

基于属性掌握概率的认知诊断计算机化 自适应测验选题策略*

罗照盛¹ 喻晓锋^{1,2} 高椿雷¹ 李喻骏¹ 彭亚风¹
王 睿¹ 王钰彤¹

(¹ 江西师范大学心理学院, 南昌 330022) (² 亳州师范高等专科学校, 亳州 236800)

摘 要 在认知诊断计算机化自适应测验(CD-CAT)中, 被试对每个属性的掌握概率更直接地反映了被试能力的当前估计值。因此, 基于被试的属性掌握概率来构建选题策略, 选择最能改变被试属性掌握概率的题目作为下一个测验项目, 这应该是一个值得尝试的方案。本文借鉴已有相关研究的数据生成模式进行探索, 模拟实验结果表明: 假设属性间相互独立, 在定长(长度为 16)、变长(长度为 16 或后验属性掌握模式概率达到 0.8)以及短测验(长度分别为 4、6、8、10)的情况下, 基于属性掌握概率的选题策略 PPWKL 和 PHKL 有较好的分类准确率, 在题目曝光率, 题库使用均匀性等方面也有较好的表现; 与研究较多的 PWKL、HKL 等策略相比, 也略有优势; 当属性间存在不同程度的相关时, 在定长、变长以及较短的测验条件下, 基于 PHKL 和 MI 的测验对知识状态估计精度较好, 基于 PPWKL 和 PHKL 的测验综合表现占优。

关键词 认知诊断计算机化自适应测验; 选题策略; 属性掌握概率; 属性掌握模式

分类号 B841

1 引言

相对于传统的测验形式, 计算机化自适应测验(Computerized Adaptive Test, CAT)由于有更高的测试效率和更好的测验精度而受到广泛关注(Barrada, Olea, Ponsoda, & Abad, 2008; Chang & Ying, 1999; Chang, Qian, & Ying, 2001; 程小杨, 丁树良, 严深海, 朱隆尹, 2011; 刘珍, 丁树良, 林海菁, 2008)。与其它测验形式相比, 认知诊断(Cognitive Diagnosis, CD)测验最大的优势在于它能提供被试在测验领域上的知识诊断报告, 这个诊断报告包含了更加丰富的评价信息, 可以对被试的进一步学习、教师开展针对性教学等提供帮助(Leighton & Gierl, 2007; Rupp, Templin, & Henson, 2010)。

认知诊断计算机化自适应测验(Cognitive Diagnostic Computerized Adaptive Testing, CD-CAT)(Cheng, 2009a, 2009b; McGlohen & Chang, 2008; Xu, Chang, & Douglas, 2003)建立在传统 CAT(指没有诊断功能的 CAT)的基础之上, 同时赋予传统 CAT 新的功效——认知诊断, 它是将认知诊断的基本理论、方法与计算机自适应测验相结合的产物。CD-CAT 结合了 CAT 和 CD 的优点, 它一方面可以对被试的知识状态进行诊断; 另一方面在诊断过程中可以有针对性、“量体裁衣”式的选择项目让被试作答, 从而有利于提高测验效率和测量精度(Cheng, 2009a)。根据 Wang (2013)的描述, CD-CAT 是结合了 CD 和 CAT 二者的优点的一种测验, 其中 CD 的目的是根据被试的知识掌握情况对被试分类, 找到被试的优

收稿日期: 2014-05-13

* 国家自然科学基金(31160203, 31100756, 31360237)、国家社会科学基金(12BYY055)、教育部人文社会科学研究青年基金项目(13YJC880060)、安徽省高校省级优秀青年人才基金重点项目(2013SQRL127ZD)、安徽省自然科学研究项目(KJ2010B123, KJ2013B151, KJ2013B250)、高等学校博士学科点专项科研基金(20113604110001)、江西省研究生创新专项基金(YC2013-B024)、安徽省哲学社会科学规划项目(AHSKY2014D102)、安徽省高等教育振兴计划重大教学改革研究项目(2014ZDJY190)和资助。

通讯作者: 喻晓锋, E-mail: andyyuxf@163.com 高椿雷, E-mail: gaochunlei51@qq.com

势和弱点,而 CAT 的算法则使这一过程尽可能更高效的实现。

在传统 CAT 研究中,选题策略是一个重要的组成部分,每次都是基于被试的当前能力估计值,根据某种信息测度(比如 Fisher 最大信息量)来选择下一个要施测的项目,从而达到采用较少的项目估计被试的能力也能达到预先设定的精度(Chang & Ying, 1996)。

关于 CD-CAT 的选题策略,已有的研究主要有 5 种:一是基于 KL 信息量的选题策略(KL),即根据被试的当前属性掌握模式估计值,每次从题库或剩余题库中选择 KL 信息量最大的项目施测(Cheng, 2009a)。二是基于香农熵的选题策略(SHE),根据被试的当前属性掌握模式估计值,每次从题库或剩余题库中选择香农熵最小的项目施测(Tatsuoka, 2002, Xu et al., 2003)。三是基于后验概率加权的 KL 信息量的选题策略(PWKL),相对于 KL 信息量选题策略, PWKL 给 KL 信息量增加了不同的权重,权重是属性掌握模式的后验概率(Cheng, 2009a)。四是基于后验概率和属性掌握模式距离加权的 KL 信息量的选题策略(HKL),相对于 PWKL 选题策略, HKL 选题策略的区别在于权重不同, HKL 选题策略同时考虑后验概率和属性掌握模式间的相似性来对 KL 信息量加权(Cheng, 2008, 2009a)。有关这 4 种选题策略的详细计算方式,请参考 Cheng (2008, 2009a),并且 Cheng 对这 4 种选题策略之间的关系进行了阐述和说明。根据 Cheng 研究结果,这 4 种选题策略的模式分类准确率最高的是 HKL,并且 PWKL 与 HKL 的模式分类准确率很接近。第五种是基于互信息 (Mutual Information) 的选题策略 (MI, Wang, 2013), Wang 研究了短测验下,比较 MI 与 KL, PWKL, SHE 等策略的表现,模拟研究表明:对被试的属性掌握模式准确率上,在 Wang 的实验条件下, MI 在多数情况下略占优。

已有的 CD-CAT 选题策略基本是基于被试属性掌握模式的当前估计值,并结合某种信息测度,比如 KL 信息量,香农熵或互信息等,从题库或剩余题库中选择某个项目来施测。属性掌握模式估计值通常是通过截断点转换(比如将属性掌握概率与 0.5 比较,大于 0.5 则认为被试掌握了某属性,否则不掌握)或者是取最大期望后验概率(Maximum A Posterior, MAP)对应的属性掌握模式而得到(Huebner & Wang, 2011)。然而,在自适应测验初期,由于对被试水平的诊断信息较少,此时的属性掌握

模式估计值可能存在较大偏差,如果采用的选题策略仅仅基于当前的属性掌握模式和作答,会不利于估计被试的属性掌握模式,进而影响到整个 CD-CAT 的测验效度和测验精度(涂冬波,蔡艳,戴海崎, 2013)。涂冬波等(2013)研究了在初始阶段选择包含“可达矩阵”的项目让被试作答,模拟实验结果表明,初始阶段的选题对被试的属性掌握模式的估计是有影响的。根据被试属性掌握概率和人为给出的截断点赋以被试的知识状态,比如截断点为 0.5,则两个知识状态某一个分量(属性掌握概率)为 0.01 和 0.49 的,都评判为该属性没有掌握,但是两者的差异是明显的。因此,被试属性掌握概率(Attribute Mastery Probability, AMP)可以更直接地反映被试的当前水平,还未发现基于被试属性掌握概率的选题策略的研究(即在测试过程中,使用属性掌握概率变化加权的选题策略)。也正是基于这种考虑,本研究基于被试的属性掌握概率,提出两种新的选题策略,并与已有的 CD-CAT 下的选题策略进行比较。

2 基于属性掌握概率的选题策略

在介绍新的选题策略之前,首先对涉及到的概念和符号进行说明。

2.1 相关的概念

属性掌握概率:它是一个由 0 到 1 之间的数,被试在测验各属性上的掌握概率就构成了属性掌握概率向量,每个元素对应了被试对该位置上的属性的掌握概率。比如:某被试的属性掌握概率向量为 $[0.21, 0.68, 0.85]$,表明该被试掌握测验中三个属性的概率分别是 0.21、0.68 和 0.85。

属性掌握模式(或称知识状态):它是一个由 0 和 1 组成的二值向量,其中向量中的 0 表示被试对该位置对应的属性没有掌握,1 表示掌握。比如:某被试的属性掌握模式为 $[0, 1, 1]$,表明测验考察了三个属性,该被试掌握了第 2 和第 3 个属性,未掌握第 1 个属性。属性掌握模式通常是通过属性掌握概率转换得到的,比如采用 0.5 为截断点,属性掌握概率为 $[0.21, 0.68, 0.85]$ 的被试的属性掌握模式为 $[0, 1, 1]$ 。

属性掌握概率变化加权:属性掌握概率变化是指被试在作答某个项目前后其属性掌握概率的差异,分为三种情况:单个属性掌握概率变化、最不确定属性掌握概率变化和属性掌握概率之和的变化。比如被试在作答某项目之前的属性掌握概率为

[0.21,0.68,0.85], 作答之后其属性掌握概率变为 [0.61,0.75,0.91]。单个属性掌握概率变化是作答前后指 3 个属性的掌握概率变化值, 分别为 0.40、0.07 和 0.06, 其中第 1 个属性的掌握概率变化最大; 最不确定属性掌握概率是指与 0.5 最接近的属性掌握概率, 比如 [0.21,0.68,0.85] 中与 0.5 最接近属性第 2 个属性, 其掌握概率是 0.68, 在作答之后变为 0.75, 需要注意的是, 这里所说的“最不确定”只是一个相对的概念, 它是以概率 0.5 作为参照; 属性掌握概率之和变化是指作答某项目前后 3 个属性的掌握概率之和变化的绝对值, 作答之前的属性掌握概率之和为 $0.21+0.68+0.85=1.74$, 作答之后为 $0.61+0.75+0.91=2.27$, 则变化值为 0.53 (可以由 $|2.27-1.74| = 0.53$ 得到)。将属性掌握概率的变化(包括上面的三种情况)值作为选题时的一个权重即为属性掌握概率变化加权。

2.2 相关的符号

假设 CD-CAT 考察的属性个数为 K , 参加测验的被试($i=1, 2, \dots, N$)已经作答 t 个项目, 其作答模式记为 $x_i^t=(x_{i1}, x_{i2}, \dots, x_{it})$, x_i^t 表示被试 i 已作答 t 个项目时的作答向量, x_{it} 表示被试 i 在第 t 个项目上的作答。将被试 i 作答完 t 个项目后的可用项目的集合记为 R_i^t (即剩余题库)。CD-CAT 中的选题策略是指基于某种规则从 R_i^t 中选择最适合被试 i 作答的第 $t+1$ 个项目。

h 为来自 R_i^t 中的某个项目, 用 $D_h(\hat{\alpha}_i^t|\alpha_c)$ ($c=1, 2, \dots, 2^K$) 表示条件分布间的 KL 信息量, $D_h(\hat{\alpha}_i^t|\alpha_c)$ 的具体计算请参考陈平、李珍和辛涛(2011), 其中, $\hat{\alpha}_i^t$ 表示被试 i 完成 t 个项目后 KS 的估计值, α_c 代表某种可能的 KS , 如果不考虑属性间的层级关系, 则所有可能的 KS 数量为 2^K 个。

2.3 基于属性掌握概率的选题策略

在认知诊断测验中, 期望后验估计(Expect A Posterior, EAP)方法常常被用来计算被试在每个属性上的掌握概率(即被试在属性上的边际掌握概率), 进一步将被试对属性的掌握概率与 0.5 相比较, 当属性掌握概率大于 0.5 时, 即认为被试掌握了该属性, 否则认为被试没有掌握该属性 (Huebner & Wang, 2011)。基于属性掌握概率的选题策略是从属性掌握概率出发, 对属性掌握概率不作 0、1 转换, 选择对被试属性掌握概率影响最大的项目作为下一个施测的项目。这样做的原因有两个: 第一是因为在测验初期, 对被试的属性掌握概率估计存在较大的偏差, 随着测验的进行, 这个属性掌握概率估

计值会逐渐趋近其真值, 我们希望新的选题策略能使测验加快这个过程, 因此选择能使被试的属性掌握概率改变最大的项目作为下一个施测的项目; 第二是由于被试的属性掌握模式是通过将属性掌握概率与截断点(通常取 0.5)进行比较, 然后转换得到的, 但是当一个属性的掌握概率与 0.5 非常接近, 比如 0.501 或 0.499, 其实此时测验对该属性的状态“非常不确定”的。模拟实验表明, 基于单个属性掌握概率变化最大、基于最不确定属性的掌握概率变化最大的策略表现不佳, 因此这里仅考虑研究基于属性掌握概率之和变化最大的策略(即选择作答某项目前后, 被试对各属性的掌握概率之和变化最大的项目)的表现。

被试 i 作答 t 个项目之后的属性掌握概率记为 $p_i^t=(p_{i1}^t, p_{i2}^t, \dots, p_{iK}^t)^T$, 上标 T 表示向量的转置。使用 EAP 方法计算被试在每个属性上的掌握概率方法如下(关于 EAP 详细的计算方法请见附录一):

$$p_{ik}^t = \sum_{c=1}^{2^K} (\pi_{i,t}(\alpha_c) \cdot \alpha_{ck}), (k=1, 2, \dots, K) \quad (1)$$

其中 $\pi_{i,t}(\alpha_c)$ 是作答完 t 个项目后, 被试 i 属于第 c 种属性掌握模式的后验概率(陈平等, 2011), α_{ck} 表示第 c 种属性掌握模式中属性 k 的取值, 可以为 0 或 1。

2.3.1 基于后验概率和属性掌握概率变化加权的 KL 选题策略

记后验概率和属性掌握概率变化加权的 KL 选题策略为 PPWKL (Posterior Probability Weighted Kullback-Leibler)。PWKL 选题策略是基于后验概率加权的 KL 信息量, 这里的 PPWKL 是在 PWKL 的基础上, 增加了属性掌握概率的变化值这一权重, 即基于后验概率和属性掌握概率变化加权的 KL 信息量。PPWKL 指标的计算方式为

$$PPWKL_h(\hat{\alpha}_i^t) = \sum_{c=1}^{2^K} \left(D_h(\hat{\alpha}_i^t|\alpha_c) \cdot \pi_{i,t}(\alpha_c) \cdot \sum_{k=1}^K |p_{ik}^{t+1} - p_{ik}^t| \right) \quad (2)$$

该选题策略可以表示为:

$$\text{第 } t+1 \text{ 个项目} = \arg \max_{h \in R_i^t} [PPWKL_h(\hat{\alpha}_i^t)] \quad (3)$$

2.3.2 基于后验概率、属性掌握概率变化和属性掌握模式距离加权的 KL 选题策略

记后验概率、属性掌握概率变化和属性掌握模式距离加权的 KL 选题策略为 PHKL (Posterior Hybrid Kullback-Leibler)。与 PPWKL 不同, 这里的 PHKL 是在 HKL 的基础上, 增加了属性掌握概率之

和的变化值这一权重,即基于后验概率、属性掌握概率变化和属性掌握模式距离加权的 KL 信息量。

PHKL 指标的计算方式为

$$PHKL_h(\hat{\alpha}_i^t) = \sum_{c=1}^{2^K} \left(D_h(\hat{\alpha}_i^t | \alpha_c) \cdot \pi_{i,t}(\alpha_c) \cdot \sum_{k=1}^K |p_{ik}^{t+1} - p_{ik}^t| \cdot \frac{1}{d(\hat{\alpha}_i^t, \alpha_c)} \right) \quad (4)$$

该选题策略可以表示为:

$$\text{第 } t+1 \text{ 个项目} = \arg \max_{h \in R_i^t} [PHKL_h(\hat{\alpha}_i^t)] \quad (5)$$

其中, $d(\hat{\alpha}_i^t, \alpha_c)$ 是属性掌握模式 $\hat{\alpha}_i^t$ 和 α_c 之间的欧氏距离, 其计算方式如下:

$$d(\hat{\alpha}_i^t, \alpha_c) = \sqrt{\sum_{k=1}^K (\hat{\alpha}_{ik}^t - \alpha_{ck})^2} \quad (6)$$

3 模拟研究 1

3.1 研究设计

为了考察基于属性掌握概率的选题策略的表现, 考虑基于 DINA 模型, 模拟定长和变长的 CD-CAT 测验。已有的研究中, CD-CAT 的测验长度经常取 12 到 24 这个范围(陈平等, 2011; 涂冬波等, 2013; Wang, 2013)。本研究中, 对于定长的 CD-CAT 测验, 测验长度采用 16。对于变长的 CD-CAT 测验, 终止规则由测验长度和属性掌握模式后验概率确定, 只要二者之一达到预先设定的值即终止测验。Hsu, Wang 和 Chen (2013)研究了变长 CD-CAT 下, 属性掌握模式后验概率取不同值对测验长度、测验精度的影响, Hsu 等的结果表明, 对于高风险和低风险测验, 属性掌握模式后验概率分别应取不低于 0.9 和不高于 0.8。考虑的测验长度(每位被试测验的最大长度)也是 16, 属性掌握模式后验概率最大值固定为 0.8(Hsu et al., 2013; Tatsuoaka, 2002), 即测验长度达到最大值或者属性掌握模式后验概率达到最大值时结束测验。

目前多数的 CD-CAT 研究是基于 4-8 个属性进行的, 其中 6 个居多(Cheng, 2009a, 2010; Wang, Chang, & Huebner, 2011; Xu et al., 2003), 这里模拟的题库考虑 6 个属性。

题库中各项目的属性向量和项目参数, 被试属性掌握模式的模拟按如下方式进行:

(1) 题库中的项目数固定为 200, 各项目按 0.2 的概率考察每个属性, 并且保证每个项目至少考察 1 个属性, 最多考察 3 个属性(Henson, 2004);

(2) 项目参数 s 和 g 都采用均匀分布, 取值区间为[0.05, 0.25];

(3) 因为在实际的测验情境下, 所考察的属性之间可能存在相关。为了比较不同情况下, 各策略的表现, 分别考虑属性间独立和属性间存在相关的情况。对于属性间相互独立的情况, 假设被试掌握每个属性的概率服从参数为 0.5 的 Bernoulli 分布, 随机生成被试的属性掌握情况。对于属性间存在相关的情况, 这里分别取 0.2, 0.35, 0.5, 0.6, 0.7, 0.8 共 6 种情况的相关, 比如 0.2 表示所有属性之间存在 0.2 左右(表示属性间的相关接近 0.2, 可能不一定是 0.2)的相关, 其它相关的含义与此相同。模拟属性间的相关可以通过 HO-DINA 模型的高阶参数来控制, 可以模拟被试总体对掌握的各属性之间存在不同大小的相关, 具体可以参考 Wang, Chang 和 Douglas (2012)。这样一来, 就存在属性间相互独立, 属性之间存在较低的相关(相关系数为 0.2 和 0.35), 中等程度的相关(相关系数为 0.5 和 0.6)和较高的相关(相关系数为 0.7 和 0.8)共 7 种情况。

一共模拟 1000 名被试, 200 个项目, 有 7 种被试总体(指属性之间独立和存在不同的相关)。基于各选题策略模拟 CD-CAT。每种选题策略重复 20 次, 结果取平均值, 所有的模拟程序采用 Java 语言编制。

3.2 CD-CAT 测验施测过程

CD-CAT 按如下过程进行模拟: (1)随机生成被试的属性掌握模式; (2)按采用的选题策略, 选择下一个要施测的项目; (3)模拟被试作答; (4)采用 EAP 方法估计被试的属性掌握概率(de la Torre, 2009)。对于涉及到 KL 信息量的选题策略, 还需估计被试的属性掌握模式, Huebner 和 Wang (2011)的研究表明, 采用 MAP 方法估计被试的属性掌握模式更好, 因此, 这里采用 MAP 方法估计被试的属性掌握模式; (5)转到步骤(2), 直到满足测验终止规则。当所有的被试完成测验后, 计算相应的评价指标。

3.3 评价指标

为了综合考察每种选题策略的表现, 采用的评价指标有: 测验的平均长度(N_{ave_len}), 项目的最小曝光率(ER_{min}), 项目的最大曝光率(ER_{max}), 项目的平均曝光率(ER_{ave}), 题库中未使用的项目个数(N_{no_use}), 曝光率大于 20% 的项目个数($N_{20\%}$), χ^2 值, 测验重叠率(TOR), 属性模式判准率(PMR)和属性边际判准率(MMR), 这些指标的计算可参见陈平等(2011), 其中: PMR, MMR, ER_{min} 三个指标是越大越好, 其余的指标是越小越好。为了节省篇幅, 下面对属性间的相关按低, 中和高三种情况来报告结果, 其中

“低”对应的结果是指属性间的相关为 0.2 和 0.35 时的平均值,“中”对应的结果是指属性间的相关为 0.5 和 0.6 时的平均值,“高”对应的结果是指属性间的相关为 0.7 和 0.8 时的平均值。

为了能全面地比较不同选题策略之间的差异,综合考虑各评价指标下不同选题策略的表现,采用统一量纲再加权求和的方法,具体做法是:对值越大越好的指标,将该评价指标上的最大值做分母,把各选题策略在该指标上的值做分子,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。统一量纲后,对某选题策略的 10 个评价指标比值分别赋加权系数。加权求和值最大的,则该选题策略在各个方面的综合效果最好;反之则最差(陈德枝, 2004; 刘珍等, 2008;)。本文中所采用的 10 个评价指标中有 2 个(PMR, MMR)是评价测验的估计精度,余下的 8 个是评价测验题库使用相关的指标,因此,为了使两类指标(评价知识状态准确性的指标和评价题库使用均匀性的指标)

在统一量纲中占有相同的比重,加权系数的设置方式为:PMR 和 MMR 指标的权重设置为 4,其余指标的权重设置为 1,这样可以保证两类指标占有相同的比重。这里举一个例子说明,比如表 1 中,对于 PHKL 的统一量纲的指标计算方式为: $4 \times 0.961 / 0.961 + 4 \times 0.992 / 0.992 + 1 \times 16 / 16 + 1 \times 92.118 / 92.118 + 1 \times 0.540 / 0.540 + 1 \times 1 \times 0.967 / 0.972 + 1 \times 1 \times 9.8 / 100 + 1 \times 22 / 25 = 15.83$,其中对 PMR 和 MMR 来说是值越大越好,计算时是将各策略对应的指标作为分子,所有 PMR 和 MMR 中最大的值作为分母,并且乘上对应的权重;其它指标是越小越好,计算时是将各策略对应的指标作为分母,所有对应指标中最小值作为分子,并且乘上相应的权重,最后对所有指标按统一量纲后求和,得到评价各策略的综合评价指标。

3.4 实验结果

3.4.1 属性之间独立时的结果

表 1 和表 2 分别对应了定长和变长 CD-CAT 测验下各评价指标的值,并且表 1 和表 2 中的最后一

表 1 各选题策略的分类准确率和题库使用均匀性(定长, 属性独立)

选题策略	知识状态准确性		题库使用均匀性								统一量纲后的综合指标
	PMR	MMR	N_{len}	χ^2	TOR	ER_{min}	ER_{max}	ER_{ave}	N_{no_use}	$N_{20\%}$	
KL	0.445	0.894	16	102.454	0.592	0	1.000	0.08	136	27	12.77
SHE	0.949	0.990	16	97.618	0.568	0	1.000	0.08	106	22[1]	15.73
PWKL	0.948	0.989	16	97.209	0.566	0	0.975	0.08	110	25	15.60
HKL	0.945	0.989	16	96.441	0.562	0	0.971	0.08	112	24	15.63
MI	0.955	0.991	16	97.906	0.569	0	1.000	0.08	104	22[1]	15.77
PPWKL	0.954	0.990	16	92.118[1]	0.540[1]	0	0.967[1]	0.08	98[1]	25	15.84
PHKL	0.961[1]	0.992[1]	16	93.357	0.546	0	0.972	0.08	100	25	15.83

注: (1) PMR, MMR, N_{len} , χ^2 , TOR, ER_{min} , ER_{max} , ER_{ave} , N_{no_use} , $N_{20\%}$ 分别表示属性模式判准率, 属性边际判准率, 测验长度(变长 CD-CAT 下对应的是平均测验长度), 卡方统计量, 测验重叠率, 最小曝光率, 最大曝光率, 平均曝光率, 未使用项目数, 曝光超过 20% 的项目数。

(2) 数据后面中括号中的数字表示该指标在所有选题策略中所处的位置, 比如: PMR 列的 **0.961[1]** 表示 PHKL 选题策略的 PMR 指标最好(PMR 越大越好), 处于第 1 位。有灰色底纹显示的选题策略是至少有一项指标排在第 1 位的选题策略。

(3) 由于每种选题策略下都有未使用的项目, 因此, 最小曝光率指标 ER_{min} 在所有选题策略下都为 0。

(4) 表中粗体显示的数据是在所有选题策略中排名第 1 位的指标值。

表 2 各选题策略的分类准确率和题库使用均匀性(变长, 属性独立)

选题策略	知识状态准确性		题库使用均匀性								统一量纲后的综合指标
	PMR	MMR	N_{len}	χ^2	TOR	ER_{min}	ER_{max}	ER_{ave}	N_{no_use}	$N_{20\%}$	
KL	0.446	0.893	15.8	103.394	0.596	0	1.000	0.079	138	24	12.05
SHE	0.834	0.970	9.4[1]	102.527	0.559	0	1.000	0.047[1]	142	14[1]	15.46
PWKL	0.843	0.971	9.8	99.213	0.545	0	0.951	0.049	125	15	15.56
HKL	0.840	0.969	9.8	97.358	0.535	0	0.956	0.049	123	15	15.60
MI	0.836	0.969	9.5	102.216	0.558	0	1.000	0.048	129	15	15.45
PPWKL	0.840	0.969	9.8	94.621	0.522	0	0.948	0.049	121	16	15.61
PHKL	0.848[1]	0.979[1]	9.8	92.872[1]	0.513[1]	0	0.944[1]	0.049	113[1]	16	15.79

列分别对应了定长和变长 CD-CAT 测验下各选题策略的综合评价指标。

根据表 1 和表 2, 在测验长度为 16 的定长 CD-CAT 下, 如果考察测验的精度, 则 PHKL, MI 和 PPWKL 的表现较好, 分别排名前三位。如果综合测验精度和题库的使用均匀性指标, 则 PPWKL, PHKL 和 MI 表现略好。在变长(最大测验长度为 16, 最大后验概率为 0.8)的 CD-CAT 下, PHKL, PWKL, PPWKL 和 HKL 的测验对被试的知识状态估计精度较好, PHKL 和 PPWKL 在题库使用均匀性的表现占优。

总的来说, 在属性独立情况下, 无论是定长, 还是变长的 CD-CAT, 考虑了属性掌握概率的选题策略在保证测验精度不损失或损失较小的情况下, 在题库使用均匀性上的表现都更好, 这些都可以从表 1 和表 2 可以很明显的看出来。

3.4.2 属性之间存在相关时的结果

表 3 和表 4 对应了属性之间存在较低的相关, 定长和变长 CD-CAT 测验各评价指标的值。

从表 3 可以看出, 在属性之间存在低相关, 长度为 16 的定长 CD-CAT 下, 单从测验精度来看, PHKL 和 SHE 略占优势, 其中 PHKL 选题策略的测验精度最好, 排在 2, 3, 4 位的分别是 SHE, MI 和

PPWKL, 它们与 PHKL 选题策略与 PMR 指标分别相差 1%, 1.5%和 1.5%; 但是如果考察题库的使用均匀性指标, PHKL 选题策略最好, 有 4 项指标(分别是 χ^2 , TOR, $ER_{\max}$, $N_{\text{no use}}$)排名第一, 1 项指标($N_{20\%}$)排名第二, 剩下的三项指标(N_{len} , $ER_{\min}$ 和 ER_{ave})都相同。综合测验精度和题库的使用均匀性来看, PHKL 的表现较好, 这一点从统一量纲后的综合指标也能看出。综合指标排在前三位的分别是 PHKL、PPWKL 和 HKL。

属性间存在中等相关和较高相关时的详细结果请参考附录二*。PHKL 在属性之间存在中等相关, 长度为 16 的定长 CD-CAT 下有 6 项指标排名第 1(分别是 PMR, MMR, χ^2 , TOR, $ER_{\max}$, $N_{\text{no use}}$)。从测验精度来说, PHKL 和 MI 很接近, 分别排在第 2 位, 但是 PHKL 在题库使用均匀性上有很大优势, 附录二表 6 中的综合指标也表明 PHKL 策略的综合表现更好。

当属性之间存在高相关, 长度为 16 的定长 CD-CAT 下, 在测验精度指标上, MI 策略最好(其 PMR 和 MMR 指标都排第 1), 但是从题库使用上来看, PHKL 策略更好(χ^2 , TOR, $ER_{\max}$, $N_{\text{no use}}$ 指标都排在第 1 位), 详细结果请参考附录二中的表 7。综合来看, PHKL 策略略占优。

表 3 各选题策略的分类准确率和题库使用均匀性(定长, 低相关)

选题策略	知识状态准确性		题库使用均匀性								统一量纲后的综合指标
	PMR	MMR	N_{len}	χ^2	TOR	$ER_{\min}$	$ER_{\max}$	ER_{ave}	$N_{\text{no use}}$	$N_{20\%}$	
KL	0.406	0.874	16	91.629	0.538	0	0.982	0.08	132	28	12.62
SHE	0.934	0.987	16	86.424	0.512	0	1.000	0.08	108	25.5	15.64
PWKL	0.925	0.986	16	88.470	0.522	0	0.974	0.08	104	26	15.60
HKL	0.927	0.986	16	87.838	0.519	0	0.973	0.08	102.5	24[1]	15.71
MI	0.929	0.986	16	86.893	0.514	0	1.000	0.08	108	25	15.63
PPWKL	0.929	0.986	16	85.355	0.506	0	0.968	0.08	98	24.5	15.80
PHKL	0.944[1]	0.994[1]	16	82.791[1]	0.493[1]	0	0.961[1]	0.08	96[1]	24.5[2]	15.98

表 4 各选题策略的分类准确率和题库使用均匀性(变长, 低相关)

选题策略	知识状态准确性		题库使用均匀性								统一量纲后的综合指标
	PMR	MMR	N_{len}	χ^2	TOR	$ER_{\min}$	$ER_{\max}$	ER_{ave}	$N_{\text{no use}}$	$N_{20\%}$	
KL	0.416	0.884	15.865	99.024	0.574	0	0.994	0.079	141.5	27	11.90
SHE	0.837	0.970	9.781	119.703	0.647	0	1.000	0.049	130	12[1]	15.36
PWKL	0.837	0.970	9.942	109.663	0.598	0	0.973	0.050	118	14	15.44
HKL	0.828	0.969	9.909	107.822	0.588	0	0.974	0.050	118	15	15.37
MI	0.829	0.968	9.701[1]	120.613	0.651	0	1.000	0.049[1]	132	12[1]	15.31
PPWKL	0.835	0.970	10.162	94.608[1]	0.523[1]	0	0.945	0.051	113.5[1]	15	15.66
PHKL	0.842[1]	0.972[1]	10.126	94.987	0.525	0	0.938[1]	0.051	113.5[1]	15	15.71[1]

*为节省篇幅, 纸质文章只呈现附录一, 电子版包含附录二、三和四。

表 4 的结果来看, 在属性间存较低相关, 最大测验长度为 16, 后验属性掌握模式概率为 0.8 的变长 CD-CAT 下, 单从测验精度来看, PHKL、PWKL 和 SHE 的表现较好, 分别处于第 1、2 和 3 位。综合来看, PHKL, PPWKL 和 MI 选题策略的总体表现较好, 分别有 4, 3 和 3 项指标排名第一, PHKL 的综合指标表现最好, 这说明在考虑了被试的属性掌握概率变化之后, 在保证测验精度的同时, 对题库使用的均匀性控制上也有了改善。

属性间存在中等相关和较高相关时也有类似的结论, 详细结果请参考附录三中的表 8 和表 9。

总体来说, 在变长 CD-CAT 下, 考虑了属性掌握概率的选题策略 PHKL, 在测验精度和题库使用均匀性指标上表现都较好, 无论是属性之间存在较低相关、中等相关或是较高相关的情况, PHKL 的综合指标都排在第 1 位。

4 模拟研究 2

认知诊断测验可能会用于日常分项诊断中, 比

如单元测验、随堂诊断测验等, 此时, 题目量可能会比较少。为了考察不同题量情形下各种选题策略的表现, 选取 PWKL、HKL、MI、PPWKL 和 PHKL 五个选题策略。考虑属性之间相互独立、存在相关时, 它们在短测验下的表现。所有的数据模拟方式与研究 1 相同, 不同的是本研究中考察的是长度分别为 4、6、8 和 10 的定长测验, 即模拟 4 种短测验, 考察这 5 种选题策略的表现。所采用的评价指标与研究 1 相同。表 5 列出了属性独立时, 4 种选题策略在 4 种定长的短测验中的表现。

从表 5 的结果可以看出, 在属性相互独立时, 所列出的所有短测验(也可以看作是长测验的测验初期)中, 大部分情况下, PHKL 的各项指标都是最好的。从统一量纲后的综合指标来看, PHKL 几乎总是优于其它几种策略, 只有在测验长度为 10 时, PPWKL 综合指标指第 1 位。总的来说, 考虑了属性掌握概率的选题策略 PHKL 和 PPWKL, 在综合测验精度和题库使用均匀性指标后略占优势。

当属性之间存在较低、中等和较高相关时, 各

表 5 三种选题策略的分类准确率和题库使用均匀性(短测验, 属性独立)

选题策略	知识状态准确性		题库使用均匀性								统一量纲后的综合指标
	PMR	MMR	N _{ave_Len}	χ^2	TOR	ER _{min}	ER _{max}	ER _{ave}	N _{no_use}	N _{20%}	
PWKL	0.187	0.778	4	95.717	0.498	0	0.754	0.020	177	5	15.829
HKL	0.180	0.772		96.096	0.500	0	0.758	0.020	178	5	15.632
PHKL	0.192	0.779		94.301	0.491	0	0.759	0.020	178	5	15.957
PPWKL	0.187	0.778		96.035	0.500	0	0.731	0.020	177	5	15.852
MI	0.185	0.777		144.415	0.742	0	1.000	0.020	192	6	14.644
PWKL	0.496	0.882	6	111.579	0.587	0	0.924	0.030	169	9	15.442
HKL	0.470	0.879		110.945	0.584	0	0.938	0.030	168	10	15.145
PHKL	0.527	0.903		108.385	0.571	0	0.933	0.030	165	8	15.952
PPWKL	0.514	0.894		106.590	0.563	0	0.932	0.030	164	9	15.740
MI	0.519	0.898		126.781	0.664	0	1.000	0.030	187	10	15.207
PWKL	0.665	0.923	8	107.519	0.577	0	0.953	0.040	158	12	15.636
HKL	0.677	0.925		106.821	0.574	0	0.957	0.040	153	13	15.675
PHKL	0.686	0.929		100.189	0.540	0	0.935	0.040	149	14	15.850
PPWKL	0.680	0.927		102.098	0.550	0	0.944	0.040	148	13	15.832
MI	0.679	0.927		109.601	0.588	0	1.000	0.040	167	13	15.527
PWKL	0.799	0.957	10	104.732	0.573	0	0.968	0.050	144	14	15.752
HKL	0.791	0.957		102.891	0.564	0	0.953	0.050	139	14	15.792
PHKL	0.806	0.959		98.611	0.543	0	0.963	0.050	136	15	15.901
PPWKL	0.804	0.958		99.320	0.546	0	0.955	0.050	133	15	15.903
MI	0.806	0.959		106.431	0.582	0	1.000	0.050	151	16	15.566

选题策略在短测验中的表现,请参考附录 4 中的表 10, 11 和 12。从表 10, 11 和 12 可以看出, PHKL 和 MI 两中策略的测验精度几乎总是排在前两位, 在长度为 4 和 6 的测验中, PHKL 占优, 在长度为 8 和 10 的测验中, MI 占优, 并且当属性之间的相关达到中等以上时, MI 在测验精度上的优势比属性之间存在低相关时略大; 在大多数情况下, PHKL 和 PPWKL 在题库使用均匀性上的表现总是排在前 2 位。综合来看, PHKL 大多数情况下, 综合指标都是排在第 1 位。

5 小结与讨论

通过对基于属性掌握概率的 2 种选题策略与 5 种已有的 CD-CAT 下的选题策略的比较研究发现: 在属性之间是独立的定长和变长的 CD-CAT 中, 考虑了属性掌握概率的 PHKL 和 PPWKL 选题策略在测验精度和题库的利用率上的综合表现优于其它选题策略; 当属性之间存在较低、中等和较高相关时, 在定长和变长的 CD-CAT 中, PHKL 和 PPWKL 在保证测验精度的同时, 对题库使用的均匀性控制上也有了改善, 它们的综合指标排在前 2 位; 当属性之间存在较低、中等和较高相关的短测验中, 在测验长度为 4 和 6 时, PHKL 的测验精度更好, 当测验长度达到 8 和 10 时, MI 的测验精度更好, 这说明 PHKL 策略更适合在测验初期使用; PHKL 和 PPWKL 策略的一大优点是在不损失或较少损失测验精度的条件下, 能改善题库的使用均匀性。

在测验初期, 因为关于被试属性掌握状态的信息较少, 采用基于属性掌握模式的选题策略可能不利于对被试的知识状态的估计, 因为这人为增加了误差。特别是在测验长度较短时, 从表 10, 11 和 12 中的结果很清楚地说明了这一点, 基于属性掌握概率的 PHKL 和 PPWKL, 以及基于互信息的 MI 的测验精度分别排在前 3 位。PHKL 和 PPWKL 选题策略在选题时考虑了属性掌握模式的后验概率和被试的属性掌握概率的变化情况, 选题时一方面考虑被试的总体分布情况, 另一方面也考虑了所选择的项目对于被试属性掌握概率的影响, 越能改变被试属性掌握概率的项目越容易被选到, 这样在测验长度较短时就有利于估计被试的属性掌握模式。

基于属性掌握概率的选题策略与基于属性掌握模式的选题策略的不同之处在于前者考虑了被试的属性掌握概率变化情况, 而后者只考虑被试的属性掌握模式估计值。被试的属性掌握概率是在区

间 0 到 1 之间的连续值, 被试作答每个题之后, 都会引起其属性掌握概率的变化, 因此, 在选题时将这个变化考虑进去比仅考虑属性掌握模式的变化更精细, 特别是在测验长度较短的测验中, 因为此时被试的属性掌握模式估计并不准确, 此时需要结合更多有用的信息来选题(这一点类似于 CAT 中的全局信息量选题, 可参考 Chang 和 Ying (1996)), 有利于提高测验的估计精度, 并且考虑了被试的属性掌握概率之后, 会对题库的使用均匀性有改善。

在本研究中没有考虑属性之间可能存在的层级关系(hierarchical relationship, 可参见 Leighton, Gierl, & Hunka, 2004), 但在实际的测验中, 属性之间有可能会存在层级关系。当属性之间存在层级关系时, 特别是很多实际测验中都涉及到的层级关系, PHKL, PPWKL, PWKL, HKL 和 MI 等策略的表现会是什么样的, 这需要进一步的深入研究。并且, 在实际的应用中, 需要根据测验的目的进行综合权衡, 选择合适的选题策略。

题库总体利用率较低是所涉及的各种选题策略都存在的问题, 这一点从陈平等(2011)的研究结果中也可以得到验证。就本研究来说, 在变长 CD-CAT 测验下, 未使用的项目数大多都大于 110, 也就是说, 题库中有超过一半以上的项目都没有被使用到, 只是使用了不到一半的项目, 这充分反映了这里所使用的选题策略在题库的利用率上的表现还有待进一步提高。

参 考 文 献

- Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *British Journal of Mathematical and Statistical Psychology*, 61, 493-513.
- Chang, H. H., Qian, J. H., & Ying, Z. L. (2001). A-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, 25(4), 333-341.
- Chang, H. H., & Ying, Z. L. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20(3), 213-229.
- Chang, H. H., & Ying, Z. L. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3), 211-222.
- 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, D. Z. (2004). *Comparison study of item selection strategies of computerized adaptive testing with the Samejima Graded Response Model*(Unpublished Master's

- thesis). Jiangxi Normal University.
- [陈德枝. (2004). *Samejima 等级反应模型下 CAT 选题策略比较研究*(硕士学位论文). 江西师范大学.]
- 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(2), 203–212.]
- Cheng, Y. (2008). *Computerized adaptive testing: new development and applications*(Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
- Cheng, Y. (2009a). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74(4), 619–632.
- Cheng, Y. (2009b). *Computerized adaptive testing for cognitive diagnosis*. Paper presented at the 2009 GMAC Conference on Computerized Adaptive Testing.
- 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(6), 902–913.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 34(1), 115–130.
- Henson, R. A. (2004). *Test discrimination and test construction for cognitive diagnostic models*(Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
- Hsu, C. L., Wang, W. C., & Chen, S. Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement*, 37(7), 563–582.
- Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement*, 71(2), 407–419.
- Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. New York: Cambridge University Press.
- 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(3), 205–237.
- 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.]
- McGlohen, M., & Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, 40(3), 808–821.
- Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods and applications*. New York: Guilford.
- Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(3), 337–350.
- Tu, D. B., Cai, Y., & Dai, H. Q. (2013). Item selection strategies and initial items selection methods of CD-CAT. *Journal of Psychological Science*, 36(2), 469–474.
- [涂冬波, 蔡艳, 戴海崎. (2013). 认知诊断 CAT 选题策略及初始题选取方法. *心理科学*, 36(2), 469–474]
- Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, 73(6), 1017–1035.
- Wang, C., Chang, H. H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods*, 44(1), 95–109.
- Wang, C., Chang, H. H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic CAT. *Journal of Educational Measurement*, 48(3), 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 the Annual Meeting of American Educational Research Association, Chicago, IL.

Item Selection Strategies Based on Attribute Mastery Probabilities in CD-CAT

LUO Zhaosheng¹; YU Xiaofeng^{1,2}; GAO Chunlei¹; LI Yujun¹; PENG Yafeng¹;
WANG Rui¹; WANG Yutong¹

(¹School of Psychology, Jiangxi Normal University, Nanchang 330022, China)

(²Computer Department, Bozhou Normal College, Bozhou 236800, China)

Abstract

Cognitive diagnostic computerized adaptive testing (CD-CAT) is a popular mode of online testing of cognitive diagnostic assessment (CDA). The key to a CD-CAT system is the item selection strategies. Some of the popular strategies are developed based on Kullback-Leibler information (KL), Shannon entropy (SHE) to select items in CD-CAT. Typically, during CD-CAT, these familiar methods would use a cutoff point to transform the attribute mastery probabilities' provisional value to binary values, but at the initial stage, the cutoff point method may lead to a larger deviation. A method that can take advantage of the probabilistic information with regard to attributes may offer a better alternative.

This paper proposed two item selection strategies based on the provisional value of the attribute mastery probabilities, as follows: (1) the first strategy, which is called as PPWKL (Posterior Probability Weighted Kullback-Leibler), is based on the KL information, and it can lead to maximum difference of the sum of attribute mastery probabilities, and it is weighted by the pattern's posterior probability as well as the difference of the attribute mastery probabilities between the $\hat{\alpha}$ and any possible latent state; (2) the PPWKL considers the fact that not all the patterns are equally likely, but overlooks the fact that the distances between different patterns and the current estimate are not all of equal importance. Therefore, the PPWKL can be weighed by the inverse of the distance between the $\hat{\alpha}$ and any possible latent state, which is called as PHKL (Posterior Hybrid Kullback-Leibler).

Then, three simulation studies were carried out, one was the fixed length of CD-CAT, and the second was the variant length CD-CAT, and the last was short length CD-CAT. Variant item selection strategies were taken into consideration in these studies, including KL, SHE, PWKL, HKL, MI, PPWKL and PHWKL. The results were compared in terms of pattern or attribute classification correct rate, item average exposure ratio, item maximum exposure ratio, item minimum exposure ratio, average test length, unused item number, number of items with exposure ratio over 20%, test overlap ratio. The simulation results indicate that: (1) the comprehensive performance of PPWKL and PHKL are better than other mentioned strategies in fixed and variant length CD-CAT; as to PHKL and MI, each has different strengths in short length CD-CAT; (2) PHKL and PPWKL can retain a good measurement accuracy, and also improve the utilization ratio of item pool.

Key words text CD-CAT; item selection strategies; attribute mastery probabilities; attribute mastery pattern

附录一 被试对属性掌握概率的期望后验估计(EAP)

首先对符号进行说明, 测验共考察 K 个属性, 在属性相互独立的情况下, 共有 2^K 种属性掌握模式。用 $x'_i = (x_{i1}, x_{i2}, \dots, x_{it})$ 表示被试 i 的作答得分向量, α_c 表示第 c 种属性掌握模式, $p(\alpha_c | x'_i)$ 表示被试 i 作答 t 个项目后属性掌握模式后验概率, 其计算公式为:

$$p(\alpha_c | x'_i) = \frac{L(x'_i | \alpha_c) p(\alpha_c)}{L(x'_i)}$$

其中: $L(x'_i | \alpha_c)$ 是被试 i 条件作答概率, 其计算公式为:

$$L(x'_i | \alpha_c) = \prod_{j=1}^t [p_j(\alpha_c, s_j, g_j)]^{x_{ij}} [1 - p_j(\alpha_c, s_j, g_j)]^{1-x_{ij}}$$

则被试 i 的边际似然函数为:

$$L(x'_i) = \sum_{c=1}^{2^K} L(x'_i | \alpha_c) p(\alpha_c)$$

被试 i 对属性 k 的掌握概率为 $p(\alpha_{ik})$, 其计算公式为:

$$p(\alpha_{ik} | x'_i) = \sum_{c=1}^{2^K} p(\alpha_c | x'_i) \alpha_{ck}$$

其中 α_{ck} 表示第 c 种属性掌握模式中属性 k 的值。

附录二 长度为 16, 属性间存在中等和较高相关时各策略的表现

表 6 各选题策略的分类准确率和题库使用均匀性(定长, 属性间存在中等相关)

选题策略	知识状态准确性		题库使用均匀性								统一量纲后的 综合指标
	PMR	MMR	N _{len}	χ^2	TOR	ER _{min}	ER _{max}	ER _{ave}	N _{no_use}	N _{20%}	
KL	0.384	0.860	16	88.012	0.520	0	0.970	0.080	132	27.5	12.57
SHE	0.938	0.989	16	84.971	0.504	0	1.000	0.080	109	27[1]	15.64
PWKL	0.927	0.985	16	86.301	0.511	0	0.963	0.080	103.5	27[1]	15.64
HKL	0.928	0.987	16	84.882	0.504	0	0.965	0.080	102.5	27[1]	15.69
MI	0.940	0.989	16	84.441	0.502	0	1.000	0.080	107	27[1]	15.68
PPWKL	0.926	0.983	16	80.924	0.484	0	0.962	0.080	100	28.5	15.72
PHKL	0.941[1]	0.991[1]	16	77.887[1]	0.469[1]	0	0.955[1]	0.080	94.5[1]	29	15.93

表 7 各选题策略的分类准确率和题库使用均匀性(定长, 属性间存在较高相关)

选题策略	知识状态准确性		题库使用均匀性								统一量纲后的 综合指标
	PMR	MMR	N _{len}	χ^2	TOR	ER _{min}	ER _{max}	ER _{ave}	N _{no_use}	N _{20%}	
KL	0.365	0.850	16	86.966	0.514	0	0.963	0.08	132.5	27.5	12.49
SHE	0.938	0.989	16	84.897	0.504	0	1.000	0.08	110.5	27.5	15.65
PWKL	0.933	0.987	16	85.506	0.507	0	0.955	0.08	104	27[1]	15.73
HKL	0.923	0.985	16	84.566	0.502	0	0.954	0.08	106	27.5	15.66
MI	0.939[1]	0.989[1]	16	84.898	0.504	0	1.000	0.08	112.5	27.5	15.64
PPWKL	0.928	0.985	16	80.271	0.481	0	0.955	0.08	103	30.5	15.70
PHKL	0.935	0.986	16	76.725[1]	0.463[1]	0	0.948[1]	0.08	100[1]	30.5	15.85

附录三 最大长度为 16 的变长测验, 属性间存在中等和较高相关时各策略的表现

表 8 各选题策略的分类准确率和题库使用均匀性(变长, 属性间存在中等相关)

选题策略	知识状态准确性		题库使用均匀性								统一量纲后的 综合指标
	PMR	MMR	N _{len}	χ^2	TOR	ER _{min}	ER _{max}	ER _{ave}	N _{no_use}	N _{20%}	
KL	0.435	0.867	15.902	96.305	0.561	0	0.984	0.080	141.5	26	11.88
SHE	0.810	0.964	9.867	121.233	0.655	0	1.000	0.049[1]	129.5	12.5[1]	15.08
PWKL	0.835	0.970	9.836	106.401	0.581	0	0.944	0.049[1]	116	13.5	15.51
HKL	0.830	0.968	9.839	105.738	0.577	0	0.950	0.049[1]	120.5	13.5	15.44
MI	0.809	0.965	9.757[1]	122.106	0.659	0	1.000	0.049[1]	134.5	12.5[1]	15.06
PPWKL	0.812	0.965	10.065	88.646[1]	0.493[1]	0	0.930[1]	0.050	110[1]	15	15.62
PHKL	0.837[1]	0.971[1]	10.126	88.964	0.495	0	0.937	0.051	111.5	16	15.68[1]

表 9 各选题策略的分类准确率和题库使用均匀性(变长, 属性间存在较高相关)

选题策略	知识状态准确性		题库使用均匀性								统一量纲后的 综合指标
	PMR	MMR	N _{len}	χ^2	TOR	ER _{min}	ER _{max}	ER _{ave}	N _{no_use}	N _{20%}	
KL	0.436	0.834	15.961	93.828	0.548	0	0.976	0.080	142	26	11.72
SHE	0.794	0.961	9.921	124.667	0.673	0	1.000	0.050	127	13	14.92
PWKL	0.805	0.965	9.691	103.462	0.565	0	0.910	0.048[1]	120	12.5[1]	15.48
HKL	0.808	0.964	9.681[1]	103.451	0.565	0	0.922	0.048[1]	116.5	13	15.47
MI	0.791	0.961	9.872	125.196	0.675	0	1.000	0.049	133.5	13	14.87
PPWKL	0.810	0.965	9.826	82.539	0.461[1]	0	0.908[1]	0.049	116	16	15.68
PHKL	0.815[1]	0.966[1]	9.900	82.356[1]	0.461[1]	0	0.912	0.049	111.5[1]	16	15.73[1]

附录四 长度分别为 4, 6, 8 和 10, 属性间存在较低、中等和较高相关时各策略的表现

表 10 三种选题策略的分类准确率和题库使用均匀性(短测验, 属性间存在较低相关)

选题策略	知识状态准确性		题库使用均匀性								统一量纲后的综合指标
	PMR	MMR	N _{ave_Len}	χ^2	TOR	ER _{min}	ER _{max}	ER _{ave}	N _{no_use}	N _{20%}	
PWKL	0.247	0.796	4	116.203	0.601	0	0.866	0.020	180.5	5	15.625
HKL	0.234	0.796		116.114	0.600	0	0.908	0.020	181	5	15.377
PHKL	0.251	0.799		99.163	0.515	0	0.897	0.020	182	7	15.672
PPWKL	0.227	0.792		111.372	0.576	0	0.898	0.020	181	6	15.155
MI	0.238	0.782		165.067	0.845	0	1.000	0.020	194	5.5	14.616
PWKL	0.512	0.886	6	124.855	0.654	0	0.930	0.030	171.5	8	15.162
HKL	0.507	0.885		121.054	0.635	0	0.941	0.030	170	8	15.175
PHKL	0.587	0.909		111.107	0.585	0	0.959	0.030	172.5	10	15.756
PPWKL	0.532	0.898		119.688	0.628	0	0.952	0.030	172	9.5	15.248
MI	0.587	0.905		133.572	0.698	0	1.000	0.030	187	9.5	15.338
PWKL	0.639	0.919	8	111.272	0.596	0	0.956	0.040	152.5	12	15.416
HKL	0.644	0.920		108.487	0.582	0	0.951	0.040	151	11	15.596
PHKL	0.686	0.927		103.592	0.558	0	0.960	0.040	148	11.5	15.922
PPWKL	0.645	0.918		109.081	0.585	0	0.955	0.040	149.5	11	15.589
MI	0.688	0.931		111.673	0.598	0	1.000	0.040	166	12.5	15.583
PWKL	0.773	0.951	10	103.368	0.566	0	0.965	0.050	136	15	15.611
HKL	0.779	0.953		100.332	0.551	0	0.951	0.050	137.5	16	15.652
PHKL	0.799	0.959		98.022	0.540	0	0.977	0.050	138.5	15	15.845
PPWKL	0.791	0.955		100.308	0.551	0	0.963	0.050	134.5	13.5	15.889
MI	0.799	0.958		103.448	0.567	0	1.000	0.050	143	15	15.685

注: 粗体显示的行对应的是这种条件下综合表现最好的选题策略

表 11 三种选题策略的分类准确率和题库使用均匀性(短测验, 属性间存在中等相关)

选题策略	知识状态准确性		题库使用均匀性								统一量纲后的综合指标
	PMR	MMR	N _{ave_Len}	χ^2	TOR	ER _{min}	ER _{max}	ER _{ave}	N _{no_use}	N _{20%}	
PWKL	0.289	0.812	4	116.259	0.601	0	0.865	0.020	180	5	15.600
HKL	0.292	0.816		115.268	0.596	0	0.896	0.020	181	5	15.627
PHKL	0.292	0.816		96.195	0.500	0	0.889	0.020	181.5	7	15.679
PPWKL	0.262	0.810		110.173	0.570	0	0.896	0.020	181	6	15.104
MI	0.259	0.779		165.766	0.849	0	1.000	0.020	194	5	14.328
PWKL	0.522	0.893	6	122.268	0.641	0	0.914	0.030	172.5	9	15.049
HKL	0.540	0.895		118.634	0.623	0	0.932	0.030	171.5	8	15.325
PHKL	0.574	0.898		106.958	0.564	0	0.937	0.030	170	9	15.658
PPWKL	0.524	0.893		116.769	0.613	0	0.944	0.030	171	9	15.119
MI	0.599	0.909		135.325	0.706	0	1.000	0.030	187	10	15.212
PWKL	0.643	0.923	8	109.212	0.586	0	0.944	0.040	151	13	15.360
HKL	0.653	0.926		106.427	0.572	0	0.947	0.040	152.5	12	15.539
PHKL	0.687	0.932		100.845	0.544	0	0.950	0.040	148.5	12	15.885
PPWKL	0.671	0.929		106.126	0.570	0	0.947	0.040	150	12	15.678
MI	0.702	0.938		113.348	0.606	0	1.000	0.040	166	14	15.482
PWKL	0.789	0.957	10	100.994	0.555	0	0.958	0.050	136	16	15.539
HKL	0.791	0.957		97.007	0.535	0	0.934	0.050	136.5	17	15.594
PHKL	0.795	0.958		93.092	0.515	0	0.967	0.050	134.5	14.5	15.814
PPWKL	0.787	0.952		96.229	0.531	0	0.955	0.050	133	14	15.744
MI	0.813	0.962		102.589	0.563	0	1.000	0.050	143	15.5	15.590

表 12 三种选题策略的分类准确率和题库使用均匀性(短测验, 属性间存在较高相关)

选题策略	知识状态准确性		题库使用均匀性								统一量纲后的综合指标
	PMR	MMR	N _{ave_Len}	χ^2	TOR	ER _{min}	ER _{max}	ER _{ave}	N _{no_use}	N _{20%}	
PWKL	0.339	0.832	4	116.504	0.602	0	0.867	0.020	180.5	5	15.542
HKL	0.314	0.821		115.078	0.595	0	0.894	0.020	181	5	15.181
PHKL	0.343	0.842		95.006	0.495	0	0.887	0.020	181.5	7.5	15.639
PPWKL	0.296	0.818		109.145	0.565	0	0.897	0.020	181	6	14.880
MI	0.296	0.819		166.729	0.853	0	1.000	0.020	194	5	14.284
PWKL	0.541	0.898	6	121.406	0.637	0	0.908	0.030	173	9.5	15.292
HKL	0.541	0.894		116.536	0.612	0	0.920	0.030	170.5	9.5	15.348
PHKL	0.558	0.899		105.857	0.559	0	0.937	0.030	172.5	9.5	15.638
PPWKL	0.546	0.898		115.432	0.607	0	0.937	0.030	172.5	8.5	15.490
MI	0.582	0.910		136.190	0.711	0	1.000	0.030	187	9.5	15.278
PWKL	0.678	0.931	8	108.552	0.582	0	0.945	0.040	151	13	15.557
HKL	0.669	0.929		104.606	0.563	0	0.938	0.040	152.5	12.5	15.598
PHKL	0.679	0.931		99.818	0.539	0	0.945	0.040	150.5	11.5	15.838
PPWKL	0.666	0.931		104.636	0.563	0	0.938	0.040	151.5	12	15.631
MI	0.701	0.938		114.624	0.613	0	1.000	0.040	166	14	15.415
PWKL	0.790	0.957	10	99.742	0.548	0	0.956	0.050	139	16	15.649
HKL	0.792	0.958		95.886	0.529	0	0.927	0.050	139	16.5	15.734
PHKL	0.795	0.959		91.514	0.507	0	0.954	0.050	136	16.5	15.832
PPWKL	0.791	0.957		94.751	0.523	0	0.942	0.050	132.5	15.5	15.837
MI	0.804	0.961		103.703	0.568	0	1.000	0.050	144.5	16	15.587