 $\pi_{ik} = P(Y_{i \bullet k} = 1 \mid \alpha_{\bullet k} = 1)$ 和 $r_{ik} = P(Y_{i \bullet k} = 1 \mid \alpha_{\bullet k} = 0)$ 。它们的其值不随被试的不同而改变。那么, $\pi_i^* = \prod_{k=1}^K \pi_{ik}^{q_{ik}}$, 它表示掌握项目 i 要求的所有属性时, 正确作答这些属性的概率。 $r_{ik}^* = r_{ik} / \pi_{ik}$, r_{ik}^* 越小, 表明没有掌握属性 k 会极大地降低正确作答项目 i 的概率。因此, r_{ik}^* 通常称为项目 i 中属性 k 的区分度参数。式(1)中 $P_{c_i}(\theta_j)$ 可通过 Rasch 模型计算, c_i 表示项目难度, θ_j 表示影响测验反应但不能被 Q 矩阵解释的那部分潜特质的水平。实际上, 大部分研究都假设 Q 矩阵完备, 即 $P_{c_i}(\theta_j)$ 等于 1 (Hsu et al., 2013; McGlohen & Chang, 2008; Wang et al., 2011)。

RUM 中同一被试正确作答包含相同属性的不同项目的概率 P 不同。另外, 项目考察的属性越多, 它能区分更多不同 KS 的被试正确作答该项目的概率 P 。Templin, Henson 和 Templin (2008)指出决定性输入, 噪音‘与’门(*the deterministic inputs, noisy ‘and’ gate*, DINA)模型对每个项目只能将被试分为两类。一类被试掌握了项目考查的所有属性, 其余为另一类。他们还指出噪音输入, 决定性‘与’门(*the noisy inputs, deterministic ‘and’ gate*, NIDA)模型中同一被试正确作答包含相同属性的不同项目的概率相等。由此可见, 与 DINA 和 NIDA 模型相比, RUM 更符合实际情况。另外, RUM 用于分析实际数据时具有较高的诊断效度(Wang et al., 2011)。此外, 它还常应用于 CD-CAT 的相关研究(Hsu et al., 2013; McGlohen & Chang, 2008; Wang et al., 2011)。于是, 本文选用 RUM, 并假设 $P_{c_i}(\theta_j)$ 等于 1。

3 CD-CAT

本文将在同时具有内容约束和项目曝光控制条件下探讨项目选择方法。

3.1 CD-CAT 中具有内容约束的选题方法

CD-CAT 中有关内容约束的选题策略主要借鉴了传统 CAT 的相关研究。例如, 基于传统 CAT 中约束加权 a -分层方法, Cheng (2010)提出最大修正全局区分度指标(*Maximum Modified Global Discrimination Index*, MMGDI)选题方法。又如, Mao 和 Xin (2013)将传统 CAT 中 Monte Carlo 方法(Belov, Armstrong, & Weissman, 2008)应用到 CD-CAT 以满足内容约束, 并指出(1) Monte Carlo 方法在满足测验约束的条件下具有较高的测量精度, 还能提高大部分项目的曝光率但不能明显降低最大项目曝光率;(2)在测量精度、项目曝光均匀性和题库利用率方面, Monte Carlo 方法的表现都比 MMGDI 方法的结果更好。因此, 本研究将运用 Monte Carlo 方法以满足测验内容约束。

Monte Carlo 方法的核心是在选择每个项目之前均匀组合多份影子测验, 使它们不仅包含已施测项目而且满足所有内容约束, 然后在这些测验所有未施测的项目中依据信息量指标选择测验项目。在选择第 $t+1(t=0, 1, \dots, L-1, L$ 表示测验长度)个项目之前, 假设(1)知识状态的估计值为 $\hat{\alpha}'$; (2) 参数 m 表示在选择第 $t+1$ 个项目时要求组合的影子测验的个数;(3) 集合 S 包含 r 个影子测验。那么, 选择第 $t+1$ 个项目包括以下几个步骤。第一, 均匀组合 $m-r$ 个影子测验, 将它们并入集合 S 。均匀组卷要求选择每个影子测验的概率相等, 它可以通过保留随机组合的包含已施测项目且满足约束条件的那些测验得到。第二, 提取 S 中每个测验所有未施测的项目组成一个序列, 一共有 $m(L-t)$ 个项目, 它可能包含重复项目; 第三, 从该项目序列中选择在 $\hat{\alpha}'$ 处具有最大信息量的项目作为第 $t+1$ 个项目; 第四, 去掉 S 中不包含第 $t+1$ 个项目的那些测验后得到选择下一个项目的初始集合 S 。

3.2 CD-CAT 中控制项目曝光率的选题方法

针对 CD-CAT 中项目曝光控制问题, Wang 等 (2011)在传统 CAT 研究基础上提出限制阈值方法(*the Restrictive Threshold method*, RT)和限制进度指标方法(*the Restrictive Progressive Method*, RPG)。他们运用 RUM 并假设 4 个独立属性条件下的研究表明, RT 方法的测量精度高于 RPG 方法的结果, 但其

项目曝光均匀性不及 RPG 方法。总体上, 它们均能较好地控制最大项目曝光率并提高曝光不足项目的使用率。

另外, 毛秀珍和辛涛(2013)在独立属性结构下基于 DINA 模型比较了项目合格方法 (*Item Eligibility*, IE)、分层方法、修正优先指标 (*Modified Priority Index*, MPI) 方法、RT 和 RPG 方法在控制项目曝光率方面的表现。结果表明, IE、MPI、RT 和 RPG 方法均能在一定程度上控制最大项目曝光率。

3.3 CD-CAT 中具有非统计约束的选题方法

基于已有研究, 当测验同时要求内容约束和曝光控制时, 本文运用下面 5 种方法选题。

3.3.1 Monte Carlo 方法与 IE 相结合, 记为 MC-IE 方法

该方法在每个被试参加测验之前运用 IE 方法 (van der Linden & Veldkamp, 2004) 得到所有有资格向该被试施测的项目构成影子题库。然后, 运用 Monte Carlo 方法在该影子题库中选择测验项目。其中, IE 方法通过以下几步确定项目 i ($i=1, 2, \dots, N$, N 表示题库容量) 是否有资格施测给被试 $j+1$ ($j=0, 1, 2, \dots, M-1$, M 表示被试总人数)。首先令 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 的关系为 $A_i \subseteq E_i$, 并且 $P(A_i) = P(A_i | E_i) \cdot P(E_i)$ 。其次计算项目 i 对被试 $j+1$ 合格的概率 $P^{j+1}(E_i)$ 。假设最大项目曝光率为 $r^{\max}$, 那么

$$P(E_i) \leq \frac{r^{\max}}{P(A_i | E_i)} = \frac{r^{\max} \cdot P(E_i)}{P(A_i)}。 \quad (2)$$

式(2)是可重复的, 即项目 i 对被试 $j+1$ 合格的概率为 $P^{j+1}(E_i) \leq \frac{r^{\max} \cdot P^{1, \dots, j}(E_i)}{P^{1, \dots, j}(A_i)}$, 等价于

$$P^{j+1}(E_i) = \min \left\{ \frac{r^{\max} \cdot P^{1, \dots, j}(E_i)}{P^{1, \dots, j}(A_i)}, 1 \right\}。 \quad (3)$$

其中, $P^{1, \dots, j}(E_i)$ 和 $P^{1, \dots, j}(A_i)$ 分别表示前 j 个被试中项目 i 是合格和施测项目 i 的频率。为保证式(3)有意义, 当前 j 个被试中施测项目 i 的频率为 0 时都有 $P^{j+1}(E_i) = 1$ 。最后, 若 $P^{j+1}(E_i)$ 大于 $[0, 1]$ 的随机数 v_i , 则项目 i 有资格施测给被试 $j+1$, 反之亦然。

3.3.2 Monte Carlo 方法与 MPI 方法相结合, 记为 MC-MPI 方法

若将 MMGDI 方法 (Cheng, 2009) 中属性平衡条件换成最大项目曝光率约束, 容易得到项目 i 的

MPI 为:

$$MPI_i = \frac{r^{\max} - n_i / M}{r^{\max}} \cdot Index_i。 \quad (4)$$

其中, $r^{\max}$ 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。“*Index*”表示项目信息量指标, 可采用 Kullback-Leibler (KL) 信息量、后验加权 KL (*Posterior Weighted Kullback-Leibler*, PWKL)、结合后验和距离加权的混合 KL (*Hybrid Kullback-Leibler*, HKL) 和 Shannon 熵 (SHE) 信息量计算。特别地, KL 和 PWKL 的计算方法如下 (Cheng, 2009):

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

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

其中 $\pi_{i,t}(\alpha_c) = \pi_{i,0}(\alpha_c) L(u^{t-1} | \alpha_c)$, $\pi_{i,0}(\alpha_c)$ 表示知识状态 α_c 的先验概率, $L(u^{t-1} | \alpha_c)$ 为似然函数, 令 u_i 代表项目 i 的作答反应, 那么 $L(u^{t-1} | \alpha_c) = \prod_{i=1}^{t-1} [P_i(\alpha_c)]^{u_i} [1 - P_i(\alpha_c)]^{1-u_i}$ 。

于是, MC-MPI 方法在选择第 $t+1$ ($t=0, 1, 2, \dots, L-1$) 个项目之前运用 Monte Carlo 方法得到候选项目序列 U_{t+1} , 然后在 U_{t+1} 中选择使 MPI 值最大的项目。

3.3.3 Monte Carlo 方法与 RT 方法相结合, 记为 MC-RT 方法

该方法在每个被试参加测验之前将曝光率大于预设最大值的那些项目从题库中去掉后形成一个影子题库。当选择第 $t+1$ ($t=0, 1, 2, \dots, L-1$) 个项目时, 首先在该影子题库中运用 Monte Carlo 方法得到候选项目序列 U_{t+1} , 然后在 U_{t+1} 中运用 RT 方法选择测验项目。具体而言, RT 方法从 U_{t+1} 中信息量落在最大信息量的一个较小区间 $[\max(Index) - \delta, \max(Index)]$ 内的项目集中随机选取一个项目。其中 $\delta = [\max(Index) - \min(Index)] * (1 - t/L)^\beta$, 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。

3.3.4 Monte Carlo 方法与 RPG 相结合, 记为 MC-RPG 方法

MC-RPG 方法在选择第 $t+1$ ($t=1, 2, \dots, L-1$)

个项目时, 首先运用 Monte Carlo 方法生成候选项目序列 U_{t+1} , 然后运用 RPG 方法依下式选择测验项目。

$$i_t = \arg \max \{ (1 - e_i / \bar{e}_i) [(1 - t/L)R_i + Index_i \times \beta t / L], i \in S_t \}. \quad (7)$$

e_i 与 $\bar{e}_i$ 分别表示项目 i 的曝光率和期望项目曝光率, 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 $R_i \sim U(0, H^*)$ 。同 RT 方法一样, β 越大, 测量越准确, 项目曝光均匀性越差。

3.3.5 Monte Carlo 方法与知识状态最大后验概率方法相结合, 记为 MC-PP 方法

变长 CAT 与定长 CAT 相比, 能在较短平均测验长度的条件下使每个被试的测量达到预定精度。若在每个被试的测量达到预设精度之后施测曝光率较小的项目可以控制项目曝光均匀性。于是, 本文将知识状态最大后验概率 $pp_{\max}$ (Hsu et al., 2013) 作为测量精度的标准, 并与 Monte Carlo 方法结合选题。具体来讲, 在选择第 $t+1$ ($t = 1, 2, \dots, L-1$) 个项目之前, 判断所有知识状态后验概率的最大值是否达到预定要求, 一旦达到预设值以后各题均在 Monte Carlo 方法选出的候选项目序列中具有较小项目曝光率的 h 个项目中随机选择, 否则选择具有最大项目信息量的项目。

MC-IE 方法运用 IE 方法在测验之前过滤项目的方式控制最大项目曝光率, MC-MPI 和 MC-RPG 方法在 Monte Carlo 方法选出的项目中控制最大项目曝光率。根据以往研究 (van der Linden & Veldkamp, 2004; Mao & Xin, 2013) 推知, 它们均可能在一定程度上违反最大项目曝光率的约束。

4 研究 1: 同时具有内容约束和曝光控制条件下选题方法的比较

研究一运用 RUM 分别在 5 种属性结构下以 Matlab (R2010a) 为工具编写 CAT 代码开展模拟研究, 比较 MC-IE、MC-MPI、MC-RT、MC-RPG 和 MC-PP 的选题表现。

4.1 研究方法

4.1.1 题库的模拟方法

在线性型(S1)、收敛(S2)、发散(S3)、无结构(S4)和独立(S5)五种属性结构下模拟题库。借鉴 Wang 等(2011)的方法, 每个题库包含 480 个试题, 一共考查 6 个属性。根据图 1 可知在 S1、S2、S3、S4 和 S5 五种结构下分别有 6、7、15、32 和 63 类 API。

于是, 不同属性结构的项目从所对应的 API 中产生, 随之可得 Q 矩阵。具体地, 对 S3、S4 和 S5 属性结构, 题库中各种不同 API 的项目服从均匀分布。由表 1 可知这 3 种属性结构下题库中考查各个属性的项目数均大于 170。特别地, S1 结构中所有 API 为 (1,0,0,0,0,0)、(1,1,0,0,0,0)、(1,1,1,0,0,0)、(1,1,1,1,0,0,0)、(1,1,1,1,1,0)和(1,1,1,1,1,1); S2 结构中所有 API 为 (1,0,0,0,0,0,0)、(1,1,0,0,0,0,0)、(1,1,1,0,0,0,0)、(1,1,0,1,0,0,0)、(1,1,1,1,0,0,0)、(1,1,1,1,1,0)和(1,1,1,1,1,1)。于是, S1 和 S2 结构中只有 API 为(1,1,1,1,1,1)的项目才考查第六个属性。因此, 若假设各种 API 服从均匀分布, 将导致考查第六个属性的项目个数远小于 100。考虑到测验内容约束, 为消除题库不充分对选题策略的影响, 将 S1 和 S2 结构中 API 为(1,1,1,1,1,1)的项目类在题库中的比例调高至 30%, 从而使考查第六个属性的项目个数在 150 左右, 其余类型的项目服从均匀分布。

采用 Wang 等(2011)和 Hsu 等(2013)的方法, 假设 RUM 中项目参数的分布为 $\pi_i^* \sim U(0.75, 0.95)$ 和 $r_{ik}^* \sim U(0.2, 0.95)$ ($i=1, 2, \dots, 480$; $k=1, 2, \dots, 6$)。

4.1.2 被试的模拟方法

在属性掌握模式分布未知的前提下, 本研究假设 KS 服从均匀分布, 并在每种属性结构下模拟 2000 个被试。运用 RUM 计算被试 j ($j = 1, 2, \dots, 2000$) 正确作答项目 i ($i = 1, 2, \dots, 480$) 的概率 P_{ij} , 然后产生 $[0, 1]$ 区间的随机数 s_{ij} 。如果 P_{ij} 大于 s_{ij} , 则被试 j 在项目 i 上的反应为 1, 否则为 0。对 Q 矩阵和被试知识状态的真值描述统计的结果见表 1 和表 2。

4.1.3 项目选择方法

分别在 5 种题库中运用 MC-IE、MC-MPI、MC-RT、MC-RPG 和 MC-PP 方法选题。Cheng (2009) 研究表明, PWKL 的测量精度优于 KL、HKL 和 SHE。另外, PWKL 要计算知识状态的后验概率, 于是各种方法除了在选择第一个项目时采用 KL 作为信息量指标外, 其它项目均采用 PWKL 作为信息量指标。考虑到在测验初期增加候选项目的个数可以提高测量精度, 测验末期减少候选项目的个数对测量精度的影响不大, 本文对 Mao 和 Xin (2013) 中 Monte Carlo 方法中参数 m 的设置进行调整, 即选择第 t ($t < L-3$) 个项目时, 令 $m = \max \{ [(2t+1)/(L-t)], 3 \}$, 即 $(2t+1)/(L-t)$ 的值四舍五入之后与 3 求最大; 当选择最后 3 个项目时, 取 $m=L$, L 为测验长度。此外, Monte Carlo 方法本身可以在一定程度上提高曝光过低项目的使用率, 所以在 MC-RT 和 MC-RPG 方

表 1 不同属性结构下考察(掌握)各个属性的项目个数(被试人数)

统计对象	属性结构	属性					
		A1	A2	A3	A4	A5	A6
项目	S1	480	413	355	286	223	159
	S2	480	420	299	309	184	147
	S3	480	335	168	377	185	177
	S4	480	237	241	240	234	238
	S5	248	248	248	239	221	255
被试	S1	1720	1444	1141	867	580	301
	S2	1743	1484	999	990	507	251
	S3	1893	1252	638	1507	762	751
	S4	1938	964	941	991	996	970
	S5	1022	992	974	1024	972	1023

注: S1: 线性型; S2: 收敛型; S3: 发散型; S4: 无结构型; S5: 独立型(下同)

表 2 不同属性结构下考查(掌握)不同属性个数的项目个数(被试人数)

统计对象	属性结构	属性个数						
		0	1	2	3	4	5	6
项目	S1	0	67	58	69	63	64	159
	S2	0	60	53	126	57	37	147
	S3	0	26	70	133	129	91	31
	S4	0	8	70	168	159	68	7
	S5	0	40	112	164	120	41	3
被试	S1	280	276	303	274	287	279	301
	S2	257	259	248	483	246	256	251
	S3	107	135	235	525	503	359	136
	S4	62	66	311	592	578	322	69
	S5	19	201	441	669	462	179	29

法中为使测量精度不显著受到项目曝光控制的影响, 令 β 等于 2。最后, MC-PP 方法中 $pp_{\max}$ 设置为 0.8, h 等于 10。在 5 种属性结构下分别运用 5 种方法选题且每种实验条件下模拟一次实验, 一共进行了 25 次模拟实验。

4.1.4 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, 以后采用极大似然(MLE)方法估计被试的知识状态。

4.1.5 测验要求

借鉴以往研究(Finkelman, Kim, Roussos, & Verschoor, 2010; Mao & Xin, 2013)本文假设测验满足以下条件: (1) 测验长度为 25; (2) 考查每个属性的项目不少于 4 个; (3) 有 4 种正确答案选项, 每个选项的项目个数介于 5 个到 9 个之间; (4) 随机确定第 49 题和第 274 题为“敌对项目”, 即所有测验不能同时包含这两个项目。另外, 最大项目曝光率不

超过 0.2。

4.1.6 评价指标

采用单个属性和 KS 的判准率衡量测量结果的准确性。其中, 第 k 个属性的判准率为第 k 个属性被正确判断为掌握或没有掌握的被试人数 l_k 与被试总人数 M 之比, 即 l_k / M ; 模式判准率为被试的 KS 得到正确判断的被试人数 f 与 M 之比, 即 f / M 。

本文选用项目曝光率的四分位数、题库利用率、过度曝光项目(即曝光率大于 0.2)的比例、 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, χ^2 统计量表示项目观察曝光率 e_i 和期望曝光率 $\bar{e}_i$ 之间的差异。 $\bar{e}_i$ 等于测验长度 L 除以题库项目数量 N , 即 $\chi^2 = \sum_{i=1}^N [(e_i - \bar{e}_i)^2 / \bar{e}_i]$ 。其值越小, 总体上项目曝光越均匀。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。本文运用公式(8) (Chen, Ankenmann, & Spray, 2003)计算测验重叠率,

$$\hat{T} = \frac{L}{N} S_{er}^2 + \frac{N}{L} \quad (8)$$

其中, 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。

最后, 采用统一量纲再加权求和的方法(陈平, 丁树良, 林海菁, 周婕, 2006; 刘珍, 丁树良, 林海菁, 2008)评价各个选题策略在测量精度和曝光均匀性的综合表现。具体做法是: 对值越大越好的指标, 将该指标的最大值作为分母, 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所作为分子, 把各选题策略在该指标上的值作为分母, 求两者的比值。统一量纲后, 对每个选题策略的 5 个评价指标(属性掌握模式的判准率, 题库利用率, 过度曝光项目的比例, 测验重叠率和 χ^2)分别赋予相等的加权系数。加权求和值越大, 表明该选题策略在这

几个方面综合表现越好, 反之亦然。

4.2 研究结果

4.2.1 单个属性和 KS 的判准率

由表 3 可知, 固定属性结构时(1)所有选题方法对单个属性的判准率较高, 例如在 S1 和 S2 结构中单个属性的判准率均在 0.99 以上; (2) 除 S5 结构外, 不同方法对单个属性的判准率之差小于 0.03; (3) 不同方法对属性掌握模式的判准率高低依次为 MC-PP、MC-IE、MC-RT、MC-MPI 和 MC-RPG 方法。相同方法在 S1、S2、S3、S4 和 S5 结构下的 KS 判准率依次降低。在 S1、S2、S3、S4 和 S5 结构下 API 的个数依次增加, 题库中考查各类 API 的项目数量依次减少, 即题库的充分性相对降低, 这可能是导致各种方法 KS 判准率逐渐降低的原因之一。另外, 对 S1、S2 和 S3 结构, MC-PP、MC-IE、MC-RT 和 MC-MPI 的判准率差异不大, 它们 KS 的判准率均大于 0.94; MC-RT 和 MC-MPI 方法在各种属性结构下的测量精度没有差异。此外, MC-PP 方

表 3 各种选题方法在不同属性结构下的判准率结果

属性结构	选题方法	属性						KS
		A1	A2	A3	A4	A5	A6	
S1	MC-PP	0.999	0.999	0.997	0.999	0.998	1.000	0.991
	MC-IE	1.000	0.998	0.997	0.998	0.999	1.000	0.991
	MC-RT	0.999	0.997	0.997	0.998	1.000	1.000	0.991
	MC-MPI	0.998	0.995	0.999	0.999	1.000	1.000	0.991
	MC-RPG	0.995	0.995	0.996	0.995	0.997	0.999	0.977
S2	MC-PP	0.999	0.997	0.997	0.997	1.000	1.000	0.988
	MC-IE	0.998	0.999	0.995	0.998	0.998	1.000	0.988
	MC-RT	0.999	0.998	0.998	0.997	0.996	1.000	0.988
	MC-MPI	1.000	0.996	0.995	0.992	0.997	0.999	0.981
	MC-RPG	0.999	0.993	0.994	0.994	0.995	0.999	0.974
S3	MC-PP	0.998	0.994	0.996	0.996	0.988	0.994	0.969
	MC-IE	1.000	0.991	0.995	0.993	0.984	0.987	0.953
	MC-RT	0.999	0.985	0.992	0.993	0.977	0.987	0.941
	MC-MPI	0.998	0.986	0.987	0.995	0.977	0.984	0.937
	MC-RPG	1.000	0.981	0.984	0.988	0.969	0.977	0.916
S4	MC-PP	0.998	0.978	0.971	0.971	0.962	0.971	0.878
	MC-IE	0.996	0.964	0.967	0.967	0.962	0.969	0.865
	MC-RT	0.992	0.966	0.963	0.965	0.952	0.964	0.847
	MC-MPI	0.995	0.967	0.956	0.966	0.953	0.960	0.842
	MC-RPG	0.997	0.963	0.949	0.945	0.946	0.946	0.816
S5	MC-PP	0.955	0.952	0.959	0.954	0.962	0.952	0.798
	MC-IE	0.956	0.945	0.945	0.938	0.943	0.948	0.751
	MC-RT	0.940	0.924	0.952	0.939	0.942	0.938	0.731
	MC-MPI	0.929	0.943	0.944	0.920	0.931	0.942	0.727
	MC-RPG	0.932	0.917	0.934	0.917	0.914	0.929	0.681

表 4 无结构型和独立型属性结构下各种方法 KS 判准率的描述统计

属性结构	选题方法	各种知识状态判准率				M1	M2
		最小	最大	平均	标准差		
S4	MC-PP	0.698	1	0.873	0.082	2.3	4.182
	MC-IE	0.667	1	0.855	0.08	2.4	4.09
	MC-RT	0.673	0.983	0.834	0.083	2.7	4.273
	MC-MPI	0.667	0.971	0.825	0.084	2.6	4.273
	MC-RPG	0.508	0.986	0.798	0.109	2.4	4.182
S5	MC-PP	0.526	0.96	0.793	0.104	1.7	4.182
	MC-IE	0.519	1	0.736	0.112	1.9	4.455
	MC-RT	0.432	0.969	0.722	0.136	1.8	4.455
	MC-MPI	0.440	1	0.710	0.129	2.2	4.546
	MC-RPG	0.414	0.967	0.667	0.136	1.8	4.455

注: M1 和 M2 分别表示判准率最小(大)的 10 类知识状态平均掌握的属性个数。

法在各种属性结构下 KS 的判准率都大于等于 0.8, 这表明该方法在达到测量精度的条件下再控制项目曝光。

为进一步分析 S4 和 S5 结构中 KS 判准率下降的原因, 表 4 统计了这两种属性结构下各类 KS 的判准率。结果表明, 每种方法对不同 KS 的判准率都存在较大差异。例如在 S4 结构下最小判准率和最大判准率的差异达 0.3 以上; 在 S5 中最小值和最大值之间相差超过 0.4; MC-RT、MC-MPI 和 MC-RPG 方法在 S5 结构下最小值和最大值之间的差异超过 0.5。若记判准率最小(大)的 10 类 KS 平均掌握的属性个数为 M1(M2)。由表 4 可知, M1 明显小于 M2。这表明, 在 S4 和 S5 结构下掌握属性个数较少的那些 KS 的判准率较低, 掌握属性个数较多的那些 KS 的判准率更高。造成这些结果的原因可能与 API 和 KS 的分布、属性结构以及模型等多种因素有关。例如, S4 和 S5 结构中 KS 总数远远大于 S1、S2 和 S3 的 KS 总数, 从而使 S4 和 S5 中各类 KS 的归类更加困难和判准率下降。另外, RUM 中, 项目考查的每个属性都会影响正确作答概率, 从表 2 可知大部分项目考查 2、3 或 4 个属性, 因此掌握较少属性的被试正确作答项目的不确定性可能比掌握更多属性的被试的不确定性高, 进而使其判准率下降。特别地, S4 和 S5 中掌握属性个数较少的那些 KS 的判准率更低的原因还有待进一步研究证实。

4.2.2 项目曝光率结果

表 5 列出各种方法在不同属性结构下项目曝光率的结果。总体上, 相同方法在不同属性结构下项目曝光率分布一致, 即各种方法项目曝光均匀性不受属性结构的影响; 不同方法在相同属性结构下的

表现具有明显差异。具体来讲, 相同属性结构下: (1) MC-PP 方法中题库利用率最低(除 S2 和 S3 外), 过度曝光项目的比例、测验重叠率和 χ^2 值最大。因此, MC-PP 方法的项目曝光均匀性最差。(2) MC-IE 方法中题库利用率(除 S2 和 S3 外)较 MC-PP 方法更高, 过度曝光项目的比例、测验重叠率和 χ^2 值显著小于 MC-PP 方法的结果, 总体上其曝光均匀性比 MC-PP 方法的结果更好。(3) MC-RT 和 MC-MPI 中过度曝光项目的比例、测验重叠率和 χ^2 值依次减少, 且均小于 MC-IE 方法的结果, MC-MPI 题库利用率稍低于 MC-RT。总体上, MC-MPI 方法的项目曝光分布比 MC-RT 方法更均匀。(4) MC-RPG 方法在所有属性结构下的题库利用率达 100%, 过度曝光项目的个数等于 1, 测验重叠率和 χ^2 值显著小于其他方法, 因此其项目曝光分布最均匀。总之, 各种方法项目曝光均匀性的优劣次序为 MC-RPG、MC-MPI、MC-RT、MC-IE 和 MC-PP 方法。同时考虑测量精度和项目曝光控制时, 根据统一量纲值可以发现 5 种方法综合表现的优劣依次为 MC-RPG、MC-MPI、MC-RT、MC-IE 和 MC-PP 方法。

5 研究 2: 仅满足内容约束或控制曝光率条件下选题方法的比较

5.1 研究方法

由表 3 可知, S1 和 S2 中同时具有内容约束和曝光控制的测量精度均达到 0.97 以上。于是, 当仅满足一种非统计约束时, S1 和 S2 中各方法的测量精度不会存在明显差异。另外, S4 中有 4 个属性是独立的, 意味着 S4 和 S5 结构很相似。因此, 研究

表 5 各种选题方法在不同属性结构下项目曝光率结果

属性结构	选题方法	曝光率				UR	NOR	OR	χ^2	统一量纲
		Q_1	Q_2	Q_3	max					
S1	MC-PP	0.003	0.021	0.080	0.496	0.898	0.044	0.146	44.786	2.425
	MC-IE	0.004	0.022	0.086	0.214	0.900	0.027	0.127	35.723	2.527
	MC-RT	0.011	0.034	0.081	0.355	0.952	0.010	0.106	25.821	2.815
	MC-MPI	0.006	0.042	0.091	0.333	0.938	0.004	0.098	21.838	3.156
S2	MC-RPG	0.027	0.042	0.071	0.320	1.000	0.002	0.075	11.018	3.926
	MC-PP	0.004	0.021	0.079	0.447	0.908	0.052	0.142	43.086	2.438
	MC-IE	0.004	0.024	0.087	0.224	0.898	0.021	0.124	34.268	2.557
	MC-RT	0.013	0.036	0.075	0.309	0.940	0.010	0.104	24.619	2.813
S3	MC-MPI	0.090	0.043	0.089	0.281	0.906	0.004	0.093	19.448	3.152
	MC-RPG	0.028	0.044	0.071	0.267	1.000	0.002	0.072	9.729	3.962
	MC-PP	0.003	0.018	0.010	0.593	0.898	0.063	0.175	59.971	2.309
	MC-IE	0.004	0.022	0.083	0.214	0.890	0.038	0.134	39.019	2.429
S4	MC-RT	0.010	0.031	0.074	0.314	0.956	0.013	0.117	31.037	2.660
	MC-MPI	0.009	0.043	0.087	0.289	0.931	0.004	0.093	19.370	3.133
	MC-RPG	0.028	0.045	0.071	0.272	1.000	0.002	0.073	9.883	3.890
	MC-PP	0.002	0.021	0.069	0.575	0.869	0.056	0.176	59.096	2.190
S5	MC-IE	0.002	0.027	0.081	0.213	0.917	0.027	0.129	36.936	2.409
	MC-RT	0.078	0.032	0.078	0.272	0.958	0.017	0.112	28.911	2.558
	MC-MPI	0.010	0.051	0.087	0.228	0.917	0.002	0.088	17.387	3.566
	MC-RPG	0.028	0.046	0.074	0.210	1.000	0.002	0.070	8.478	3.832
	MC-PP	0.002	0.017	0.065	0.750	0.838	0.054	0.201	71.224	2.030
	MC-IE	0.002	0.002	0.081	0.224	0.854	0.031	0.131	37.845	2.213
	MC-RT	0.009	0.029	0.077	0.362	0.944	0.017	0.120	32.400	2.385
	MC-MPI	0.011	0.048	0.085	0.317	0.913	0.004	0.090	18.069	2.928
	MC-RPG	0.029	0.046	0.070	0.295	1.000	0.002	0.071	8.948	3.681

注：UR：题库利用率；NOR：曝光率大于 0.2 的项目比例；OR：测验重叠率。

二假设属性结构为 S3 和 S4 时比较不同约束条件对选题的影响。另外，研究中二采用的题库和被试、知识状态的估计方法和评价指标与研究一相同。

当测验只要求内容约束时，约束条件同 4.1.5，选题方法采用 Monte Carlo 方法；当测验只要求控制项目曝光率时，选题方法为 IE、MPI、RT、RPG 以及 PP 方法。

5.2 研究结果

5.2.1 单个属性和 KS 判准率

根据表 6，当仅要求测验满足内容约束时运用 Monte Carlo 方法选题在 S3 和 S4 属性结构下测量精度分别达到 0.96 和 0.88。当测验仅考虑曝光控制时各个方法在相同属性结构下的测量精度从高到低依次为 PP、IE、RT、MPI 和 RPG。两种属性结构下 Monte Carlo 方法的测量精度都大于等于 RT、MPI 和 RPG 方法，小于等于 PP 和 IE 方法。对比表 6 和表 3 可知，仅考虑某类约束条件的结果比同时

考虑两类约束的测量精度更高。这表明，测验约束条件越多，测量精度越低。

5.2.2 项目曝光率结果

表 6 与表 7 表明，Monte Carlo 方法在满足内容约束时的题库利用率达 0.87 以上，最大项目曝光率大于 0.55，Monte Carlo 方法的表现与 Belov 等(2008)与 Mao 和 Xin (2013)的结果一致；PP 方法的测量精度最高，但该方法题库利用率最低，约 0.5 左右，最大项目曝光率大于 0.86；IE 方法能在一定程度上控制最大项目曝光率，不能提高曝光过低项目的使用率，结果与 van der Linden 和 Veldkamp (2004)一致；RT 和 RPG 方法均能较好地控制最大项目曝光率和提高题库利用率，它们在控制项目曝光率方面的表现与 Wang 等(2011)的研究结果一致。将表 7 与表 5 对比，可发现 Monte Carlo 方法和各方法结合使用能进一步提高项目曝光均匀性，但同时也降低了测量精度。

表 6 S3 和 S4 结构中不同测验条件下单个属性和 KS 的判准率

属性结构	测验条件	选题方法	属性						KS
			A1	A2	A3	A4	A5	A6	
S3	OCC	MC	0.998	0.994	0.994	0.991	0.988	0.991	0.960
	OEC	PP	0.999	0.994	0.999	0.998	0.990	0.996	0.977
	OEC	IE	1.000	0.995	0.991	0.996	0.983	0.990	0.960
	OEC	RT	1.000	0.998	0.993	0.993	0.990	0.992	0.960
	OEC	MPI	1.000	0.989	0.991	0.995	0.985	0.990	0.960
	OEC	RPG	1.000	0.986	0.986	0.993	0.975	0.983	0.941
S4	OCC	MC	0.998	0.976	0.970	0.975	0.965	0.975	0.883
	OEC	PP	0.999	0.988	0.988	0.982	0.976	0.982	0.930
	OEC	IE	0.999	0.970	0.976	0.976	0.970	0.971	0.900
	OEC	RT	0.998	0.978	0.971	0.974	0.955	0.965	0.874
	OEC	MPI	0.999	0.972	0.964	0.966	0.958	0.971	0.870
	OEC	RPG	0.994	0.960	0.951	0.964	0.954	0.951	0.825

注: OCC: 表示仅内容约束; OEC: 表示仅曝光控制, MC: Monte Carlo 方法。(下表同)

表 7 S3 和 S4 结构中不同测验条件下的项目曝光率结果

属性结构	测验条件	选题方法	曝光率				UR	NOR	OR	χ^2	统一量纲
			Q_1	Q_2	Q_3	max					
S3	OCC	MC	0.004	0.019	0.07	0.571	0.896	0.071	0.175	58.62	2.319
	OEC	PP	0.000	0.000	0.054	0.865	0.456	0.075	0.316	126.6	1.711
	OEC	IE	0.00	0.002	0.10	0.225	0.55	0.083	0.162	52.528	2.003
	OEC	RT	0.009	0.023	0.073	0.246	0.979	0.004	0.127	35.936	3.031
	OEC	MPI	0.012	0.047	0.094	0.293	0.972	0.004	0.097	21.55	3.911
	OEC	RPG	0.024	0.044	0.077	0.280	1.00	0.002	0.076	11.67	3.917
S4	OCC	MC	0.002	0.02	0.071	0.554	0.870	0.054	0.175	58.78	2.224
	OEC	PP	0.002	0.002	0.468	0.874	0.533	0.077	0.311	124.2	1.743
	OEC	IE	0.000	0.000	0.009	0.085	0.606	0.067	0.156	49.92	2.020
	OEC	RT	0.000	0.011	0.026	0.072	0.979	0.004	0.123	33.759	2.961
	OEC	MPI	0.000	0.009	0.051	0.092	0.967	0.002	0.092	18.875	3.644
	OEC	RPG	0.01	0.025	0.044	0.076	1.000	0.002	0.072	9.494	3.852

注: UR: 题库利用率; NOR: 曝光率大于 0.2 的项目比例; OR: 测验重叠率。

6 结论和讨论

目前, CD-CAT 选题策略主要针对某一类测验要求如测量精度、项目曝光控制或内容约束开展研究。实际上, 许多大规模测验往往具有成百上千个约束条件(van der Linden, 2005)。因此, 在具有多类约束条件下比较选题方法的表现具有重要实践意义。本文针对 CD-CAT 中项目曝光控制和内容约束, 运用 RUM 在不同属性结构下模拟被试和题库比较 5 种选题方法的表现。研究一表明, MC-PP 方法在各种属性结构下的测量精度最高, 项目曝光均匀性最差; MC-IE 方法的测量精度稍次于 MC-PP 方法,

其项目曝光均匀性比 MC-PP 方法好; MC-RT 和 MC-MPI 方法的测量精度没有明显差异, 稍次于 MC-IE 方法, MC-MPI 方法的项目曝光均匀性优于 MC-RT 方法和 MC-IE 方法; MC-PRG 方法的项目曝光近似服从均匀分布, 但其测量精度最差。由统一量纲值表明, MC-RPG 方法的综合表现最好, MC-MPI 方法次之。特别地, PP 方法在测量精度达到 0.8 时转向施测曝光率较低的项目来控制项目曝光, 所以 MC-PP 方法在各种属性结构下的测量精度都达到 0.8, 但是该方法并没有提高项目曝光均匀性。另外, 相同方法在不同属性结构下项目曝光率分布一致, 不同方法在相同属性结构下的表现具

有明显差异。这表明,相同方法在不同条件下具有稳定的曝光率分布,因而容易推广到其它实验条件;不同方法项目曝光率分布的差异源自方法本身的差异,不受实验条件的影响。这与毛秀珍和辛涛(2013)以及 Wang 等(2011)的研究结果类似。研究一与研究二可知,IE、RT、RPG 和 MPI 方法本身可以控制最大项目曝光率,当与 Monte Carlo 方法结合后能进一步提高项目曝光均匀性,同时也在 Monte Carlo 方法基础上降低测量精度。此外,IE、RT、MPI 与 RPG 方法与 Monte Carlo 方法结合的选题结果与毛秀珍和辛涛(2013)中 IE、RT、MPI 与 RPG 方法的表现一致。

虽然本文获得了一些有意义的研究结果,但仍存在一些不足之处。例如,模拟项目时没有考虑项目考查的属性个数与项目参数之间的关系,这与实际情况可能存在一些差异。第二,研究假设 KS 服从均匀分布,而 S1 和 S2 属性结构下分别有 7 种和 8 种 KS,所以这两种结构中一个属性都没有掌握的被试会占到 10%以上,这与实际存在一定差异。今后研究可假设属性掌握个数服从正态分布的条件下模拟被试 KS。第三,本文仅假设题库考查 6 个属性,且它们具有单一属性结构。实际上,题库往往考查更多属性,且属性结构通常由多种简单结构混合构成。因此,今后研究可以在考查更多属性、属性结构更复杂的情况下进行研究。此外,模拟研究虽然具有客观、可重复、节约和可控制误差等优点,但真实测验情景的影响因素较多,模拟研究不可能考虑周全。于是,要将这些方法运用于实践,还有待考证它们在实际情况中的表现。

本文介绍的选题方法都不局限于 RUM,它们还可以用到 DINA、NIDA 等认知诊断模型。另外,这些方法结合了同时选题和序列选题。因此,如何将测验组卷中满足约束的 0-1 线性规划(binary programming, BP)和遗传算法(genetic algorithm, GA)算法(Finkelman et al., 2010)以及分支有界(van der Linden, 2005)等算法用到具有非统计约束的 CAT 选题,或者探索满足约束的测验组卷算法都是有价值的研究问题。最后,CD-CAT 相关研究主要借鉴了传统 CAT 的思想,既要考虑内容约束和项目曝光控制,还要求测量精度时,不妨针对 CD-CAT 自身的特点研究选题方法。

参 考 文 献

Belov, D. I., Armstrong, R. D., & Weissman, A. (2008). A

- Monte Carlo approach for adaptive testing with content constraints. *Applied Psychological Measurement*, 32, 431-446.
- Chang, H. H., & Ying, Z. L. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222.
- 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.
- Chen, P., Ding, S. L., Lin, H., & Zhou, J. (2006). Item selection strategies of computerized adaptive testing based on graded response model. *Acta Psychologica Sinica*, 38, 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, 461-467.]
- 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.
- 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.
- Finkelman, M. D., Kim, W., Roussos, L., & Verschoor, A. (2010). A binary programming approach to automated test assembly for cognitive diagnosis models. *Applied Psychological Measurement*, 34, 310-326.
- 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.
- Hsu, C., Wang, W., & Chen, S. (2013). Variable-Length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement*, 37, 563-582.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuo's rule-space approach. *Journal of Educational Measurement*, 41, 205-237.
- Lin, H. J., & Ding, S. L. (2007). An exploration and realization of computerized adaptive testing with cognitive diagnosis. *Acta Psychologica Sinica*, 39, 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, 747-753.]
- 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, 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, 618-625.]
- Mao, X. Z., & Xin, T. (2013). A comparison of item selection methods for controlling exposure rate in cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica*, 45, 694-703.
- [毛秀珍, 辛涛. (2013). 认知诊断 CAT 中项目曝光控制方法的比较. *心理学报*, 45, 694-703.]
- Mao, X. Z., & Xin, T. (2013). The application of the monte

- carlo approach to cognitive diagnostic computerized adaptive testing with content constraints. *Applied Psychological Measurement*, 37, 482–496.
- McGlohen, M. K., & Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, 40, 808–821.
- Tatsuoka, K. K., & Tatsuoka, M. M. (1997). Computerized cognitive diagnostic adaptive testing: effect on remedial instruction as empirical validation. *Journal of Educational Measurement*, 34, 3–20.
- Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Taylor & Francis Group.
- Templin, J., Henson, R. A., & Templin, S. E. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement*, 32, 559–574.
- Tu, D. B. (2009). *Compilation of CD_CAT about preliminary children's math problem solving under AIG context*. (Unpublished doctoral dissertation). Jiangxi Normal University.
- [涂冬波. (2009). 项目自动生成的小学儿童数学问题解决认知诊断 CAT 编制 (博士学位论文). 江西师范大学.]
- van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27–52). Norwell, MA: Kluwer.
- van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.
- 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.
- Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, 73, 1017–1035.
- 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.
- Xu, X. L., Chang, H. H., & Douglas, J. (2003). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

A Comparison of Item Selection Methods for Cognitive Diagnostic Computerized Adaptive Testing with Nonstatistical Constraints

MAO Xiuzhen¹; XIN Tao²

(¹ Institute of Education Science, Sichuan normal university, Chengdu 610066, China)

(² Institute of Developmental Psychology, Beijing normal university, Beijing 100875, China)

Abstract

It is well known that items in the bank of computerized adaptive testing (CAT) are always expected to be used equally. For one thing, a good deal of manpower and financial resources spent on constructing the item bank will surely be wasted if a large proportion of items are seldom exposed or even never be used. For the other, works for ensuring the test security and maintaining the item bank will become serious for test practitioners if items are exposed extremely skewed. In addition to controlling the item exposure, tests which assembled for different examinees are usually required to satisfy many constraints, such as (a) the well-proportional of each content domain; (b) the “enemy items” could not be appeared in the same test, and (c) the appropriate balance of item keys. Supposing some constraints are violated, it will give some unexpected reactions during the test and result in inaccuracy of trait estimates. Therefore, both item exposure control and content constraints are important non-statistical constraints. They have great influence on the test validity, measurement accuracy and comparability among examinees. So, they need to be incorporated into the designing of item selection for CAT in practical settings.

When cognitive diagnostic theory is used in CAT, examinees can receive more detailed diagnostic information regarding their mastery of every attribute. Therefore, cognitive diagnostic CAT (CD-CAT) is a promising research area and has gained much attention because it integrates both the cognitive diagnostic method and adaptive testing. The present study compared the performances of five item selection methods in CD-CAT with item exposure control and content constraints. The item selection methods applied are (a) incorporating the Monte Carlo approach into the item eligibility approach (MC-IE); (b) incorporating the

maximum priority index method into the Monte Carlo approach (MC-MPI); (c) incorporating the restrictive threshold method into the Monte Carlo approach (MC-RT); (d) incorporating the restrictive progressive method into the Monte Carlo approach (MC-RPG), and (e) incorporating the maximum post probability of knowledge states method into the Monte Carlo approach (MC-PP). The reparameterized unified model was implemented in the simulation experiments to generate item responses with respect to five item banks constructed according to attribute structures of linear, convergent, divergent, unstructured and independent, respectively.

Results indicate that (a) the distributions of item exposure produced by the same item selection method in different item banks are similar, (b) the measurement precisions of each item selection method yield in attribute structures of linear, convergent, divergent, unstructured and independent are decreased gradually; (c) the performances of different item selection methods ordered by the measurement accuracy in each test condition are methods of the MC-PP, the MC-IE, the MC-MPI, the MC-RT, and the MC-RPG; their performances in terms of item exposure control are sorted in the opposite order.

According to the value of uniformly dimensional, the MC-RPG method yields a best balance between item exposure control and test accuracy while satisfying some content constraints, and then followed by the MC-MPI method.

Key words cognitive diagnostic theory; computerized adaptive testing; measurement accuracy; item exposure rate; content constraints.