

# 分类视角下认知诊断测验项目区分度指标及应用\*

汪文义<sup>1</sup> 宋丽红<sup>\*\*2</sup> 丁树良<sup>1</sup>

(<sup>1</sup>江西师范大学计算机信息工程学院, 南昌, 330022) (<sup>2</sup>江西师范大学初等教育学院, 南昌, 330022)

**摘要** 在认知诊断中还没有指标能在无作答数据情况下直接评价项目的属性分类准确率或属性判断率。项目水平上的属性分类准确率, 与项目属性向量、项目参数、先验分布和作答反应等有关。综合各个影响因素定义了项目水平上的属性期望分类准确率指标, 并将其用于组卷。模拟研究显示: 新指标可十分准确地评价项目的属性判断率, 新指标对于项目筛选十分重要; 以模式分类准确率为评价指标, 基于新指标的组卷方法与经典的组卷方法表现相当。

**关键词** 分类准确率 项目属性期望分类准确率 组卷 确定性输入噪音与门模型

## 1 引言

如何评价和选择富含诊断信息的项目, 对构建具有较高的属性分类准确率或属性判断率的测验十分重要。要判断项目是否有利于诊断分类, 就必须开发项目质量评价指标。有研究发现项目所测的属性向量不同, 会影响测验分类准确率, 由此测验 Q 阵设计成为研究热点 (丁树良, 汪文义, 罗芬, 熊建华, 2016; 丁树良, 汪文义, 杨淑群, 2011; 丁树良, 杨淑群, 汪文义, 2010; Chiu, Douglas, & Li, 2009; Liu, Huggins-Manley, & Bradshaw, 2017)。特别是, 单位阵中测量单个属性的项目 (Chiu et al., 2009) 和可达阵中项目 (丁树良等, 2010) 对于诊断分类十分重要。丁树良等人提出了理论构念效度指标 (丁树良, 毛萌萌, 汪文义, 罗芬, Cui, 2012), 但没有考虑猜测和失误等因素。

在认知诊断中, 项目参数也是刻画项目质量和影响属性分类准确率的重要因素。众所周知, 项目区分度是经典测验理论和项目反应理论下重要的项目质量评价指标之一。为了刻画项目对属性模式或属性的区分度, 在认知诊断中也有一系列项目区分度指标 (item discrimination index; Rupp, Templin, & Henson, 2010): 认知诊断指标 (cognitive diagnosis index, CDI; Henson & Douglas, 2005)、属性区分度指标 (attribute discrimination index, ADI; Henson,

Roussos, Douglas, & He, 2008)、修改的 CDI 和 ADI 指标 (modified CDI/ADI, MCDI/MADI; Kuo, Pai, & de la Torre, 2016)、项目区分度指标 (郭磊, 郑蝉金, 边玉芳, 宋乃庆, 夏凌翔, 2016; Rupp et al., 2010)。

CDI 是基于相对熵信息量 (Kullback-Leibler information, KLI; Chang & Ying, 1996) 而得到, 其中 KLI 广泛应用于选题算法构建 (罗照盛等, 2015; 汪文义, 丁树良, 宋丽红, 2014; Chen, Xin, Wang, & Chang, 2012; Wang, 2013; Zheng & Chang, 2016)。ADI 可反映项目对特定属性的区分力, 尽管与分类准确率存在非线性关系, 但这种关系会随属性掌握概率变化而变化 (Henson et al., 2008)。确定性输入噪音与门 (DINA) 模型下 CDI 和选题算法 (郭磊等, 2016; 汪文义等, 2014; Wang, 2013) 中均含有项目区分度指标。另外, 项目质量还会随知识状态分布变化而变化, 如考虑后验分布的选题算法效果不错 (汪文义等, 2014; Cheng, 2009; Zheng & Chang, 2016)。

综上所述分析发现, 项目的属性分类准确率与项目属性向量、项目参数和知识状态分布有关。这种关系能否直接用数学公式进行描述, 能否应用于组卷及效果如何? 这是本研究重点要解决的问题。该问题的解决, 可让测验开发者在分类视角下获得项目质量的解释, 对于项目筛选具有重要价值, 将为预测测验的分类准确率、提出新选题算法、评价

\* 研究得到国家自然科学基金项目 (31500909, 31360237, 31160203)、全国教育科学规划教育部重点课题 (DHA150285)、江西省自然科学基金项目 (20161BAB212044)、江西省教育科学 2013 年度一般课题 (13YB032)、江西省社会科学规划项目 (17JY10)、国家社会科学基金项目 (16BYY096)、江西师范大学青年成长基金和江西师范大学博士启动基金的资助。

\*\* 通讯作者: 宋丽红。E-mail: viviansong1981@163.com

DOI:10.16719/j.cnki.1671-6981.20180234

诊断测验的优良性等奠定基础。

## 2 分类视角下项目区分度指标

### 2.1 确定性输入噪声与门模型

DINA 模型的项目反应函数为:

$$f(X_j = 1 | \alpha_i) = g_j^{1-\eta_{ij}} (1-s_j)^{\eta_{ij}}. \quad (1)$$

$$f(X_j = 1 | \alpha_i) \text{ 和 } f(X_j = 0 | \alpha_i) = 1 - f(X_j = 1 | \alpha_i)$$

表示知识状态为  $\alpha_i$  的被试  $i$  在项目  $j$  上的正确和错误作答概率; 而  $\eta_{ij} = \prod_{k=1}^K \alpha_{ik}^{q_{kj}}$  表示知识状态为  $\alpha_i$  的被试在项目  $j$  上的理想反应;  $s_j$ ,  $g_j$  和  $q_j$  为项目  $j$  的失误参数、猜测参数和所考查的属性向量。并记  $K_j = \sum_{k=1}^K q_{kj}$  表示项目  $j$  所考查的属性数量,  $K$  为属性数量,  $Q_S$  表示知识状态全集且  $\alpha_i \in Q_S$ 。

### 2.2 项目属性期望分类准确率

引言中提到, 项目的属性分类准确率与项目属性向量、项目参数和知识状态分布等有关。如果项目没有考查属性  $k$ , 即  $q_{jk}=0$ , 该项目上项目参数和作答反应并不会影响属性  $k$  的分类准确率。如果采用极大似然方法估计知识状态, 属性  $k$  上的分类正确或错误与否, 完全仅凭机会而定, 即正确分类概率为 .5。这也可从 ADI 的定义得到解释, 项目  $j$  上属性的 ADI 为 0(Henson et al., 2008)。为便于理解, 下面通过三个例子引入分类视角下项目区分度指标之项目属性期望分类准确率, 记为 EAMR, 再通过定理 1 和定理 2 给出一般性结论。

例 1 若  $K_j=1$ ,  $q_{jk}=1$ ,  $s_j$  和  $g_j$  均为 .25, 在被试不失误或猜测情况下, 项目  $j$  对考查的属性  $k$  分类正确, 此时项目的  $EAMR_{jk}=.75$ 。

例 2 若  $K=K_j=2$ ,  $q_j=[1,1]$ ,  $s_j$  和  $g_j$  均为 .25, 可知  $EAMR_{jk}=.625$ 。这是因为, 给定知识状态全集  $Q_S=\{[1,1], [1,0], [0,1], [0,0]\}$  和各种知识状态的概率均为 .25: 若被试的知识状态为  $[1,1]$  且在项目  $j$  上不失误, 则属性 1 上分类正确, 对应概率为  $p_{11}=.25(1-.25)$ ; 若被试的知识状态为  $[1,0]$  且猜测正确, 则属性 1 上分类正确, 对应概率为  $p_{12}=.25 \times .25$ ; 若被试的知识状态为  $[0,0]$  或  $[0,1]$  且被试不猜测, 则属性 1 上分类正确, 对应概率为  $p_{13}=.5(1-.25)$ ; 三个正确分类概率相加即为 .625, 其他情形均会分类错误。

例 3 若  $K=K_j=3$ ,  $q_j=[1,1,1]$ ,  $s_j$  和  $g_j$  均为 .25, 可知  $EAMR_{jk}=.5625$ 。类似可计算得到  $p_{11}=.125(1-.25)$ ,  $p_{12}=.375 \times .25$  和  $p_{13}=.5(1-.25)$ , 三者之和 .5625 与  $EAMR_{jk}$  相等。

定理 1 在属性相互独立和知识状态分布为均

匀分布条件下, 对于 DINA 模型, 若  $(1-s_j) > g_j$  且  $q_{jk}=1$ , 如果采用极大似然估计方法估计属性  $k$  的状态, 项目  $j$  对属性  $k$  的期望分类准确率为:

$$EAMR_{jk} = (1-s_j - g_j) / 2^{K_j} + 0.5. \quad (2)$$

易知 DINA 模型下项目对考查的各个属性的期望分类准确率相同。从图 1 可以看出, 猜测与失误越小且项目所测属性数量越小, EAMR 越大。这可以很好地解释单位阵中考查单个属性的项目为何重要。注意到对于其他非独立型的属性层级结构, 在 Q 矩阵中, 可达阵的列相对于扩张出来的列包含非零元是最少的, 这个定理也解释了为什么说可达阵在认知诊断测验编制中十分重要的原因。这个定理可能可以解决除可达阵之外, 其他项目应该如何选取的问题。为了保证阅读流畅性, 在此略去证明 (其实定理 1 是下面定理 2 的推论)。

定理 2 如果项目  $j$  采用 0-1 记分规则 ( $X_j=0$  或  $X_j=1$ ), 某种分类方法下属性  $k$  的分类结果分别为  $\hat{\alpha}_{k0}$  和  $\hat{\alpha}_{k1}$ , 在考虑先验分布信息下, 项目  $j$  对属性  $k$  的期望分类准确率为:

$$EAMR_{jk} = P(\alpha_k = \hat{\alpha}_{k1} | X_j=1) P(X_j=1) + P(\alpha_k = \hat{\alpha}_{k0} | X_j=0) P(X_j=0). \quad (3)$$

定理 2 可以从属性水平的分类准确率指标 (Wang, Song, Chen, Meng, & Ding, 2015) 得到很好的解释和证明。在给定一个项目下被试观察得分  $X_j$ , 根据认知诊断测验中属性  $k$  的分类准确率 (Wang et al., 2015), 即属性估计状态的后验分布  $\hat{p}_{ik} = P(\hat{\alpha}_k | X_i)$ , 知  $EAMR_{jk}$  公式中两项  $P(\alpha_k = \hat{\alpha}_{k1} | X_j=1)$  和  $P(\alpha_k = \hat{\alpha}_{k0} | X_j=0)$  分别是给定得分  $X_j=0$  和  $X_j=1$  时属性的分类准确率。再考虑项目反应的随机性, 对项目  $j$  上得分取期望, 正好是项目  $j$  对属性  $k$  的分类准确率。由于该指标是属性  $k$  的分类准确率的期望, 故命名为项目属性期望分类准确率。

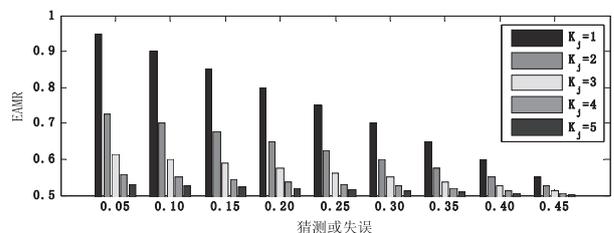


图 1 DINA 模型下不同猜测与失误和属性数量下 EAMR

## 3 组卷方法

组卷为 NP 难问题 (Chen, 2016), 类似于旅行商问题, 这类问题的大型实例迄今为止不能用精确

算法求解，在有限计算时间条件必须放弃寻求全局最优解，而寻求这类问题的近似算法或启发式算法 (heuristics; Givens & Hoeting, 2013)。为解决组合优化问题，引言提到的 CDI, ADI 和 MCDI/MADI 组卷方法，均是采用贪心算法，依次从题库中选择指标最大的项目，最终将这些项目组成试卷。

### 3.1 基于 CDI 的组卷方法

在项目水平上，KLI 可用于评价任意两种知识状态  $\alpha_u$  和  $\alpha_v$  下项目反应概率分布  $f(X_j | \alpha_u)$  和  $f(X_j | \alpha_v)$  之间的距离  $D_{juv}$ ：

$$D_{juv} = KLI_j(f(X_j | \alpha_u), f(X_j | \alpha_v)) = \sum_{x=0}^1 f(X_j = x | \alpha_u) \ln \left[ \frac{f(X_j = x | \alpha_u)}{f(X_j = x | \alpha_v)} \right]. \quad (4)$$

两个概率分布完全相同时，KLI 为 0，即 KLI 下限；两个概率分布差异越大，KLI 越大。

在实际应用组卷中，需要重点考虑测验对相近知识状态的区分力。基于这样的考虑，有研究者提出了 CDI(Henson & Douglas, 2005)：

$$CDI_j = \frac{1}{\sum_{u \neq v} h(\alpha_u, \alpha_v)^{-1}} \sum_{u \neq v} h(\alpha_u, \alpha_v)^{-1} D_{juv}. \quad (5)$$

其中相似性采用欧氏距离  $h(\alpha_u, \alpha_v) = \sum_{k=1}^K (\alpha_{uk} - \alpha_{vk})^2$  进行度量。

### 3.2 基于 EAMR 的组卷方法

先约定以下记号， $B = \{1, 2, \dots, M\}$  表示用于某次组卷时题库中所有项目的集合； $I$  表示已经从  $B$  中选择的  $m$  个项目的集合，组卷开始时  $I = \emptyset$ 。新组卷方法以 EAMR 之和为准则，采用贪心算法选择满足下列条件的项目进入试卷：

$$j = \arg \max_{j \in B-I} \sum_{k=1}^K EAMR_{jk} \quad (6)$$

从 EAMR 计算公式或图 1 来看，猜测和失误参数小且考查的属性数量少的项目更易被贪心算法所选择。

## 4 模拟研究

### 4.1 研究 1 EAMR 指标的表现

#### 4.1.1 研究目的

由于被试知识状态具有潜在不可观察性，只能通过模拟研究，评价 2.2 节提出的 EAMR 指标能否准确估计单个项目的属性分类准确率。

#### 4.1.2 研究设计

固定属性数量为  $K=5$ 。根据定理 1 和定理 2，

考虑的影响因素包括：项目参数、项目所考查的属性数量、被试知识状态分布。模拟  $M=1000$  个项目。根据已有研究参数设置情况 (Chen et al., 2012; Henson & Douglas, 2005)，DINA 模型的项目参数水平分为三个水平，即分别服从均匀分布： $U(.05, .25)$ 、 $U(.25, .45)$  和  $U(.05, .40)$ 。为了反映不同属性数量情况下指标的返真性，并让各种属性向量的项目较均匀出现，各个项目所考查的属性向量从所有可能的属性向量集中随机抽取；为了准确得到模拟的属性判断率，被试数  $N$  设为 10,000 (Henson & Douglas, 2005)。

潜在能力向量服从多维正态分布，即  $\theta \sim MVN(\mathbf{0}, \Sigma)$ 。 $\Sigma$  为相关矩阵，主对角线和非对角线元素分别为 1 和  $\rho$ ，其中  $\rho$  设置了四个水平：0, .5, .75 和 .95 (Henson & Douglas, 2005; Henson et al., 2008)。当  $\rho=0$  时，潜在能力划界值设为 0，以得到满足定理 1 条件的均匀分布的知识状态；当  $\rho$  为 .5, .75 和 .95 时，采用下式模拟被试属性状态 (Chiu et al., 2009)：

$$\alpha_{ik} = \begin{cases} 1 & \text{如果 } \theta_{ik} \geq \phi^{-1}\left(\frac{k}{K+1}\right) \\ 0 & \text{其他.} \end{cases} \quad (7)$$

其中  $\phi^{-1}$  为标准正态分布的分布函数  $\phi$  的逆函数。

#### 4.1.3 评价指标

采用期望后验概率 (expected a posteriori, EAP) 估计方法得到被试  $i$  的知识状态 ( $\hat{\alpha}_i$ )，再根据模拟的知识状态  $\alpha_i$ ，可计算模拟的属性判断率 (AMR)：

$$AMR_k = \frac{1}{N} \sum_{i=1}^N \mathbf{I}(\hat{\alpha}_{ik} = \alpha_{ik}). \quad (8)$$

其中  $\mathbf{I}$  为示性函数。EAP 估计是基于被试  $i$  在单个项目上模拟作答反应和先验分布得到，并采用 .5 进行划界。知识状态先验分布是按照 Henson 和 Douglas (2005) 的方法，通过统计各分布下模拟的 10,000 名被试各知识状态的频率而得到的；EAMR 也是直接根据经验先验分布计算。

采用偏差 (BIAS)、绝对偏差 (ABS) 和均方误差 (root mean square errors, RMSE) (张淑梅, 辛涛, 曾莉, 孙佳楠, 2011) 评价 AMR 与 EAMR 之间的误差：

$$BIAS = \frac{\sum_{k=1}^K \sum_{j=1}^M (EAMR_{jk} - AMR_{jk})}{KM}, \quad (9)$$

$$ABS = \frac{\sum_{k=1}^K \sum_{j=1}^M |EAMR_{jk} - AMR_{jk}|}{KM}, \quad (10)$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^K \sum_{j=1}^M (EAMR_{jk} - AMR_{jk})^2}{KM}} \quad (11)$$

#### 4.1.4 实验结果

表 1 给出了三种项目参数分布、两种被试分布下 (其他结果类似, 故略) EAMR 的估计精度。

表 1 EAMR 的估计精度

| 相关    | 属性数量 | 题数  | s,g~U(.05,.25) |       |       | s,g~U(.25,.45) |       |       | s,g~U(.05,.40) |       |       |
|-------|------|-----|----------------|-------|-------|----------------|-------|-------|----------------|-------|-------|
|       |      |     | BIAS           | ABS   | RMSE  | BIAS           | ABS   | RMSE  | BIAS           | ABS   | RMSE  |
| ρ=0   | 1    | 148 | -.0004         | .0034 | .0042 | -.0021         | .0065 | .0076 | -.0010         | .0046 | .0054 |
|       | 2    | 333 | -.0001         | .0038 | .0046 | -.0018         | .0058 | .0070 | -.0005         | .0047 | .0055 |
|       | 3    | 324 | .0000          | .0036 | .0046 | -.0012         | .0053 | .0065 | -.0009         | .0046 | .0055 |
|       | 4    | 168 | -.0006         | .0037 | .0046 | -.0011         | .0052 | .0063 | -.0014         | .0045 | .0054 |
|       | 5    | 27  | -.0010         | .0040 | .0049 | -.0005         | .0045 | .0057 | -.0015         | .0051 | .0061 |
| ρ=.75 | 1    | 148 | -.0009         | .0036 | .0046 | -.0008         | .0064 | .0075 | .0000          | .0028 | .0038 |
|       | 2    | 333 | -.0016         | .0039 | .0049 | -.0007         | .0062 | .0073 | .0004          | .0027 | .0036 |
|       | 3    | 324 | -.0021         | .0041 | .0051 | -.0007         | .0063 | .0074 | .0007          | .0027 | .0037 |
|       | 4    | 168 | -.0022         | .0041 | .0051 | -.0011         | .0063 | .0075 | .0005          | .0027 | .0036 |
|       | 5    | 27  | -.0017         | .0044 | .0053 | -.0014         | .0063 | .0075 | .0003          | .0028 | .0038 |

EAMR 指标估计误差基本上没有差异; (3) 不同项目参数下 EAMR 指标估计精度也没有呈现规律性的差异。

## 4.2 研究 2 EAMR 在组卷中的应用

### 4.2.1 研究目的

研究 2 主要开展 EAMR 指标在组卷中的应用, 并与经典的组卷方法进行比较。

### 4.2.2 研究设计

属性数量、项目参数、知识状态分布与研究一的设置相同。模拟含 300 个项目的题库。题库 Q 阵下各个项目以概率 .3 独立考查各个属性 (Cui, Gierl, & Chang, 2012), 题库中所有项目考查属性数量的均值接近理论期望 1.8 (由于项目至少考查一个属性, 形成了二项分布的截尾分布, 该截尾分布的期望比二项分布的期望 1.5 高)。为了清晰显示测验分类准确率与测验长度之间的关系, 测验长度水平不按单个项目的增减而设置, 考虑了 10 个水平的测验长度, 分别为 5, 10, ..., 50。

组卷算法主要考虑随机组卷、CDI 组卷 (Henson & Douglas, 2005)、基于可达阵的组卷 (丁树良等, 2010, 2011) 和基于 EAMR 的组卷方法, 分别记为 RD, CDI, R 和 EAMR。基于可达阵的组卷算法, 从 300 个项目中有放回式取一个或多个单位阵 (独立结构下的可达阵) 组成测验 Q 阵, 该 Q 阵中均是测

总体上来看, 所有条件下 BIAS、ABS 和 RMSE 都十分接近于 0, 说明 EAMR 可以准确地估计其 AMR。从各个因素对 EAMR 指标估计精度的影响来看: (1) 因为 EAMR 指标估计中考虑了真实分布的经验分布, 相关或被试分布并没有对估计误差造成明显影响; (2) 对于考查不同属性数量的项目而言,

量单个属性的项目。CDI 组卷算法也倾向于选择考查属性数较少的项目 (Kuo et al., 2016)。

### 4.2.3 评价指标

类似于 AMR, 可计算测验的模式判准率 (PMR):

$$PMR = \frac{1}{N} \sum_{i=1}^N \mathbf{I}(\hat{\alpha}_i = \alpha_i) \quad (10)$$

向量  $\hat{\alpha}_i$  和  $\alpha_i$  对应分量全部相等时, 示性函数  $\mathbf{I}$  为 1, 否则为 0。

### 4.2.4 实验结果

图 2 至图 4 分别给出了三种项目参数分布和四种被试分布下各种组卷方法所得测验的 PMR。结果显示: (1) 基于 EAMR 的组卷方法和 CDI 组卷方法表现基本相当, 说明 EAMR 可以较好地应用于组卷。特别是测验长度较长时, 这两种组卷方法表现十分接近。在测验较短时, 基于 EAMR 的组卷方法甚至比 CDI 组卷方法的 PMR 要高; (2) 在项目参数分布为 U(.05,.25) 时, 基于可达阵的组卷方法表现不错。而在项目参数分布为 U(.25,.45) 和 U(.05,.40) 时, 基于可达阵的组卷方法表现逊于基于 CDI 和基于 EAMR 组卷方法; (3) 随机组卷方法表现最差。

为什么在某些条件下基于可达阵的组卷方法竟然与随机组卷方法表现相当呢? 这主要是由以下两方面原因共同决定: (1) 可达阵设计是一种理想状态

的设计, 可达阵组卷并没有考虑猜测与失误参数的影响, 因此当所得测验的项目猜测和失误均值较大时 (见表 2), 噪音比较大准确率自然不可能太高。另外, 可达阵组卷时具有一定的随机性, 在相同项目参数分布和不同相关水平下, 模式准确率本身也

会存在一定的波动; (2) 实验中考虑的是独立结构, 而独立结构层级关系松散, 可达阵对独立型结构表现不是很好。在独立结构下, 含单个属性的可达阵列项目的 EAMR, 并不一定比某些非可达阵列项目高。如猜测和失误参数稍小而考查属性数量较多 (如

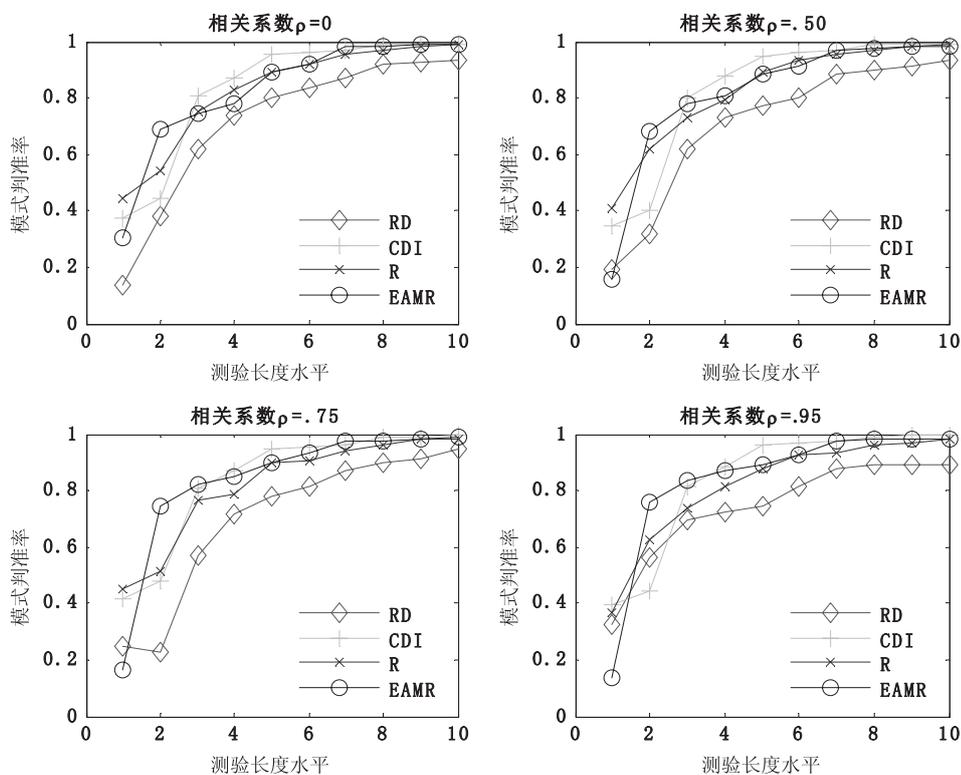


图 2 U(.05, .25) 下四种组卷和相关下测验的模式准确率

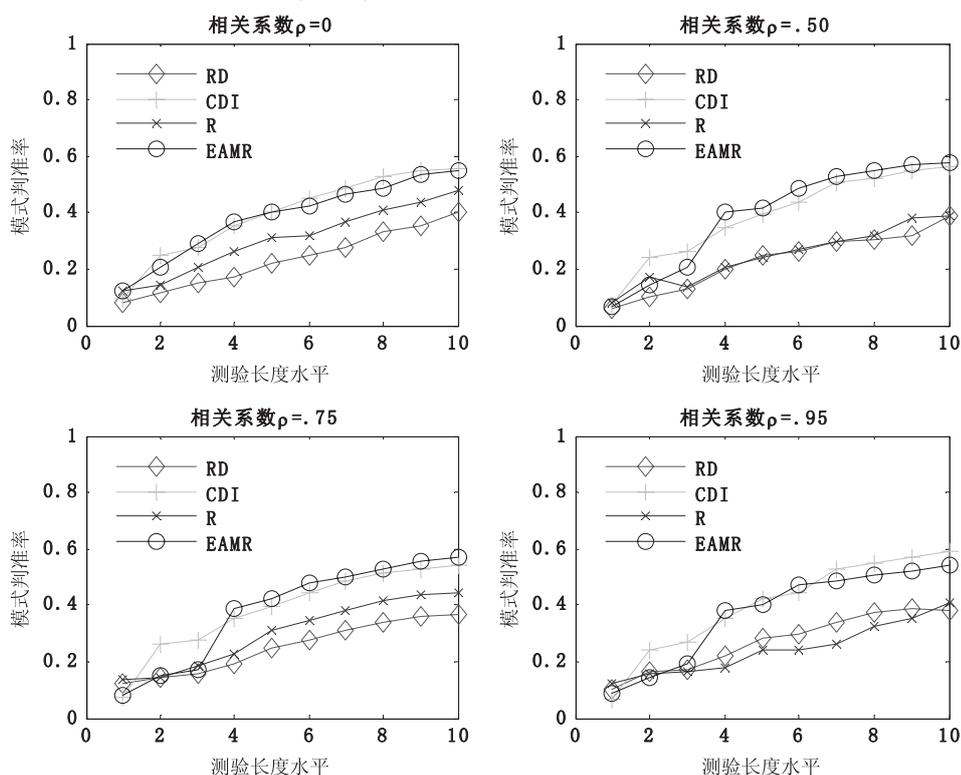


图 3 U(.25, .45) 下四种组卷和相关下测验的模式准确率

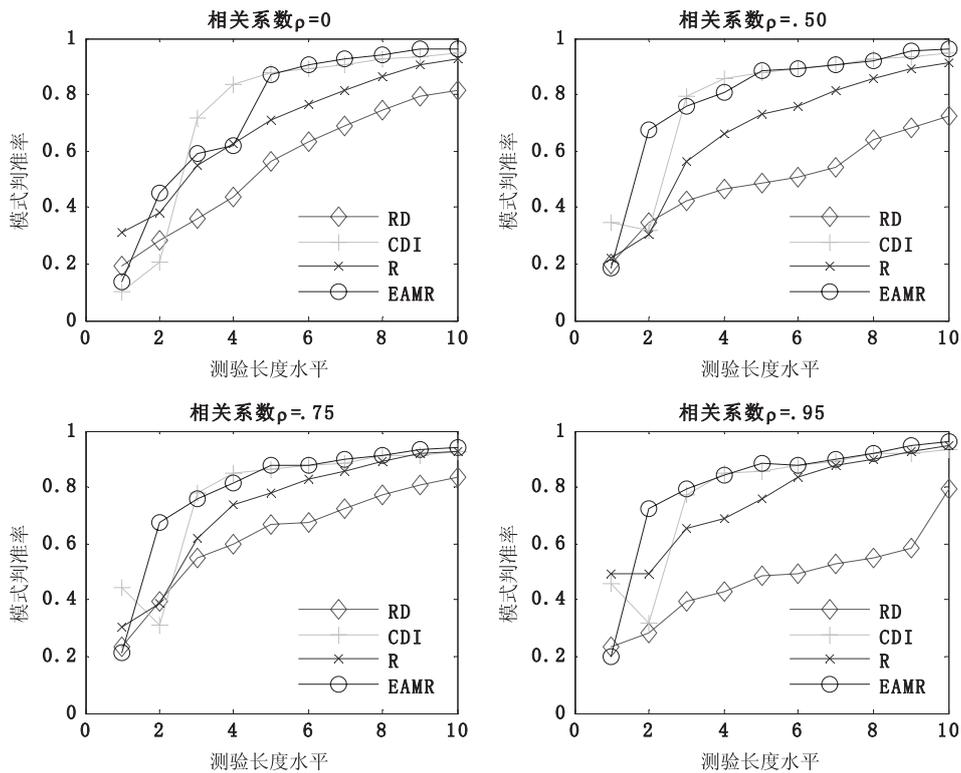


图 4 U(.05,.40) 下四种组卷和相关下测验的模式判准率

表 2 各组卷方法对应测验的各属性数量的项目数和参数统计结果

| 项目参数       | 组卷方法 | 测量各属性数量的项目数 |      |      |     |    | 测验项目参数均值 |       |
|------------|------|-------------|------|------|-----|----|----------|-------|
|            |      | 1           | 2    | 3    | 4   | 5  | 失误       | 猜测    |
| U(.05,.25) | RD   | 24.0        | 15.8 | 9.0  | .8  | .5 | .1471    | .1454 |
|            | CDI  | 44.0        | 6.0  | .0   | .0  | .0 | .0949    | .0941 |
|            | R    | 50.0        | .0   | .0   | .0  | .0 | .1416    | .1452 |
|            | EAMR | 33.0        | 17.0 | .0   | .0  | .0 | .0919    | .0898 |
| U(.25,.45) | RD   | 23.3        | 15.3 | 10.0 | 1.3 | .3 | .3462    | .3465 |
|            | CDI  | 34.0        | 16.0 | .0   | .0  | .0 | .2916    | .2925 |
|            | R    | 50.0        | .0   | .0   | .0  | .0 | .3490    | .3611 |
|            | EAMR | 28.0        | 22.0 | .0   | .0  | .0 | .2888    | .2933 |
| U(.05,.40) | RD   | 19.3        | 21.8 | 8.0  | 1.0 | .0 | .2350    | .2314 |
|            | CDI  | 36.0        | 13.0 | 1.0  | .0  | .0 | .1307    | .1418 |
|            | R    | 50.0        | .0   | .0   | .0  | .0 | .1974    | .2304 |
|            | EAMR | 32.0        | 18.0 | .0   | .0  | .0 | .1281    | .1447 |

2 或 3) 的项目的 EMAR, 甚至高于猜测和失误参数稍大而考查属性数量为 1 项目的 EMAR (见图 1)。另外, 相对于 U(.05,.25) 分布, U(.25,.45) 分布下属性数量对项目的 EMAR 的影响程度急剧下降 (见图 1)。

猜测和失误小且考查的属性数量少的项目更易被 EAMR 和 CDI 组卷方法所选择, 这一结论得到了实验证实。从测验不同属性数量的项目分布来看 (见表 2), R、CDI、EAMR 和随机组卷方法, 分别选择了约 50、38、31 和 22.2 个单个属性的项目, 以及分别选择了约 0、11.7、19 和 17.6 个 2 个

属性的项目, 除随机组卷方法选择了 9 个含 3 个属性的项目外, 其他属性数量的项目选择非常少。从项目参数来看, CDI 和 EAMR 组卷方法均选择了猜测和失误较小的项目。如在 U(.05,.25) 题库, CDI 和 EAMR 组卷方法的测验项目的参数均值在 .09 左右; 而 R 和 RD 所得测验的参数均值与题库均值基本相等。

下面分析 EAMR 和 CDI 组卷方法的表现存在上述差异的原因。图 5 给出了不同项目参数水平下项目的 EAMR 之和, 及其基于相对熵的 D 值平均值, 其中基于相对熵的 D 值平均值可视为等权重的

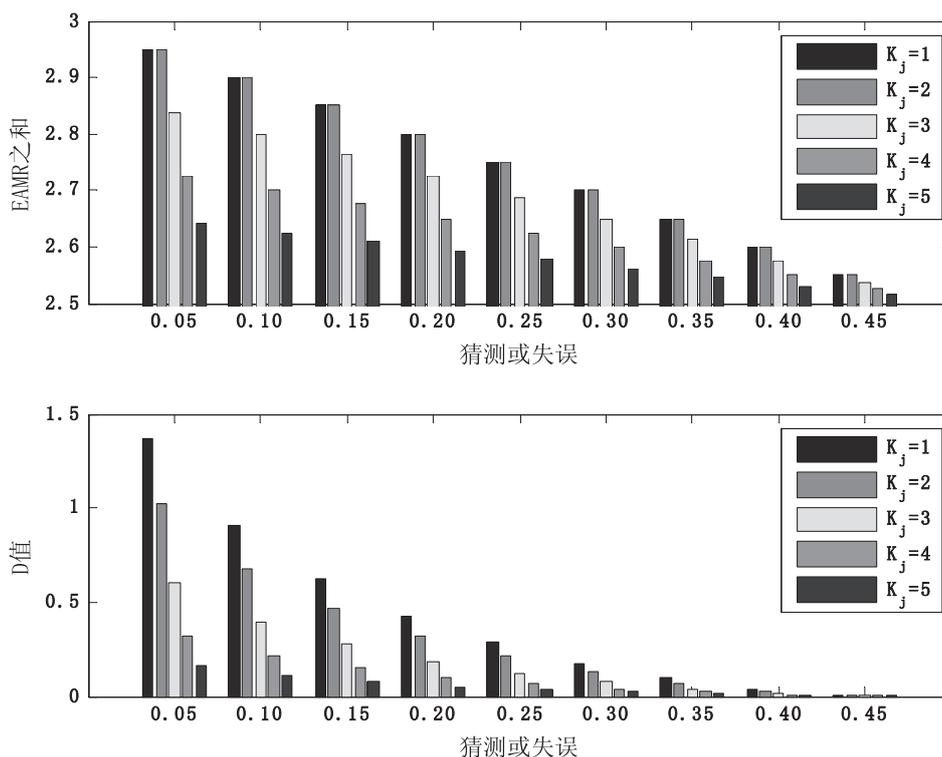


图5 项目 EAMR 之和与 D 值变化趋势比较

CDI (Henson & Douglas, 2005)。从图 5 可以看出，在相同项目参数情况下，含 1 或 2 个属性数的项目的 EAMR 之和相等，而基于相对熵的 D 值平均随项目所考查的属性数增加而严格递减，这可以很好地解释为什么 EAMR 较集中选择含 1 或 2 个属性数的项目，而 CDI 集中选择含一个属性数的项目。

## 5 结论与讨论

综合考虑影响项目分类效果的因素，从理论上导出了认知诊断测验的 EAMR 指标，同时给出了 DINA 模型下 EAMR 计算公式，并基于 EAMR 构建了新的组卷方法。模拟研究结果显示：EAMR 指标可以十分准确地估计项目的模拟属性判准率，说明该指标可以作为 DINA 模型下项目的分类准确率的度量指标；基于 EAMR 之和的组卷方法，与基于 CDI 组卷方法表现相当，并且 EAMR 组卷方法可以提高题库使用率，说明 EAMR 可以很好地应用于认知诊断测验组卷。

认知诊断测验的基本目标是根据被试作答反应准确地对被试知识状态进行分类，才能有效地根据分类结果对被试进行个性化的补救教学，提高学习效率和增强学习动机。这就要求诊断项目的区分度指标能与诊断测验的分类目标十分契合。从分类视角下，本文创新地提出了衡量认知诊断项目区分度

的指标。该指标不同于认知诊断下源于经典测验理论和项目反应理论下传统的项目区分度指标（CDI 和 ADI 指标），新指标可以看作是认知诊断测验特有的第三类项目区分度指标。

基于 EAMR 组卷方法根据项目对属性的分类准确率进行选题，相比 CDI 组卷方法，目标更为明确而直接。由于属性数量权重差异，CDI 组卷方法比 EAMR 组卷方法，更侧重于属性数量较少的项目。EAMR 组卷方法倾向于选择含 1 或 2 个属性数的项目。一般认为含单个属性的项目难度相对较小，而 2 个或多个属性的项目难度相对较大，选择 2 个或多个属性的项目有利于区分掌握属性较好的被试知识状态。为简单起见，模拟研究并没有考虑各种组卷约束，如各个属性分类准确率的平衡、测验长度变化等，这在实际应用中有待考虑。

根据本文的研究结论，下面针对几种实际中可能遇到的其他情况，提出几点建议供读者参考：(1) 对于测验编制者或测验开发者而言，可适当多编写含属性数较少的题目，并且尽量少采用猜测和失误较大的题型，如判断题和选项数较少的选择题；(2) 对于题库中各个属性的平均期望分类准确率不太均衡情况，在组卷时需要考虑不同属性分类准确率的平衡问题，或在组卷前补充相关试题，以提高测验的整体分类准确率；(3) 对于组卷中曝光过度或不太

使用的题目,可通过分库、分层、休眠、增加新题、控制曝光、题目另作他用(练习题)等方法常态化地维护题库;(4)新指标可供测验开发者作为项目质量评价指标使用,用于评价项目分类质量。

尽管模拟研究中只给出 DINA 模型不同条件下 EAMR 指标的表现, EAMR 指标也可用于其他认知诊断模型下项目质量评价。如常见的确定性输入噪音或门模型等,基于先验分布、项目参数和项目反应函数,就可以直接使用通式公式(3)计算项目 EAMR 指标。在具有诊断功能的计算机化自适应测验中,结合知识状态先验分布的通式公式(3),将来可用于开发新的选题算法。如何使用项目 EAMR 预测认知诊断测验的分类准确率,值得进一步研究。

综上所述,本研究的贡献或潜在贡献主要在于:综合了各种影响因素,从理论上定义了诊断项目的属性期望分类准确率指标,可以直接反映认知诊断项目对于各个属性的分类效果,这与认知诊断分类目标十分契合;该指标将作为重要的项目质量评价指标,结合传统的项目区分度指标,可供测验开发者评价和选择项目,更好地控制测验的信度和效度;基于 EAMR 的组卷方法,可以较好地提高题库项目使用率;EAMR 应该适合构建新的选题算法,但仍有待研究。

### 参考文献

- 丁树良,毛萌萌,汪文义,罗芬, Cui. (2012). 教育认知诊断测验与认知模型一致性的评估. *心理学报*, *44*(11), 1535-1546.
- 丁树良,汪文义,罗芬,熊建华. (2016). 可达阵功能的不可替代性. *江西师范大学学报(自然科学版)*, *40*(3), 290-294, 298.
- 丁树良,汪文义,杨淑群. (2011). 认知诊断测验蓝图的设计. *心理科学*, *34*(2), 258-265.
- 丁树良,杨淑群,汪文义. (2010). 可达矩阵在认知诊断测验编制中的重要作用. *江西师范大学学报(自然科学版)*, *34*(5), 490-494.
- 郭磊,郑蝉金,边玉芳,宋乃庆,夏凌翔. (2016). 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. *心理学报*, *48*(7), 903-914.
- 罗照盛,喻晓锋,高椿雷,李喻骏,彭亚凤,王睿,王钰彤. (2015). 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. *心理学报*, *47*(5), 679-688.
- 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.
- 张淑梅,辛涛,曾莉,孙佳楠. (2011). 2PL 模型的 EM 缺失数据处理方法研究. *应用概率统计*, *27*(3), 241-255.
- Chang, H. H., & Ying, Z. L. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*(3), 213-229.
- Chen, P. H. (2016). Three-element item selection procedures for multiple forms assembly: An item matching approach. *Applied Psychological Measurement*, *40*(2), 114-127.
- 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*(2), 201-222.
- Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*(4), 619-632.
- Chiu, C. Y., Douglas, J. A., & Li, X. D. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*(4), 633-665.
- Cui, Y., Gierl, M. J., & Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*(1), 19-38.
- Givens, G. H., & Hoeting, J. A. (2013). *Computational statistics*. Hoboken, NJ: John Wiley & Sons, Inc.
- 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.
- Kuo, B. C., Pai, H. S., & de la Torre, T. (2016). Modified cognitive diagnostic index and modified attribute-level discrimination index for test construction. *Applied Psychological Measurement*, *40*(5), 315-330.
- Liu, R., Huggins-Manley, A. C., & Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement*, *77*(2), 220-240.
- Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: The Guilford Press.
- 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, W. Y., Song, L. H., Chen, P., Meng, Y. R., & Ding, S. L. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, *52*(4), 457-476.
- Zheng, C. J., & Chang, H. H. (2016). High-efficiency response distribution-based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *40*(8), 608-624.

# An Item Discrimination Index and Its Application in Cognitive Diagnostic Assessment on a Classification-Oriented View

Wang Wenyi<sup>1</sup>, Song Lihong<sup>2</sup>, Ding Shuliang<sup>1</sup>

(<sup>1</sup>School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, 330022)

(<sup>2</sup>Elementary Educational College, Jiangxi Normal University, Nanchang, 330022)

**Abstract** The existing studies suggested that item quality is closely relevant to the number of attributes required by an item, item parameters, and the prior distribution of attribute patterns in cognitive diagnostic assessment. Several studies focused on the design of Q-matrix and showed that items required only one attribute are important for classification. There are some works that provided two basic sets of item discrimination index to measure discriminatory power of an item. The first one is based on descriptive measures from classical test theory, such as the global item discrimination index, and the second index is based on information measures from item response theory, including cognitive diagnosis index (CDI), attribute discrimination index (ADI), modified CDI and ADI. Results showed a strong relationship between these indices and the average correct classification rates of attributes. But their relationship to the indices may change as a function of the distribution of attributes.

There lacks an item quality index as a measure of item's correct classification rates of attributes. The purpose of this study was to propose an item discrimination index as a measure of correct classification rate of attributes based on Q-matrix, item parameters, and the distribution of attributes. Firstly, an attribute-specific item discrimination index, called item expected attribute matched rate (EAMR), was introduced. Secondly, a heuristic method was presented using EAMR for test construction.

The first simulation study was conducted to evaluate the performance of EAMR under the deterministic input noisy "and" gate (DINA) model. Several factors were manipulated for five independent attributes in this study. Four levels of correlation between latent attributes,  $\rho=0.00$ ,  $\rho=0.50$ ,  $\rho=0.75$ , and  $\rho=0.95$ , were considered. Items were categorized into five groups according to the number of attributes measured by each item. Item discrimination power was set at three levels, high, medium, and low. High level meant relatively smaller guessing and slip parameters, which were randomly generated from a uniform distribution  $U(0.05, 0.25)$ . Medium-level and low-level item parameters were randomly drawn from uniform distributions  $U(0.05, 0.40)$  and  $U(0.25, 0.45)$ . Next, 1000 items were simulated with the q-vector randomly selected from all possible attribute patterns measuring at least one attribute. Results showed that the new index performed well in that their values matched closely with the simulated correct classification rates of attributes across different simulation conditions.

The second simulation study was conducted to examine the effectiveness of the heuristic method for test construction. The test length was fixed to 50 and simulation conditions are similar to those used in the first study. Results showed that the heuristic method based on the sum of EAMRs yielded comparable performance to the famous CDI.

These indices can provide test developers with a useful tool to evaluate the quality of the diagnostic items. The attribute-specific item discrimination index will provide researchers and practitioners a way to select the most appropriate item and test that they want to measure with greater accuracy. It will be valuable to explore the applications and advantages of using the EAMR for developing item selection algorithm or termination rule in cognitive diagnostic computerized adaptive testing.

**Key words** correct classification rate, item expected attribute matched rate, test construction, the DINA model