心理学报 2016, Vol. 48, No.7, 903−914
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2016.00903

认知诊断计算机化自适应测验中新的选题策略：结合项目区分度指标*

郭 磊[1,2,3] 郑蝉金[4] 边玉芳[5] 宋乃庆[3,6] 夏凌翔[1]

([1]西南大学心理学部, 重庆 400715) ([2]西南大学统计学博士后科研流动站, 重庆 400715)
([3]中国基础教育质量监测协同创新中心西南大学分中心, 重庆 400715) ([4]江西师范大学心理学院, 南昌 330022)
([5]北京师范大学中国基础教育质量监测协同创新中心, 北京 100875) ([6]西南大学基础教育研究中心, 重庆 400715)

摘 要 当前国内外大部分认知诊断计算机化自适应测验(CD-CAT)主要采用 PWKL 作为选题策略进行研究。PWKL 结合后验分布信息对 KL 指标进行加权, 提高了判准率, 但该方法仅利用个体层面信息加权, 忽视了项目本身能够提供的信息, 属于单源指标。本研究结合认知诊断中的项目区分度信息, 对 PWKL 进行修正, 提出了 4 种新的多源选题策略：GIDPWKL、AIDPWKL、CIDPWKL 和 KLEDPWKL 方法, 并在加入曝光控制下与 PWKL 和互信息法(MIM)进行比较。模拟研究结果表明：(1)在定长测验情景下的绝大多数实验结果表明, 测验长度越短, 新方法的判准率越高。平均属性/模式判准率最高的是 GIDPWKL, 之后是 AIDPWKL, 而 CIDPWKL、KLEDPWKL 和 MIM 方法的优势随实验条件不同而不同。(2)在定长测验情景下的绝大多数实验结果表明, 题目质量越高, 新方法的优势越明显。(3)Q 矩阵结构的复杂性会影响不同选题策略的表现。(4)在变长测验情景下, 4 种新方法和 MIM 的平均测验长度均要低于 PWKL 方法, 表现最好的是 GIDPWKL 方法。因此, 若实际测验情景与本研究的模拟情景相似, 推荐 GIDPWKL 方法。

关键词 认知诊断计算机化自适应测验; 选题策略; 项目区分度; 曝光控制

分类号 B841

1 引言

近些年, 国内外心理测量着重于形成性评估(*formative assessment*), 它要求提供给教育工作者和学生更多的测验信息, 以帮助教师教学和学生改进。基于此, 认知诊断评估(*Cognitive Diagnostic Assessment,* CDA)通过测查学生是否掌握了某一知识领域内的认知属性和技能而蓬勃发展。计算机化适应性测验(*Computerized adaptive testing,* CAT)是量体裁衣式的新型测验形式, 在美国得以广泛运用, 例如研究生入学考试(*Graduate Record Examination*, GRE)、美国护士资格考试(*The National Council of State Boards of Nursing,* NCSBN)等。和传统纸笔测验相比, CAT 测验长度更短, 能力估计精度更高。将 CDA 和 CAT 结合兼具二者优势, 能够快速精准地得到学生知识状态(*Knowledge State,* KS), 该测验形式被称作认知诊断计算机化自适应测验(*Cognitive Diagnostic Computerized adaptive testing,* CD-CAT; Cheng, 2009)。和传统 CAT 一样, CD-CAT 同样具有 5 个重要组成部分(郭磊, 2014)。其中, 研究最多的当属选题策略(Cheng, 2009; Wang, Chang, & Douglas, 2012; Wang, Chang, & Huebner, 2011; Xu, Chang, & Douglas, 2003; 毛秀珍, 辛涛, 2011; 尚志勇, 丁树良, 2011; 汪文义, 丁树良, 宋丽红, 2014), 因为选题策略的好坏不仅影响测验效率, 还影响题库的使用情况, 通常被视作 CAT 系统的核心成分。CD-CAT

收稿日期: 2014−10−08

* 中央高校基本科研业务费专项资金资助, 项目批准号: SWU1409433。教育部人文社会科学研究青年基金项目, 项目批准号: 15YJC190003。自立人格与社区心理(PI)研究室科研基金资助。

通讯作者: 郭磊, E-mail: happygl1229@swu.edu.cn

在形成性评估中的一个重要作用是让教师能在课堂中快速地掌握学生的学习动态。例如, 上课前几分钟, 教师用较短的测验可以初步掌握学生的知识状态, 便于接下来有针对性地进行课堂教学。因此, 如何能在较短测验中准确地估计学生的知识状态尤为重要, 这就跟选题策略息息相关。

目前, 在众多选题策略中, 效果较好并且应用较多的是后验加权库尔贝克-莱布勒信息量法(*Posterior-Weighted Kullback-Leibler,* PWKL; Cheng, 2009)。该方法将每次更新后的被试知识状态的后验概率作为权重融合到库尔贝克-莱布勒信息量(*Kullback-Leibler information*)指标中, 大大提高了被试 KS 的估计精度。但 PWKL 指标仅从个体层面(*person-level*)[1]对 KL 信息量进行加权, 并未考虑项目质量对估计精度的影响, 属于单源指标(*single-source index*)。在经典测验理论(*Classical Test Theory*, CTT)和项目反应理论(*Item Response Theory*, IRT)中, 题目的质量决定着测验的质量, 而题目质量中比较关键的指标之一就是项目区分度(*item discrimination*)。项目区分度较高, 表明该题目能够较好地区分出高能力被试和低能力被试, 这也是测验编制所追求的目标之一。正是基于项目区分度如此重要的作用, Chang 和 Ying (1999)在传统 CAT 中提出了著名的 a 分层选题法。他们建议在测验初期使用区分度较低的项目, 因为测验初期对被试能力值的估计还不是很精确, 无需使用项目信息量较高的项目, 等到测验后期需要对被试能力值进行精确估计时, 再使用高区分度的项目。同样, 在 CDA 领域, 我们仍需考虑项目质量的问题。若项目区分度较高, 则题目能够区分出掌握该题目所考察属性的被试和未掌握该题目所考察属性的被试的能力(*power*)就较大(Rupp, Templin, & Henson, 2010)。可以看出, 不论测验理论是 CTT, IRT, 还是 CDA, 项目区分度均是用来衡量题目能否有效区分出高能力被试和低能力被试(或不同知识状态)的关键指标。Rupp 等(2010)书中第 13 章总结了当前 CDA 中常用的一些项目区分度指标, 主要包括两大类: 一类是基于 CTT 思想提出的项目区分度指标, 另一类是基于 KL 信息量提出的项目区分度指标。另一方面, Wang (2013)基于互信息理论提出了互信息选题方法(*Mutual Information Method*, MIM), 模拟研究结果表明 MIM 在大多数实验条件下的判准率要优于 PWKL, 特别是在测验长度较短(5 题)时, 但 MIM 并未考虑项目区分度信息。

与传统 CAT 一样, 在 CD-CAT 的实际应用中, 不容忽视的一个重要问题是项目曝光问题。当前 CD-CAT 着重于测量精度的实现, 较少考虑项目曝光问题, 导致题库使用极其不均匀, 优质题目曝光十分严重(Wang et al., 2011)。在选题策略的研究中, 估计精度和项目曝光度往往是相互制约的。因此, 要全面考察一个选题指标的好坏, 并与实际应用情景相符, 对项目过度曝光的控制是很重要的。但即使是在 Wang (2013)的研究中, 也未曾考虑曝光控制问题。

查阅国内外相关文献, 将区分度信息纳入 CD-CAT 选题过程的研究并不多, 据我们所知, 汪文义等(2014)基于 CTT 的思想将项目区分度信息纳入选题策略中进行了研究, 但该方法不仅在加权形式上与 Rupp 等(2010)提出的加权形式不同, 而且也不是对 PWKL 指标的加权。除此之外, 尚未见到基于 KL 信息量提出的项目区分度加权指标。因此, 本文以确定性输入, 噪音"与"门(*the Deterministic Inputs, Noisy "and" Gate*, DINA)模型为例(DINA 模型是认知诊断研究中最常使用的模型, 由于 DINA 模型参数较少、简单易懂、方便解释, 因此成为了许多研究者修正和拓展的基础模型), 将项目区分度信息融入选题策略中, 对 PWKL 指标进行修正, 提出 4 个新的多源选题指标(*multiple-source index*), 分别称作: 基于经典测验理论的项目区分度加权法(*CTT-analogous item-discrimination-posterior-weighted Kullback-Leibler*, CIDPWKL)、基于 KL 信息量的全局项目区分度加权法(*KLI-based global-item-discrimination-posterior-weighted Kullback-Leibler,* GIDPWKL)、基于 KL 信息量的属性层面项目区分度加权法(*KLI-based attribute-specific-item-discrimination-posterior-weighted Kullback-Leibler,* AIDPWKL)、以及使用汪文义等(2014)提出的权重加权方法(本文将该方法称作 KLEDPWKL 法), 并在加入曝光控制技术下, 将 4 种新方法和 PWKL、MIM 在不同实验条件下进行系统比较, 以验证新方法的优越性。

本文按如下方式组织。首先对 DINA 模型进行简单介绍, 其次对 PWKL、4 种新的选题方法、以及 MIM 方法进行详细介绍。第四部分和第五部分分别进行两个模拟研究, 最后部分为本文的研究结

[1] 个体层面是指 PWKL 指标在 KL 指标基础上融入的被试 KS 后验概率信息。

论, 讨论及展望。

2　DINA 模型简介

DINA 模型是具有显式项目特征函数的诊断模型(Haertel, 1989; Junker & Sijtsma, 2001), 其数学表达式为：

$$P(X_{ij}=1|\alpha_i)=(1-s_j)^{\eta_{ij}}g_j^{1-\eta_{ij}} \quad (1)$$

其中, $\eta_{ij}=\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}$ 是潜在反应指标, 若被试 i 掌握了项目 j 所考察的全部属性, 则 $\eta_{ij}=1$, 否则 $\eta_{ij}=0$。α_{ik} 表示被试 i 是否掌握了第 k 个属性(k=1,2,…,K), $\alpha_{ik}=1$表示掌握, $\alpha_{ik}=0$ 表示未掌握。q_{jk} 表示项目 j 是否考察了属性 k, 若 $q_{jk}=1$表示考察了, $q_{jk}=0$ 表示未考察。s_j 是项目的失误参数, 它表示被试掌握了项目 j 考察的全部属性却答错的概率; g_j 是项目的猜测参数, 它表示被试未全部掌握项目 j 考察的属性却答对的概率。一个质量较好的项目, 应该具有较小的 s_j 和 g_j 参数, 并且要满足 $1-s_j>g_j$ (Templin & Henson, 2006)。

3　相关选题策略介绍

Cheng (2009)在提出了 PWKL 指标的同时还提出了 HKL 指标, 并通过模拟研究发现两者的表现相差无几。由此表明, 仅仅在 PWKL 指标中融入 KS 之间距离的加权做法收效不大。受 Cheng 的研究和项目区分度效能的启发, 一个更加合适的权重应该是 CDA 中的项目区分度(请见本文 3.3 和 3.4 部分介绍)。因为项目区分度不仅包含了项目是如何考察 K 个属性的信息(即 Q 矩阵中的 $\boldsymbol{q_j}$ 向量), 还包含了项目参数以及被试 KS 之间不同组合提供的信息, 提供的信息更加丰富。下面将分别对本文涉及的 6 种选题方法进行介绍。

3.1　PWKL 指标

KL 信息量是衡量两个概率分布差异的统计量, 具有非对称性。若两个分布相差越小, KL 值就越小, 特别地, 当且仅当两个分布一样时, KL 值等于 0。令 j 表示剩余题库 S_j 中的项目, x 为第 $t+1$ 个项目的作答反应, $\hat{\alpha}_i^t$ 为被试 i 作答完 t 个项目后的知识状态估计值, α_l 为知识状态集合中的任意一种 ($l=1,2,\cdots,2^K$), 则项目 j 的 KL 信息量为：

$$KL_j(\hat{\alpha}_i^t)=\sum_{l=1}^{2^K}\left[\sum_{x=0}^{1}\log\left(\frac{P(X_{ij}=x|\hat{\alpha}_i^t)}{P(X_{ij}=x|\alpha_l)}\right)\cdot P(X_{ij}=x|\hat{\alpha}_i^t)\right] \quad (2)$$

但 KL 选题策略中 KL 指标是计算当前估计的 KS 与所有可能 KS 之间的 KL 距离的等权之和, 该做法不太合理。Cheng (2009)认为, 随着被试作答项目数量的增长, 被试能提供更多的诊断信息, 因此各种可能的KS之间的后验概率差异会越来越大, 即该被试从属于某类 KS 的可能性会逐渐增大。于是, 她利用后验概率对 KL 信息量进行修正, 提出了 PWKL 方法, PWKL 指标为：

$$PWKL_j(\hat{\alpha}_i^t)=\sum_{l=1}^{2^K}\left\{\left[\sum_{x=0}^{1}\log\left(\frac{P(X_{ij}=x|\hat{\alpha}_i^t)}{P(X_{ij}=x|\alpha_l)}\right)\cdot P(X_{ij}=x|\hat{\alpha}_i^t)\right]\cdot\pi_{i,t}(\alpha_l)\right\} \quad (3)$$

其中, $\pi_{i,t}(\alpha_l)\propto L(x_1,x_2,\cdots,x_t|\alpha_l)\cdot p(\alpha_l)$ 表示知识状态 α_l 的后验概率。$L(x_1,x_2,\cdots,x_t|\alpha_l)$ 为似然函数, $p(\alpha_l)$ 为 KS 的先验概率, 一般取均匀分布, 即 $p(\alpha_l)=1/2^K$, 其余符号同前。PWKL 指标选择题目的标准是：从剩余题库中选择具有最大 PWKL 信息量的题目给被试作答。

3.2　CIDPWKL 指标

在CTT中, 二值计分的项目区分度 d_j的一种求法为：$d_j=p_u-p_l$。其中, p_u 和 p_l 分别是高分组和低分组[2]的通过率, 其思想是尽可能拉大真正能够答对被试与真正未能答对被试之间的差距, 差距越大, 项目区分度越大。但在 CDA 中, 测评的结果是以多维离散潜在变量呈现, 不像 CTT 那样能够基于总分找到高分组和低分组。因此, Rupp 等(2010)仿照 CTT 的思想, 定义了 CDA 中的项目区分度：$d_j=p_{\alpha_h}-p_{\alpha_l}$。其中, p_{α_h} 表示掌握题目 j 考察的较多属性的正确作答概率, p_{α_l} 表示掌握题目 j 考察的较少属性的正确作答概率。其含义是：“该项目区别掌握‘较多’属性被试和掌握‘较少’属性被试的能力”, 因此, 若一道题目能够较好地区分掌握所有属性和未掌握所有属性的被试, 则 d_j 的值就会很大。在 DINA 模型中有：$p_{\alpha_h}=(1-s_j)^1g_j^{1-1}=1-s_j$ 和 $p_{\alpha_l}=(1-s_j)^0g_j^{1-0}=g_j$。由此有,

$$d_j=(1-s_j)-g_j=1-(s_j+g_j) \quad (4)$$

可以看出, 一个题目的猜测参数和失误参数的和越小, 该题目的区分度就越大。因此, 结合了基于 CTT 思想推导出的项目区分度后, CIDPWKL 指标的公式如下：

[2] 高分组指按总分排名位于前 25%或 30%的所有被试, 低分组指按总分排名位于后 25%或 30%的所有被试。

$$CIDPWKL_j = (1-(s_j+g_j))\cdot PWKL_j \quad (5)$$

CIDPWKL 指标选择题目的标准是：从剩余题库中选择具有最大CIDPWKL信息量的题目给被试作答。

3.3 GIDPWKL 指标

Henson 和 Douglas (2005)指出，如果某个项目能够很好地区分相似的 KS，那么它也能够较好地区分差异较大的 KS。基于此，他们提出了全局项目区分度指标 CDI (*Cognitive Diagnostic Index*)。题目 j 的 CDI 计算公式如下：

$$C_j = \frac{\sum_{u\neq v} h(\boldsymbol{\alpha}_{\mathbf{u}},\boldsymbol{\alpha}_{\mathbf{v}})^{-1}\cdot D_{j,uv}}{\sum_{u\neq v} h(\boldsymbol{\alpha}_{\mathbf{u}},\boldsymbol{\alpha}_{\mathbf{v}})^{-1}} \quad (6)$$

其中，C_j 是 CDI 的缩写，$\boldsymbol{\alpha}_{\mathbf{u}}$ 和 $\boldsymbol{\alpha}_{\mathbf{v}}$ 表示两个不同的 KS，$h(\boldsymbol{\alpha}_{\mathbf{u}},\boldsymbol{\alpha}_{\mathbf{v}})$ 为两个 KS 之间的汉明距离(*hamming distance*)。例如，若 $\boldsymbol{\alpha}_{\mathbf{u}}=(0,0,1,1)$ 和 $\boldsymbol{\alpha}_{\mathbf{v}}=(1,0,1,0)$，那么汉明距离 $h(\boldsymbol{\alpha}_{\mathbf{u}},\boldsymbol{\alpha}_{\mathbf{v}})=|0-1|+|0-0|+|1-1|+|1-0|=2$。若用 D_j 表示两两 KS 之间 KL 信息量在项目 j 上构成的方阵(其大小为 $2^K\times 2^K$)，则 $D_{j,uv}$ 为 D_j 方阵中与 $\boldsymbol{\alpha}_{\mathbf{u}}$ 和 $\boldsymbol{\alpha}_{\mathbf{v}}$ 相关的元素。因此，结合了全局项目区分度指标后，GIDPWKL 指标的公式如下：

$$GIDPWKL_j = C_j\cdot PWKL_j \quad (7)$$

GIDPWKL 指标选择题目的标准是：从剩余题库中选择具有最大 GIDPWKL 信息量的题目给被试作答。

3.4 AIDPWKL 指标

Henson, Roussos, Douglas 和 He (2008)提出了属性层面(*attribute-specific*)的项目区分度指标 C_{j_k}，该指标表示项目 j 能够区分掌握属性 k 和未掌握属性 k 的效能(*power*)。基于 D_j 矩阵，C_{j_k} 关注只在属性 k 上有差异的那些元素。例如，测验考察 3 个属性，那么在第一个属性上有差异的元素共包括 8 组：000 和 100、100 和 000、010 和 110、001 和 101、110 和 010、101 和 001、011 和 111、111 和 011。类似地，可以在 D_j 矩阵中找出在第二个和第三个属性上有差异的元素。由此，项目 j 在第 k 个属性上的区分度计算公式如下：

$$C_{j_k} = \frac{1}{2^K}\sum_{all\ relevant\ cells} D_{j,uv} \quad (8)$$

若项目 i 考察的属性个数多于项目 j，则项目 i 的属性区分度个数也要多于项目 j，因此，项目 i 能够贡献的效能就越多。基于此，结合了属性层面的项目区分度指标 C_{j_k} 后，AIDPWKL 指标的公式如下：

$$AIDPWKL_j = \sum_{k=1}^{K} C_{j_k}\cdot PWKL_j \quad (9)$$

AIDPWKL 指标选择题目的标准是：从剩余题库中选择具有最大 AIDPWKL 信息量的题目给被试作答。

3.5 MI 指标

给定两个随机变量 X 和 Y，互信息为两变量边际分布的乘积 $f(x)f(y)$ 与它们联合分布 $f(x,y)$ 的 KL 距离，其表达式为：

$$I(X;Y) = \sum_x\sum_y f(x,y)\log\left[\frac{f(x,y)}{f(x)f(y)}\right] \quad (10)$$

$I(X;Y)$ 测量了 X 和 Y 之间的依赖程度，X 能够提供给 Y 越多信息(或 Y 能够提供给 X 越多信息【互信息的对称性】)，$I(X;Y)$ 越大。在 CD-CAT 中，互信息可以看作是临近两次后验概率分布的期望 KL 距离(*expected KL distance*)。Wang (2013)将 KS 为 $\boldsymbol{\alpha}$ 的被试作答完 t–1 题的后验概率 $\pi(\boldsymbol{\alpha}\,|\,\mathbf{x}_{t-1})$ 替换公式(10)中的 $f(y)$，将给定作答完 t–1 题在第 t 题上反应的二项分布 $p(x_t\,|\,\mathbf{x}_{t-1})$ 替换公式(10)中的 $f(x)$，并通过简单的运算，得到了互信息指标为：

$$MI_j = \sum_{x=0}^{1} p\left(X_t=x\,|\,\mathbf{x}_{t-1}\right)\left[\sum_{l=1}^{2^K}\pi\left(\boldsymbol{\alpha}_l\,|\,\mathbf{x}_{t-1},X_t=x\right)\log\left(\frac{\pi\left(\boldsymbol{\alpha}_l\,|\,\mathbf{x}_{t-1},X_t=x\right)}{\pi\left(\boldsymbol{\alpha}_l\,|\,\mathbf{x}_{t-1}\right)}\right)\right] \quad (11)$$

其中，$p\left(X_t=x\,|\,\mathbf{x_{t-1}}\right)=\sum_{l=1}^{2^K} p(X_t=x\,|\,\boldsymbol{\alpha}_l)\pi\left(\boldsymbol{\alpha}_l\,|\,\mathbf{x}_{t-1}\right)=\frac{\sum_{l=1}^{2^K} p(\mathbf{x}_{t-1},X_t=x\,|\,\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)}{\sum_{l=1}^{2^K} p(\mathbf{x}_{t-1}\,|\,\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)}$，

$$\pi\left(\boldsymbol{\alpha}_l\,|\,\mathbf{x}_{t-1},X_t=x\right)=\frac{p(\mathbf{x}_{t-1},X_t=x\,|\,\boldsymbol{\alpha}_l)\,p(\boldsymbol{\alpha}_l)}{\sum_{l=1}^{2^K} p(\mathbf{x}_{t-1},X_t=x\,|\,\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)}$$

此时，规定：$h_1=\sum_{l=1}^{2^K} p\left(\mathbf{x}_{t-1}\,|\,\boldsymbol{\alpha}_l\right)p\left(\boldsymbol{\alpha}_l\right)$，$h_2=\sum_{l=1}^{2^K} p(\mathbf{x}_{t-1},X_t=x\,|\,\boldsymbol{\alpha}_l)\,p\left(\boldsymbol{\alpha}_l\right)$。那么，原始公式(11)可改写为：

$$MI_j = \sum_{x=0}^{1}\frac{h_2}{h_1}\left[\sum_{l=1}^{2^K}\frac{p(\mathbf{x}_{t-1},X_t=x\,|\,\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)}{h_2}\log\frac{p(X_t=x\,|\,\boldsymbol{\alpha}_l)h_1}{h_2}\right]$$

$$=\frac{1}{h_1}\left[\sum_{x=0}^{1}\sum_{l=1}^{2^K}p(\mathbf{x}_{t-1},X_t=x\,|\,\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)\log\frac{p(X_t=x\,|\,\boldsymbol{\alpha}_l)}{h_2}\right.$$

$$\left.+\log h_1\sum_{x=0}^{1}\sum_{l=1}^{2^K}p(\mathbf{x}_{t-1},X_t=x\,|\,\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)\right] \quad (12)$$

其中，$\sum_{x=0}^{1}\sum_{l=1}^{2^K}p(\mathbf{x}_{t-1},X_t=x\,|\,\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)=\sum_{l=1}^{2^K}p(\mathbf{x}_{t-1}\,|\,\boldsymbol{\alpha}_l)$ $p(\boldsymbol{\alpha}_l)=h_1$，因此，公式(12)可进一步改写为：

$$MI_j$$

$$=\sum_{x=0}^{1}\frac{h_2}{h_1}\left[\sum_{l=1}^{2^K}\frac{p(\mathbf{x}_{t-1},X_t=x\,|\,\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)}{h_2}\log\frac{p(X_t=x\,|\,\boldsymbol{\alpha}_l)h_1}{h_2}\right]$$

$$=\frac{1}{h_1}\left[\sum_{x=0}^{1}\sum_{l=1}^{2^K}p(\mathbf{x}_{t-1},X_t=x\,|\,\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)\log\frac{p(X_t=x\,|\,\boldsymbol{\alpha}_l)}{h_2}\right.$$

$$\left.+h_1\cdot\log h_1\right] \quad (13)$$

MI 指标选择题目的标准是：从剩余题库中选择具有最大 MI 值的题目给被试作答。

3.6　KLEDPWKL 指标

汪文义等(2014)提出了 KLED 选题方法，在 DINA 模型下可以将其换算为：

$$KLED_j=\arg\max_{j\in R_i^t}(w_j\cdot\pi_j(i,t)\cdot(1-\pi_j(i,t))) \quad (14)$$

其中，R_i^t 表示被试 i 作答完 t 题后的剩余题库，$\pi_j(i,t)=\sum_{l:\,a_l'q_j\geqslant q_j'q_j}\pi(\alpha_l\,|\,i,t)$ 表示掌握题目 j 所有属性的知识状态的后验概率之和，$1-\pi_j(i,t)$ 表示至少有一个属性未掌握的知识状态后验概率之和。$w_j=(1-\mathrm{s}_j-\mathrm{g}_j)\left(\log\frac{1-s_j}{\mathrm{g}_j}+\log\frac{1-\mathrm{g}_j}{s_j}\right)$ 为权重，它是题目参数的函数。

基于此，将 KLED 中的权重 w_j 与 PWKL 结合后的 KLEDPWKL 计算公式如下：

$$KLEDPWKL_j=w_j\cdot PWKL_j \quad (15)$$

KLEDPWKL 指标选择题目的标准是：从剩余题库中选择具有最大 KLEDPWKL 值的题目给被试作答。

4　模拟研究 1

4.1　研究目的

采用蒙特卡洛模拟方法，在固定测验长度条件下比较 6 种选题策略：PWKL 法、CIDPWKL 法、GIDPWKL 法、AIDPWKL 法、KLEDPWKL 法和 MIM 法，重点考察不同选题策略对被试 KS 估计精度的影响。其中，PWKL 法作为基线。所有程序采用 Matlab 2012b 进行编程。需要指出的是，在进行新方法的编程时，可以提前将 D_j 矩阵以及相应的 C_j 和 C_{j_k} 计算好。在每次使用项目区分度信息时，直接从该矩阵中调取即可，这样可以有效提高选题速度。除了一开始计算 D_j 矩阵等需要较短的时间以外，整个选题过程所用时间和 PWKL 所用时间基本相同。

4.2　研究设计

本研究中的 Q 矩阵包括两种结构(如表 1 所示)：(1)简单结构(S)中包含 800 题，考察相互独立的 6 个属性，每个属性有 20%的项目考察，每个项目至少考察一个属性；(2)复杂结构(C)中每个属性有 50%的项目考察，其余条件同简单结构。项目的 s 参数和 g 参数越小，项目的质量越高，本研究的题目质量包括两个水平：(1)高质量题目的 s 参数和 g 参数被定义为平均数为 0.1，波动范围为 0.05，因此均从 U(0.05,0.15)中抽取；(2)低质量题目的 s 参数和 g 参数的波动范围与高质量题目相同，其平均数被定义为 0.2，因此均从 U (0.15,0.25)中抽取。测验长度为 5 题和 10 题，分别表示较短测验长度和中等测验长度(Wang, 2013)。1000 个被试 KS 的真值按照高阶 DINA 模型生成(Wang, 2013)，其中高阶能力值 θ 从标准正态分布 N (0,1)中抽取，斜率 λ_{1k} 从对数正态分布中抽取，截距 λ_{0k} 从标准正态分布中抽取。在高阶 DINA 模型中，被试 i 在属性 k 上的掌握情况为：

$$\alpha_{ik}=\begin{cases}1 & if\ P(\alpha_{ik}=1\,|\,\theta_i)\geqslant 0.5\\ 0 & otherwise\end{cases} \quad (16)$$

因此，本研究共包括 2(Q 矩阵结构) × 2(题目质量) × 2(测验长度) × 6(选题策略) = 48 种实验条件。每个实验条件重复 30 次以减小随机误差。

表 1　简单结构和复杂结构中每个属性的项目比例

Q 矩阵结构	属性 1	属性 2	属性 3	属性 4	属性 5	属性 6
简单结构的项目比例	20%	20%	20%	20%	20%	20%
复杂结构的项目比例	50%	50%	50%	50%	50%	50%

Wang 等(2011)提出了两种 CD-CAT 中的曝光控制方法：限制进度法(*Restrictive Progressive method*, RP)和限制阈值法(*Restrictive Threshold method*, RT)。估计精度和项目曝光度往往是相互制约的，比起 RT 法，RP 法在平衡估计精度和项目曝

光度方面做得更好, 而且本文的目的也并非比较不同曝光控制方法之间的差异, 因此, 本文借用 RP 法的思想作为曝光控制方法(由于篇幅所限, RP 法请参见相关文献)。具体而言, 当采用本文提出的新选题策略时, 按 RP 法的思想将 Wang 等(2011)提出的原始公式中的 PWKL 指标分别替换成 CIDPWKL、GIDPWKL、AIDPWKL、KLEDPWKL 和 MIM 指标, 从而实现对题目的曝光控制。其中, 允许的最大曝光率设置为 0.2, $\beta = 2$。

4.3 评价指标

(1)平均属性判准率(*Average Attribute Correct Classification Rate*, AACCR)

$$AACCR = \frac{\sum_{k=1}^{K}\left(\sum_{i=1}^{N} g_{ik} / N\right)}{K} \tag{17}$$

AACCR 考察所有属性平均返真性情况。假设测验共考察了 K 个属性, 有 N 个被试参加了测验, 现在考察第 k 个属性, 如果被试 i 掌握(未掌握)第 k 个属性, 今诊断其掌握(未掌握)该属性, 则表明对第 k 个属性判准了一次, 记为 $g_{ik} = 1$, 否则 $g_{ik} = 0$。

(2)模式判准率(*Pattern Correct Classification Rate*, PCCR)

$$PCCR = \sum_{i=1}^{N} n_i / N \tag{18}$$

PCCR 考察被试属性掌握模式($\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_K)$)的返真性。假设测验共考察了 K 个属性, 有 N 个被试参加了测验, 被试 i 真实的属性掌握向量记为 $\mathbf{X_i}$, 但把该被试归类为 $\mathbf{Z_i}$, 如果有 $\mathbf{X_i} = \mathbf{Z_i}$, 记 $n_i = 1$; 否则记 $n_i = 0$。

(3)测验重叠率

测验重叠率被定义为两个随机抽取的被试作答相同题目的期望数除以测验长度, 计算公式如下:

$$T = \sum_{j=1}^{J} M_j (M_j - 1) \Big/ (L \cdot N \cdot (\mathrm{N}-1)) \tag{19}$$

其中, T 表示测验重叠率, M_j 是第 j 个题目被调用的次数, J 是题库大小, L 是测验长度, N 是被试人数。测验重叠率越小, 说明两个随机抽取的被试作答相同题目的比例越小。

(4)题库使用均匀性指标, 卡方 χ^2

$$\chi^2 = \sum_{j=1}^{J} (\mathrm{er}_j - L / J)^2 \Big/ (L / J) \tag{20}$$

其中, er_j 是第 j 个题目的曝光率, 其大小等于作答题目 j 的被试人数除以参加测验的总被试人数, 其余符号定义同测验重叠率指标。χ^2 越小越好, χ^2 越小, 说明整个题库使用越均匀。

除上述指标外, 研究结果还记录了题库中未使用的题目数量。

4.4 研究结果

表 2 和表 3 分别是简单结构和复杂结构中 6 种选题策略在不同测验长度和不同题目质量下的平均属性判准率和模式判准率。由结果可知, 在各种实验条件下, 与 PWKL 方法相比, 其余 5 种选题策略的 AACCR 和 PCCR 均有不同程度的提高。整体上来看, 表现最好的是 GIDPWKL 指标, 其判准率的增长幅度均是最大的, 如表 2 和表 3 中粗体数值所示。表现次之的是 AIDPWKL 方法。而 CIDPWKL、KLEDPWKL 和 MIM 方法并未呈现出一致的表现结果。例如在简单结构×5 题×高质量题目实验条件下, KLEDPWKL 的判准率要高于 CIDPWKL 和 MIM, 但在简单结构×5 题×低质量题目实验条件下, MIM 的判准率要高于其余两种方法。具体来看, 在绝大多数实验条件下, 测验长度越短, GIDPWKL 和 AIDPWKL 方法的优势越明显, 且均要优于其余方法。例如, 在简单结构中题目质量较高时, 测验长度为 5 题条件下, 与 PWKL 相比, GIDPWKL 和 AIDPWKL 的 AACCR 值分别提高了 0.025 和 0.019; PCCR 值分别提高了 0.051 和 0.049; 当测验长度增加至 10 题时, GIDPWKL 和 AIDPWKL 的 AACCR 值分别提高了 0.017 和 0.014; PCCR 值分别提高了 0.033 和 0.022。而 CIDPWKL、KLEDPWKL 和 MIM 之间并没有展现出一致的优势结果, 但三者的表现相差无几。

大部分实验结果表明, 题目质量越高, GIDPWKL 和 AIDPWKL 方法的优势越明显, 且均要优于其余方法。例如, 在简单结构中测验长度为 5 题时, 高题目质量条件下, 与 PWKL 相比, GIDPWKL 和 AIDPWKL 的 AACCR 值分别提高了 0.025 和 0.019; PCCR 值分别提高了 0.051 和 0.049; 低题目质量条件下, GIDPWKL 和 AIDPWKL 的 AACCR 值分别提高了 0.024 和 0.005; PCCR 值分别提高了 0.037 和 0.032。而 CIDPWKL、KLEDPWKL 和 MIM 之间并没有展现出一致的优势结果, 但三者的表现相差无几。

Q 矩阵结构的复杂性也会影响不同选题方法的表现。在大部分实验条件下, Q 矩阵越复杂, 不同选题方法的 AACCR 和 PCCR 的增长幅度也越大。例如, 测验长度为 10 题的高质量题目条件下, 在复

表 2　简单结构下不同选题策略的判准率及题库使用情况

测验长度	题目质量	选题方法	AACCR	ImproveA	PCCR	ImproveP	T	NU	Chi
5	高	PWKL	0.835	—	0.380	—	0.021	390	12.21
		GIDPWKL	0.860	**0.025**	0.431	**0.051**	0.033	492	22.20
		AIDPWKL	0.854	0.019	0.429	0.049	0.031	488	20.92
		CIDPWKL	0.848	0.013	0.420	0.040	0.024	424	15.08
		KLEDPWKL	0.853	0.018	0.422	0.042	0.028	466	19.08
		MIM	0.842	0.007	0.410	0.030	0.021	398	12.79
	低	PWKL	0.769	—	0.315	—	0.021	380	12.44
		GIDPWKL	0.793	**0.024**	0.352	**0.037**	0.033	506	22.04
		AIDPWKL	0.774	0.005	0.347	0.032	0.032	501	21.28
		CIDPWKL	0.773	0.004	0.328	0.013	0.026	438	16.54
		KLEDPWKL	0.774	0.005	0.331	0.016	0.028	454	18.16
		MIM	0.772	0.004	0.334	0.019	0.030	482	20.92
10	高	PWKL	0.952	—	0.804	—	0.026	113	11.34
		GIDPWKL	0.969	**0.017**	0.837	**0.033**	0.087	244	19.65
		AIDPWKL	0.966	0.014	0.826	0.022	0.084	233	19.52
		CIDPWKL	0.962	0.010	0.823	0.019	0.029	156	14.17
		KLEDPWKL	0.960	0.008	0.818	0.014	0.026	135	13.40
		MIM	0.963	0.011	0.822	0.020	0.028	151	13.87
	低	PWKL	0.858	—	0.529	—	0.026	103	11.27
		GIDPWKL	0.879	**0.021**	0.561	**0.032**	0.046	263	24.88
		AIDPWKL	0.867	0.009	0.558	0.029	0.037	241	20.05
		CIDPWKL	0.864	0.006	0.553	0.024	0.031	172	15.40
		KLEDPWKL	0.866	0.007	0.556	0.027	0.035	216	17.95
		MIM	0.865	0.007	0.549	0.020	0.028	139	13.23

注：ImproveA 表示其余指标与 PWKL 指标的 AACCR 之差, ImproveP 表示其余指标与 PWKL 指标的 PCCR 之差, T 为测验重叠率, NU 为题库中未使用的题目数量, Chi 为卡方值

表 3　复杂结构下不同选题策略的判准率及题库使用情况

测验长度	题目质量	选题方法	AACCR	ImproveA	PCCR	ImproveP	T	NU	Chi
5	高	PWKL	0.812	—	0.341	—	0.029	338	19.22
		GIDPWKL	0.844	**0.032**	0.399	**0.058**	0.050	530	35.49
		AIDPWKL	0.827	0.015	0.385	0.044	0.045	494	31.56
		CIDPWKL	0.821	0.009	0.378	0.037	0.033	369	24.83
		KLEDPWKL	0.816	0.004	0.366	0.025	0.029	344	20.04
		MIM	0.826	0.014	0.374	0.033	0.032	358	23.97
	低	PWKL	0.721	—	0.235	—	0.032	330	21.49
		GIDPWKL	0.738	**0.017**	0.281	**0.046**	0.052	531	37.25
		AIDPWKL	0.735	0.014	0.270	0.035	0.046	493	32.90
		CIDPWKL	0.726	0.005	0.263	0.028	0.041	372	24.80
		KLEDPWKL	0.722	0.001	0.261	0.026	0.038	345	22.99
		MIM	0.729	0.008	0.269	0.034	0.045	475	31.12
10	高	PWKL	0.924	—	0.735	—	0.030	141	14.77
		GIDPWKL	0.945	**0.021**	0.791	**0.056**	0.051	294	31.59
		AIDPWKL	0.939	0.015	0.778	0.043	0.046	281	27.65
		CIDPWKL	0.931	0.007	0.769	0.034	0.042	157	19.33
		KLEDPWKL	0.937	0.013	0.776	0.041	0.044	246	24.50
		MIM	0.939	0.015	0.773	0.038	0.042	212	22.74
	低	PWKL	0.824	—	0.484	—	0.034	143	17.89
		GIDPWKL	0.843	**0.019**	0.523	**0.039**	0.053	308	33.52
		AIDPWKL	0.837	0.013	0.518	0.034	0.048	280	28.96
		CIDPWKL	0.829	0.005	0.505	0.021	0.037	169	20.55
		KLEDPWKL	0.834	0.010	0.501	0.017	0.035	154	18.28
		MIM	0.832	0.008	0.512	0.028	0.042	233	24.49

注：同表 2

杂结构中, GIDPWKL、AIDPWKL、CIDPWKL、KLEDPWKL 和 MIM 的 AACCR 值分别提高了 0.021, 0.015、0.007、0.013 和 0.015; PCCR 值分别提高了 0.056、0.043、0.034、0.041 和 0.038; 在简单结构中, GIDPWKL、AIDPWKL、CIDPWKL、KLEDPWKL 和 MIM 的 AACCR 值分别提高了 0.017, 0.014、0.010、0.008 和 0.011; PCCR 值分别提高了 0.033、0.022、0.019、0.014 和 0.020。

在题库使用情况上, 由于GIDPWKL和AIDPWKL方法的判准精度更高, 因此这两种方法的测验重叠率, 未使用的题目数量以及卡方值也是最大的, 其余 3 种方法虽然判准精度比 GIDPWKL 和 AIDPWKL 低, 但它们的题库使用情况要更好。该结果正是 CAT 形式测验中精度与题库使用情况的权衡(*trade-off*)问题的体现。由于本研究加入了曝光控制, 因此题库使用情况是可以控制在预期范围之内的。

5 模拟研究 2

5.1 研究目的

采用蒙特卡洛模拟方法, 在固定测验精度(Hsu, Wang, & Chen, 2013; Tatsuoka, 2002; 郭磊, 2014), 即变长终止规则条件下比较 6 种选题策略。重点考察不同选题策略下的测验使用情况, 主要包括平均测验长度 Mean, 测验长度的标准差 SD, 最大测验长度 Max 和最小测验长度 Min。其中, PWKL 法作为基线。所有程序采用 Matlab 2012b 进行编程。将测验的使用情况作为该研究的评价指标是因为：比较不同的选题策略质量差异时(控制其他条件均相同), 若使用定长终止规则, 那么判准率高的选题方法较好; 若使用变长终止规则, 即在固定终止精度时, 主要看平均用题量, 即平均用题量少的选题方法较好。因此, 在研究 2 中, 我们不再关注判准精度, 而是比较不同方法的测验使用情况。

5.2 研究设计

由于研究 2 采用变长终止规则, 一个比较简单可行的做法是通过改变被试 KS 后验概率分布中的最大值(记作 P_{1st})来控制终止精度(Tatsuoka, 2002)。本研究的终止精度包括 3 个水平：P_{1st} = 0.7, P_{1st} = 0.8 和 P_{1st} = 0.9, 其余条件同研究 1。

郭磊、郑蝉金和边玉芳(2015)提出了 3 种变长 CD-CAT 的项目曝光控制方法, 研究结果表明, 修正的 RT 法和修正的 RP 法在项目曝光率的控制上存在过度控制现象, 而 simple 法不存在该现象, 并且操作更加简洁, 因此, 本文选用 simple 法作为变长 CD-CAT 中的曝光控制方法。同时为了不让变长 CD-CAT 的题目过长, 与实际情况更加贴近, 本文将测验长度上限设置为 30 题(郭磊等, 2015)。simple 法是在选题指标前乘以曝光控制因子 f_j, 计算公式如下：

$$f_j = \frac{r_{\max} - m_j / N}{r_{\max}} \tag{21}$$

其中, $r_{\max}$ 为允许的最大项目曝光率(本研究设置为 0.2), m_j 为第 j 个项目当前的被调用次数, N 为参加测验的总人数。

5.3 评价指标

由于研究 2 和研究 1 的目的不同, 因此, 本研究的评价指标主要是测验的使用情况, 主要包括平均测验长度 Mean, 测验长度的标准差 SD, 最大测验长度 Max 和最小测验长度 Min。

5.4 研究结果

表 4 和表 5 是 6 种选题策略的测验使用情况。由结果可知, 与 PWKL 方法相比, 其余 5 种方法的平均测验长度更少, 其中表现最好的依然是 GIDPWKL 方法。

从表 4 结果可以看出, 除了按照最大测验长度终止以外, 大部分的实验条件下, 其余 5 种方法的最大测验长度要低于 PWKL 方法, 最小测验长度和 PWKL 相差无几。该结果表明其余 5 种方法较 PWKL 方法的优势所在：在具有相同测量精度时, 可以有效降低被试作答的最大测验长度。

与 AIDPWKL、CIDPWKL、KLEDPWKL 和 MIM 相比, GIDPWKL 的平均测验长度与 PWKL 的平均测验长度之差是最大的(除表 5 中最后一行以外), 节约的平均题目数量介于 0.47~0.87 之间, 如表 5 中粗体数值所示。该结果表明, 4 种新方法和 MIM 的选题效率更高, 在相同的测验情景中, 新方法能够用更少的题目达到与 PWKL 方法相同的测量精度。

值得注意的是, 不论采用何种方法, 随着终止精度 P_{1st} 的增大, 平均测验长度和最大测验长度均增大, 该结果和 Hsu 等(2013)的研究结果一致。Q 矩阵结构和题目质量均会影响这几种选题策略的测验使用情况。例如, 当固定 Q 矩阵结构时, 题目质量越高, 平均测验长度和最大测验长度越小; 当固定题目质量时, Q 矩阵结构越简单, 平均测验长度和最大测验长度越小。该结果表明, 在实际编制 Q 矩阵和题目时, 应注重提高题目的质量和适当减小 Q 矩阵的复杂性。

表 4　变长终止规则下测验长度的最大值和最小值

结构和质量	P_{1st}	PWKL		GIDPWKL		AIDPWKL		CIDPWKL		KLEDPWKL		MIM	
		Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min
S-H	0.7	18	5	18	6	20	5	17	4	18	5	18	6
	0.8	21	5	19	6	21	6	21	5	20	5	21	6
	0.9	28	5	24	7	25	7	25	6	24	6	25	7
S-L	0.7	30	7	30	8	30	8	30	7	30	7	30	8
	0.8	30	7	30	9	30	9	30	7	30	8	30	9
	0.9	30	10	30	10	30	10	30	9	30	9	30	11
C-H	0.7	27	3	23	5	25	4	27	4	24	4	26	5
	0.8	30	3	25	5	29	4	26	3	30	4	28	6
	0.9	30	5	27	6	28	5	30	3	30	4	30	6
C-L	0.7	30	4	30	6	30	5	30	5	30	4	30	6
	0.8	30	6	30	7	30	5	30	6	30	5	30	7
	0.9	30	7	30	8	30	7	30	5	30	7	30	8

注：S-H 表示简单结构和高题目质量组合, S-L 表示简单结构和低题目质量组合, C-H 表示复杂结构和高题目质量组合, C-L 表示复杂结构和低题目质量组合。

表 5　变长终止规则下测验长度的平均值和标准差

结构和质量	P_{1st}	PWKL		GIDPWKL		AIDPWKL		CIDPWKL		KLEDPWKL		MIM	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
S-H	0.7	8.77	1.92	7.90(**0.87**)	1.67	7.98(0.79)	1.69	8.09(0.68)	1.93	8.05(0.72)	1.89	8.11(0.66)	2.02
	0.8	10.31	2.26	9.55(**0.76**)	2.11	9.62(0.69)	2.16	9.64(0.67)	2.43	9.58(0.73)	2.17	9.77(0.54)	2.61
	0.9	12.29	2.80	11.66(**0.63**)	2.55	11.81(0.48)	2.68	11.89(0.40)	2.91	11.83(0.46)	2.69	11.85(0.44)	2.80
S-L	0.7	16.16	4.85	15.50(**0.66**)	4.35	15.73(0.43)	4.39	15.71(0.45)	4.81	15.73(0.43)	4.80	15.79(0.37)	4.92
	0.8	18.38	5.49	17.84(**0.54**)	5.10	18.01(0.37)	5.07	18.31(0.07)	5.33	18.17(0.21)	5.24	18.06(0.32)	5.23
	0.9	22.92	5.45	22.15(**0.77**)	5.34	22.35(0.57)	5.34	22.53(0.39)	5.30	22.34(0.58)	5.39	22.61(0.31)	5.40
C-H	0.7	8.92	2.65	8.45(**0.47**)	2.52	8.51(0.41)	2.56	8.74(0.18)	2.85	8.72(0.20)	2.73	8.66(0.26)	2.64
	0.8	10.47	3.32	9.79(**0.68**)	2.83	9.80(0.67)	3.08	9.81(0.66)	3.20	9.86(0.61)	3.25	9.83(0.64)	3.11
	0.9	12.33	3.59	11.77(**0.56**)	3.30	11.85(0.48)	3.52	12.06(0.27)	3.78	11.94(0.39)	3.60	11.96(0.37)	3.61
C-L	0.7	16.89	5.95	16.19(**0.70**)	5.97	16.53(0.36)	5.63	16.72(0.17)	5.89	16.67(0.22)	5.70	16.51(0.38)	5.73
	0.8	19.26	6.27	18.53(**0.73**)	5.96	18.67(0.59)	6.10	18.90(0.36)	6.26	18.74(0.52)	6.35	18.97(0.29)	6.45
	0.9	22.68	6.27	22.18(0.50)	6.01	22.13(0.55)	6.04	22.14(0.54)	6.31	22.34(0.34)	6.42	22.12(**0.56**)	6.34

注：S-H 表示简单结构和高题目质量组合, S-L 表示简单结构和低题目质量组合, C-H 表示复杂结构和高题目质量组合, C-L 表示复杂结构和低题目质量组合, 括号中数值表示 PWKL 指标下平均测验长度与新指标下平均测验长度之差。

6　研究结论

本文首先指出了传统的 PWKL 指标仅考虑了被试 KS 后验分布所提供的信息, 并未关注在选题过程中题目能够提供的项目层面的信息, 因此, PWKL 属于单源指标。随后, 本文将能够提供更加丰富信息的项目区分度融入到 PWKL 指标中, 对 PWKL 指标进行了修正, 提出了 4 种新的多源选题指标：GIDPWKL、AIDPWKL、CIDPWKL 和 KLEDPWKL 指标。另一方面, 根据 Wang (2013) 的研究结果表明：MIM 在大部分实验条件下的表现要优于 PWKL, 特别是在测验长度较短时。但 Wang本人并未考虑在曝光控制条件下 MIM 的表现, 目前也没有新方法与 MIM 之间的比较研究。因此, 本文通过两个模拟研究, 在控制项目曝光基础上, 系统比较了这 6 种方法在不同实验条件下的表现, 并得到以下结论：

(1)在定长测验情景下, 不论实验条件如何改变, 4 种新方法以及 MIM 方法的平均属性/模式判准率均要高于原始的 PWKL 方法。4 种新方法中表现最好的是 GIDPWKL, PCCR 最大增幅高达 5.8 个百分点(复杂结构×高质量题目×5 题), 这意味着在 1000 人参加的较短测验中, 比 PWKL 方法可以多判准 58 人;

(2)在定长测验情景下的绝大多数实验结果表明, 测验长度越短, 新方法的优势越明显。表现最好的是 GIDPWKL 方法, 之后是 AIDPWKL 方法, 而 CIDPWKL、KLEDPWKL 和 MIM 方法的优势随实验条件不同而不同。该结果表明, 新的选题策略在测验初期就会收到较大成效, 能够加快对被试 KS 判准的速度;

(3)在定长测验情景下的绝大多数实验结果表明, 题目质量越高, 新方法的优势越明显。表现最好的是 GIDPWKL 方法, 之后是 AIDPWKL 方法, 其余 3 种方法(CIDPWKL、KLEDPWKL 和 MIM)之间并没有展现出一致的优势结果, 但三者的表现相差无几。该结果表明, 项目区分度信息的确可以, 也应该作为另一方面的信息源加入到选题过程中, 以此提高被试 KS 的判准率;

(4) Q 矩阵结构的复杂性影响着不同选题策略的表现。从实验结果可以看出, 与简单结构相比, 复杂结构的 Q 矩阵更能体现出新方法的优势, 表明新方法更能有效处理复杂的测验情景;

(5)在变长测验情景下, 4 种新方法及 MIM 的平均测验长度要低于 PWKL 方法, 表现最好的是 GIDPWKL 方法。该结果表明新方法能够用更少的题目达到与 PWKL 方法相同的测量精度, 效率更高。

(6)整体来看, 4 种新方法以及 MIM 均比 PWKL 表现好。但相对而言, 在 4 种新方法中, CIDPWKL 和 KLEDPWKL 的表现不如 GIDPWKL 和 AIDPWKL。这是因为 CIDPWKL 和 KLEDPWKL 指标的项目区分度比较简单, 只考虑了项目参数的信息(即 s 和 g 参数), 而其余二者是基于 $D_{j,uv}$ 计算得到的项目区分度, 能提供的区分信息更加丰富。

本文提出的 4 种新方法通过将项目区分度作为权重融入 PWKL 指标中, 提高了选题效率。一个良好的选题方法的标准应该是在固定测验长度时, 具有较高的判准率; 或在固定测验精度时, 具有较少的测验长度, 而不是看该指标/方法应该有多复杂。根据实验结果表明, 本文提出的 4 种新方法在较短测验长度时, 比 PWKL 更加高效。根据上述结论, 多源指标是更加有效的选题策略。在定长测验中, GIDPWKL 方法的判准率是最高的; 在变长测验中, GIDPWKL 方法的平均测验长度是最少的, 因此, 在实际应用中应该首选测验效率最高的 GIDPWKL 方法。

7 讨论及展望

在选题过程中同时考虑项目提供的区分度信息(*multiple-source*), 能够加快对被试 KS 判准的速度, 这种优势尤其体现在测验初期, 其加快速度甚至超过 MIM (MIM 是当前在较短测验中表现最好的方法)。这是因为项目区分度属于项目水平(*item-level*)特征, 不会受到测验过程中 KS 估计值 $\hat{\alpha}$ 的影响, 能够弥补 PWKL 依赖 $\hat{\alpha}$ 的缺陷。由于新的选题指标属于多源指标, 从个体层面和题目层面对原有选题方法进行了重构, 能够比 PWKL 提供更多的信息, 提供的信息越多, 对被试 KS 的判准精度就会越高。而在 4 种新方法中, GIDPWKL 指标的综合表现最佳, 这是因为该指标不仅考虑了两两 KS 之间的 KL 信息量, 还将两两 KS 之间的汉明距离纳入考量, 包含了更多信息在内。同时, 根据结果可以看出, 题目质量越好, 融合了项目区分度信息的新指标的优势就会越明显, 这与理论上的分析结果是一致的。尤其是在变长测验中, 新方法能够用更少的题目达到与 PWKL 方法相同的测量精度, 这在实际应用中具有非常重要的意义和作用: 可以有效降低被试的作答负荷。

本文成功地将项目区分度信息融入到传统的 PWKL 指标中, 取得了令人满意的结果, 但仍有继续可以研究的地方:

(1)本研究仅选用了 DINA 模型作为认知诊断模型进行研究, 而融合模型(*Fusion Model*, FM)被认为是目前最优的诊断模型, 本文提出的 4 种新方法在 FM 中表现如何, 特别是 CIDPWKL 表现如何值得进一步研究。在 FM 中, 基于 CTT 思想的项目区分度指标不再是公式(4)所示, 而是下式:

$$d_j = \pi_j^* - \pi_j^* \prod_{k=1}^{K} r_{jk}^{*\,q_{jk}} \tag{22}$$

其中, π_j^* 为项目 j 的难度, r_{jk}^* 为项目 j 在第 k 个属性上的区分度。

(2)本研究并未考虑一些非统计约束条件, 例如内容平衡(Mao & Xin, 2013), 答案平衡和属性平衡(Cheng, 2010)等因素对新方法的影响, 未来可以进行这方面的研究。

(3)本研究是从项目区分度角度对 PWKL 进行的改进, 未来研究可以考虑其他加权方法。例如, 可以根据 Rupp 等(2010; P242)提出的计算属性标准误的方法, 将计算出来的属性标准误作为权重, 考察利用属性标准误进行加权方法的效果。

参 考 文 献

Chang, H. H., & Ying, Z. L. (1999). α-stratified multistage

computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 211–222.
Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4), 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*(6), 902–913.
Guo, L. (2014). *Variable-length cognitive diagnostic computerized adaptive testing: Termination rules, exposure control and quality monitoring technique* (Unpublished doctorial dissertation). Beijing Normal University.
[郭磊. (2014). *变长认知诊断计算机化自适应测验: 终止规则、曝光控制及题库质量监控技术*(博士学位论文). 北京师范大学.]
Guo, L., Zheng, C. J., & Bian, Y. F. (2015). Exposure control methods and termination rules in variable-length cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica, 47*(1), 129–140.
[郭磊, 郑蝉金, 边玉芳. (2015). 变长 CD-CAT 中的曝光控制与终止规则. *心理学报, 47*(1), 129–140.]
Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301–321.
Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*(4), 262–277.
Henson, R., Roussos, L., Douglas, J., & He, X. M. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement, 32*(4), 275–288.
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.
Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.
Mao, X. Z., & Xin, T. (2011). Improvement of item selection method in cognitive diagnostic computerized adaptive testing. *Journal of Beijing Normal University (Natural Science), 47*(3), 326–330.
[毛秀珍, 辛涛. (2011). 认知诊断 CAT 中选题策略的改进. *北京师范大学学报 (自然科学版), 47*(3), 326–330.]
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*(6), 482–496.
Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.
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.]
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.
Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305.
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 computerized adaptive testing. *Journal of Educational Measurement, 48*(3), 255–273.
Wang, W. Y., Ding, S. L., & Song, L. H. (2014). Item selection methods for balancing test efficiency with item bank usage efficiency in CD－CAT. *Journal of Psychological Science, 37*(1), 212–216.
[汪文义, 丁树良, 宋丽红. (2014). 兼顾测验效率和题库使用率的 CD－CAT 选题策略. *心理科学, 37*(1), 212–216.]
Xu, X. L., Chang, H. H., & Douglas, J. (2003). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.

New item selection methods in cognitive diagnostic computerized adaptive testing: Combining item discrimination indices

GUO Lei[1,2,3]; ZHENG Chanjin[4]; BIAN Yufang[5]; SONG Naiqing[3,6]; XIA Lingxiang[1]

([1] *Faculty of Psychology, Southwest University, Chongqing* 400715, *China*) ([2] *Postdoctoral Research Center for Statistics, Southwest University, Chongqing* 400715, *China*) ([3] *Southwest University Branch, Collaborative Innovation Center of Assessment toward Basic Education Quality, Chongqing* 400715, *China*) ([4] *School of Psychology, Jiangxi Normal University, Nanchang* 330022, *China*)

([5] *Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing* 100875, *China*)

([6] *Center for Basic Education Research, Southwest University, Chongqing* 400715, *China*)

Abstract

Interest in developing computerized adaptive testing (CAT) under cognitive diagnostic models has increased recently. Cognitive diagnostic computerized adaptive testing (CD-CAT) attempt to classify examinees into the correct latent class profile so as to pinpoint the strengths and weaknesses of each examinee whereas CAT algorithms choose items from the item bank to achieve that goal as efficiently as possible. Most of the

research in CD-CAT uses the posterior-weighted Kullback-Leibler (PWKL) index due to its high efficiency.

The PWKL index integrated the posterior probabilities of examinees' latent class profiles into the KL information, and thus improved item selection efficiency considerably. However, the PWKL index only used examinee-based information to assess the relative importance of each latent class profile. The current study attempted to take advantage of not only the examinee-base information but also the item-based information that could be readily obtained from items. In a sense, the PWKL index should be regarded as single-source index. This paper introduced four new multiple-source item selection methods, GIDPWKL, AIDPWKL, CIDPWKL, and KLEDPWKL respectively, which can be modified from the PWKL index by combining the item discrimination information. Two simulation studies were conducted to evaluate the new methods' efficiency against the PWKL index and mutual information (MI) index in the DINA model with the exposure control. The effects of different factors were investigated: the Q matrix structure (simple vs. complex), item quality (high vs. low) and test length (moderate vs. short).

Simulation results indicated that: (1) In most cases, the shorter the test length was, the higher AACCR and PCCR values the four new methods would have in the fix-length test. The GIDPWKL index had the highest average attribute correct classification rate and pattern correct classification rate among the six methods, and followed by AIDPWKL index. The performance among the CIDPWKL, KLEDPWKL, and MI depends on the experimental conditions. (2) In most cases, the higher the item quality was, the more advantage the four new methods would have in the fix-length test. (3) The structure of the Q matrix affected the performance of different item selection methods. (4) In the variable-length test, the mean of test length across all examinees for the four new methods and MI method were all smaller than those in the PWKL method. As a whole, the performance of the GIDPWKL index was the best, and should be recommended in practice where had the similar testing scenarios.

Key words cognitive diagnostic computerized adaptive testing; item selection strategy; item discrimination; exposure control