心理学探新 2023,Vol.43,No.6,550 - 557
PSYCHOLOGICAL EXPLORATION

认知诊断测验的自动组卷方法*

马大付[1,2],秦春影[1,3],杨建芹[3],徐新爱[3],喻晓锋[1]
(1.江西师范大学心理学院,南昌 330022;2.山东省济南市教育教学研究院,济南 250002;3.南昌师范学院数学与统计学院,南昌 330032)

摘 要:认知诊断评估的重要一环是构建符合约束(统计约束与非统计约束)的诊断测验。其中构建测验的认知诊断自动测验组卷(cognitive diagnosis automated test assembly,CD - ATA)方法十分关键,受到广泛关注。文章通过梳理有关 CD - ATA 的研究,从组卷方法的类别、开发思路、组卷过程、方法对比及选用进行阐述与评析。最后,在已有方法基础上指出未来可从融合测验设计、非参数组卷、平行测验组卷、开发组卷软件、开展实证研究五个方面进行研究与探讨。

关键词:认知诊断测验;自动组卷方法;组卷精度;组卷效率
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003 - 5184(2023)06 - 0550 - 08

1 前言

测验项目是心理测量学中对被试潜在特质进行间接测量的工具。根据被试在项目上的作答结果,选取合适的模型与分析方法可实现对被试潜在特质的定量化评估(Rupp et al.,2010)。组卷是指从一个已校准的题库中选取一组同时满足统计(如测验长度和精度)与非统计约束(如内容平衡,答案平衡)的项目(Becker et al.,2021;Henson & Douglas,2005)的过程。如不考虑任何约束,从题库中组卷的数量将是巨大的,例如在一个包含 20 题的题库中选择其中 10 题进行组卷,共有 184756 种不同的组卷情况(Finkelman et al.,2009)。而随着题库项目数量的增加和试题管理约束的复杂化,人工组装高质量测验成为一项艰巨的任务(Lin et al.,2019)。自动测验组卷(automated test assembly,ATA;Finkelman et al.,2020)通过将组卷算法与计算机程序相结合,使根据测验需求进行自动组卷成为可能。

认知诊断评估(cognitive diagnosis assessment,CDA;von Davier & Lee,2019)作为新一代的心理测量理论,与项目反应理论(item response theory,IRT)关注被试的连续能力(θ)不同,其更关注对被试离散知识状态(knowledge state,KS)即属性的测量,这使得二者在构建测验的方法上不尽相同。首先,由于 θ 的连续性,IRT 自动测验组卷(IRT - ATA)常采用费舍尔信息量(fisher information,FI)作为测验组卷的方法。测验信息量为测验项目信息量的和(罗照盛,2012),测验信息量越高,测量误差越小,测验信度越高。而 KS 的离散性不满足 FI 的对数似然函数具有二阶导数的必要假设(Finkelman et al.,2009),因此基于 FI 的组卷方法无法直接推广至 CD - ATA(Finkelman et al.,2009;Henson & Douglas,2005);其次,二者组卷的复杂程度不同。IRT - ATA 与项目参数、被试 θ 有关,而 CD - ATA 则受认知诊断模型(cognitive diagnosis model,CDM)、项目 q 向量、项目参数与 KS 分布等因素的影响(de la Torre,2011;Song & Wang,2019),并且诊断测验项目 q 向量之间存在复杂的交互作用(丁树良 等,2010;Lin et al.,2017),这使得即使测验项目的参数相同,q 向量的不同组合也会产生不同的诊断结果。最后,即使 CD - ATA 成功组卷,也不存在精确的数学表达式能够描述测试项目与诊断准确性之间的关系(Lin et al.,2017;Wang et al.,2019),从而无法探知组卷结果的优劣。总之,因认知诊断测量对象的独特性,使得 CD - ATA 较 IRT - ATA 而言更加复杂。

为将诊断测验推向实际应用,国内外研究者针对 CD - ATA 问题提出多种组卷方法。Lin 等人(2017)将 CD - ATA 方法分为:基于指标组卷与基于模拟组卷两类,但却并未对各类组卷方法的发展脉络、组卷思想等进行更深入的探讨。文章通过阅读相关 CD - ATA 文献,结合国内外最新研究发现 CD - ATA 方法在整体上有着清晰的发展脉络,不同方法在组卷思想上存在诸多共性之处,且由于技术

* **基金项目**:国家自然科学基金(62341207,32360208),教育部教育考试院“十四五”规划支撑专项课题“高考实施过程中的科目跨年分数的转换研究”(NEEA2021050),江西省教育厅科技项目(GJJ2202013,GJJ212608,GJJ212602,GJJ2202018)。
通讯作者:喻晓锋,E - mail:xyu6@jxnu.edu.cn。

的发展，当前研究越来越面向实际应用，出现第三类组卷方法。起初，为沿用IRT－ATA使用FI组卷的方式，研究者提出基于信息量指标的组卷方法，并开发多种适用于CDA的信息量指标（汪文义 等，2018；Henson et al.，2008；Henson & Douglas，2005；Song & Wang，2019）。此后，基于作答模拟的方法被提出，该类方法在组卷前模拟一批作答数据，基于该批数据，使用启发式算法（heuristic algorithm）寻求合适的测验项目（Henson & Douglas，2005）。当前，研究者越发关注诊断测验的实际应用，在组卷时考虑更多与实际测验有关的信息，开发基于项目多信息的组卷方法。因此，文章拟对现有的CD－ATA方法进行论述，首先介绍组卷方法的发展脉络及其组卷思想，阐述不同方法之间的联系。其次对比不同类组卷方法之间的组卷思路、方法特征、优缺点，为使用者在方法选用上提供参考；最后，在现有组卷方法的基础上进行研究展望。

2　认知诊断测验自动组卷方法

2.1　基于信息量指标的组卷方法

信息量指标组卷方法试图沿用IRT基于信息量函数的组卷方式，因此定义CDA信息量指标是研究者开发组卷方法时首要解决的问题。根据CDA信息量指标能否直接反映项目的分类准确性，可将其分为间接信息量指标（下称间接指标）与直接信息量指标（下称直接指标）两类。间接指标采用项目对不同KS的区分能力作为项目的信息量，直接指标使用项目的期望分类准确率表示项目的信息量。上述两类指标均采用程序性组卷的方式，组卷时首先选择题库中信息量最高的项目进入测验，而后根据约束条件（如属性最少测量次数）筛选出题库中满足约束的项目，选取剩余题库中最高信息量的项目进入测验，以此类推，直至达到组卷长度。

2.1.1　间接信息量指标

（1）CDI和ADI

相对熵信息量（Kullback－Leibler information，KLI；Chang & Ying，1996）可用于描述两个概率分布的差异而不假设分布连续。项目j上任意两种知识状态α_u与α_v之间的反应概率分布距离可以描述为：

$$D_{juv} = \sum_{y=0}^{1} p(Y_{ij} = l \mid a_u) \log\left(\frac{p(Y_{ij} = l \mid a_u)}{p(Y_{ij} = l \mid a_v)}\right). \tag{1}$$

属性相互独立时，D_{juv}为一个$T * T(T = 2^K)$的D矩阵，K为属性数量。Henson和Douglas（2005）基于D矩阵提出认知诊断指标（cognitive diagnosis index，CDI）：

$$CDI_j = \frac{\sum_{u \neq v} (h(\alpha_u, \alpha_v)^{-1} D_{juv}}{\sum_{u \neq v} h(\alpha_u, \alpha_v)^{-1}}, \tag{2}$$

其中，$h(a_u, a_v)^{-1}$为a_u与a_v之间的海明距离倒数。CDI_j体现了项目j对所有KS的整体区分能力，项目CDI值越高表示项目的区分能力越强。

Henson等人（2008）认为CDI无法体现项目对单个属性的区分能力，只有当项目考察了某些属性，该项目才在该属性上存在区分能力，且当某些KS对之间的差异较大时，容易对项目的区分能力造成“虚高”的假象。因此可不必考虑差异较大的KS对，仅考虑在单个属性上存在差异的KS对。基于此定义了属性层面的区分度指标（attribute diagnosis index，ADI）：

$$ADI_j = \sum_{k=1}^{K} \left(\frac{1}{2^K} \sum_{allrelevantcells} D_{juv}\right)^{q_{jk}}, \tag{3}$$

其中$q_{jk} \in \{0,1\}$，0表示项目未考察该属性，1表示考察。ADI指标反映了项目在属性层面（attribute－specific）上的区分能力。

测验水平的CDI与ADI可表示为：

$$CDI = \sum_{j=1}^{J} CDI_j. \tag{4}$$

$$ADI = \sum_{j=1}^{J} ADI_j. \tag{5}$$

使用CDI与ADI指标组卷时，通常设置目标函数为$Maximize(CDI)$，$Maximize(ADI)$，即从题库中选择能使CDI与ADI和最大的项目组合，该项目组合有着最大区分能力。Zeng等人（2010）根据可达矩阵能够提高诊断测验准确性的原理，提出在使用CDI编制测验时添加可达矩阵，该方法提高了CDI组卷的诊断准确性。

（2）MCDI和MADI

Kuo等人（2016）对CDI与ADI展开修正，在原有指标的基础上增加属性层级结构权重与属性最少测量次数权重。校正后的MCDI与MADI（modified CDI；modified ADI）为：

$$MCDI_j = w_j^L w_j^H CDI_j, \tag{6}$$

$$MADI_j = w_j^L w_j^H ADI_j, \tag{7}$$

其中$w_j^L = (1 + I(r_L < 3) \sum_{v=1}^{V} I(q_j^* = s_v))^{-1}$，$w_j^H = (1 + r_H \sum_{v=1}^{V} I(q_j^* = s_v))^{-1}$。$V$为已选择的项目数量；$r_L = L / \sum_{k=1}^{K} L_k$，$L$为测验长度，$r_L < 3$表示

保证每个属性水平至少被测量 3 次；s_v 表示所有已选项目的 q 向量，若该 q 向量的项目已经被测量过，则对相同 q 向量项目施加惩罚权重，测量次数越多，获得的惩罚越多。r_H 为属性层级结构权重，其计算公式为：

$$r_H = \frac{\sum R_P^*}{K(K+1)/2}, \quad (8)$$

公式(8) 中的 R_P^* 为某种属性层级结构的可达矩阵(丁树良等,2010)，分母部分为属性二水平时线性层级结构的 R_P^* 之和。$r_H \in (0,1]$，当属性为线性层级结构时，r_H 为 1。当属性为独立结构时，$r_H = 2K/[K(K+1)]$。MCDI 与 MADI 同样采用程序性组卷的方式，但与 CDI 和 ADI 不同的是，MCDI 与 MADI 在组卷过程中项目的区分度值是不断变化的而非固定的值。

与 MCDI 和 MADI 组卷思想相同的是唐小娟等人(2013) 同样根据 CDI 未考虑属性层级结构，导致其整体判准率不高的缺点，提出一种基于属性层级结构的组卷方法。该方法通过定义项目类指标：$c_j(\alpha_u,\alpha_v) = \frac{\sum_{\alpha_u,\alpha_v} \neq jw_k \mid \alpha_{uk} - q_{jk} \mid}{\sum_{\alpha_u,\alpha_v} \neq jw_k \mid \alpha_{vk} - q_{jk} \mid}$，$c_j$ 值越大或越小均表示项目 j 越能区分 α_u 与 α_v，选择最能区分不同 KS 对的 q 向量进入测验，并在测验中适当加入可达矩阵，模拟研究结果表明，在属性数量 K 达到 8 个时，该方法的组卷效果要优于 CDI，但整体效果较差。

(3)RCDI 和 RADI

为保证诊断测验中不同属性具有类似的测量次数，基于属性平衡的方法，Su 和 Chu(2021) 对 MCDI 与 MADI 的 r_L 部分进行改进，将 $r_L < 3$ 修订为 $r_L = L/K$，提出修订的 CDI 和 ADI(revised CDI，revised ADI；RCDI，RADI)。组卷时，若单个属性的被测量次数未达到 r_L，$rCDI_j = w_j^H CDI_j$，$rADI_j = w_j^H ADI_j$；当属性测量次数达到 r_L，则设定包含该属性的项目的 $rCDI_j = 0$，$rADI_j = 0$，避免该属性被再次选中。

2.1.2 直接信息量指标

基于间接指标的组卷结果仅能表明测验项目具有较高的区分能力和可能具有较高的诊断准确率，却无法直接判断组卷结果的属性或模式判准情况。汪文义等人(2018) 以及 Song 和 Wang(2019) 提出一种可在无作答数据的情况下对项目各属性分类准确性进行预测的直接指标：期望属性分类准确率指标(expected attribute match rate，EAMR)：

$$EAMR_{jk} = P(\alpha_k = \hat{\alpha}_{k1} \mid X_J = 1)P(X_J = 1) + P(\alpha_k = \hat{\alpha}_{k0} \mid X_J = 0)P(X_J = 0), \quad (9)$$

其中，$\hat{\alpha}_{k1}$ 与 $\hat{\alpha}_{k0}$ 分别表示掌握与未掌握属性 k。当诊断模型为 DINA 模型时：

$$EAMR_{jk} = \frac{1 - s_j - g_j}{2^{K_j}} + 0.5. \quad (10)$$

$EAMR_{jk}$ 表示掌握属性 k 的被试在项目上正确作答并最终分类正确与未掌握属性 k 的被试在项目上错误作答并最终分类正确的概率之和。当属性 k 未被项目考察时，项目对该属性的期望正确分类率为 0.5。项目 j 的 $EAMR_j$ 为各属性的 $EAMR_{jk}$ 之和。

同样的，测验水平的 EAMR 为：

$$EAMR = \sum_{j=1}^{J} EAMR_j. \quad (11)$$

2.1.3 信息量指标优化算法组卷

除上述两类信息量指标方法外，Finkelman 等人(2010) 认为，在定义 CDA 项目信息量指标后，CD-ATA 应回归 IRT-ATA 使用优化算法的整体性组卷方式，优化算法的组卷结果可被证明是满足条件下的最优信息量指标项目组合。

(1)0-1 整数线性规划组卷

0-1 整数线性规划法(binary integer liner programming，BILP) 常用于在给定目标函数与多个约束条件的情况下，优化目标函数值。Finkelman 等人(2010) 将 BILP 用于 CD-ATA。以 ADI 指标为例(也可使用其他指标)，设定目标函数：

$$Z = maximize(\sum_{j=1}^{M} ADI_j x_j), \quad (12)$$

其中 M 为题库大小，ADI_j 为单个项目的属性区分度值，可通过公式(3) 得到；$x_j \in \{0,1\}$，为项目入选与否的指示变量。在约束条件的设置上可设置如组卷长度约束：$\sum_{j=1}^{M} x_j = J$ 及其他约束：$L_i \leqslant \sum_{x_j=1} C_{ij} \leqslant U_i, i = 1,\cdots,I$，$i$ 表示某具体约束，L_i 与 U_i 为约束 i 的下限与上限。目标函数可使用类似于 CPLEX 或 LINGO 等商用线性规划(linear programming，LP) 求解软件中的分支-切割方法(branch-and-cut) 求解。与 Finkelman 等人(2010) 使用方法相同的是 Kim(2004) 认为在 IRT-ATA 时认为可将项目的诊断信息作为一种非统计约束条件考虑进组卷过程。基于融合模型(Fusion Model，Hartz，2002)，添加属性最少被测量次数限制：$L_k \leqslant \sum_{x_j=1} q_{jk} \leqslant U_k$，区分度参数限制：$L_k \leqslant \sum_{x_j=1} r_{jk}^k \leqslant U_k$，同样地，组卷时可使用 0-1 整数线

性规划方法求解。

(2) 混合整数线性规划组卷

混合整数线性规划方法(Mix Integer Linear Programming,MILP) 的目标函数中既包括整数型决定变量,也包括连续型决定变量。Wang 等人(2021) 将该方法与项目 D 矩阵相结合,将其用于 CD - ATA。该方法首先去除项目 D 矩阵中对角线为 0 的元素,后将 D 矩阵转换为长度为 $T-1$ 的矩阵,再将其转换为列向量后按行拼接。经上述三步处理,将 D 矩阵转换为行为 $T(T-1)$,列为 1 的项目列矩阵。将题库中所有项目列向量按列合并为一个大小为行为 $T(T-1)$,列为 M 的题库矩阵:V 矩阵。设置目标函数为:

$$\min(f_1 x + f_2 y),$$

其中 $f_1 = (f_{11}, f_{12}, \cdots, f_{1M})^T, f_1 j = \sum D_j / [T(T-1)]$,表示 V 矩阵中每列的均值;$x = (x_1 x_2 \cdots x_M)^T$,为 0 - 1 向量集合,表示项目的整数决定变量。$f_2 = J \sum_{u=1}^{2^K(2^K-1)} \sum_{v=1}^{M} V_{uv} / (TM(T-1))$,为 y 的权重;$y = b_t - V_t x$,b_t 为组卷时设定的对第 t 对 α_u 与 α_v 的目标区分能力,V_t 为 V 矩阵第 t 行中的元素。

当不考虑 $f_2 y$ 部分时,MILP 方法与 BILP 方法类似,两者均是基于项目的 KLI,不同的是 MILP 基于项目的 D 矩阵,而 BILP 则是基于项目的 CDI 值。当考虑 $f_2 y$ 部分时,相较于 BILP 方法,MILP 方法保证了对每对 KS 进行足够的区分度测量,即区分度平衡。

2.1.4　基于信息量指标的组卷方法评价

基于信息量指标的组卷方法的结果与所定义的 CDA 信息量指标密切相关,由于属性的离散性,现有研究在定义 CDA 信息量指标时始终沿用一种如何将不同 KS 充分区分的思路。在得到信息量指标后,根据测验信息量最大化的组卷思想进行确定性组卷,即在确定题库项目、组卷指标、测验要求后,任一基于信息量指标的组卷方法从题库中所选择的项目是确定的。因仅进行一次组卷,而未与其他可能的组卷结果进行比较,这导致其组卷结果未必是全局最优。

2.2　基于作答模拟的组卷方法

该类方法通过事先模拟被试在项目上的作答数据,通过设立目标函数,将 CD - ATA 问题转换为在已有数据上寻求一组最符合目标函数的项目组合。由于能为诊断目的设立不同的目标函数,因此相较指标组卷方法,作答模拟组卷方法灵活度更高(Finkelman et al. ,2009)。

2.2.1　遗传算法组卷

遗传算法(generic algorithm,GA) 模拟自然界优胜劣汰的进化过程:具有更强适应能力的个体将在个体竞争中存活,并产生具有更强生存能力的后代。Finkelman 等人(2009) 将该方法用于 CD - ATA。GA 将题库中测验项目组合被视为单个个体,通过比较不同个体符合目标函数的程度,选择当前数据下接近最优的测验组合。GA 的具体组卷过程包括以下几步:① 产生一批包含 S 组初始项目的测验即父代,每个测验中包含数量为 J 的项目组合,初始项目组合可随机产生也可通过使用CDI的组卷方式产生;② 使用"变异" 策略,随机改变每个初始解中的一个项目,产生 $S*J$ 个子代;③ 评估包含父代在内的 $S*(J+1)$ 组解符合目标函数的程度;④ 根据③ 步的评估结果,选择最符合目标函数的前 S 组测验项目组合进入下一轮迭代;⑤ 重复步骤 ② - ④,直至达到最大迭代次数;⑥ 选择最后一次迭代中最优项目组合做为最优测验。

为使组卷结果更加符合实际,Finkelman 等人(2009) 提出三种目标函数:

$$F_1 = \sum_{\alpha} P(\alpha) E_a \left(\sum_{k=1}^{K} | \alpha_k - \hat{\alpha}_k | \right), \quad (13)$$

$$F_2 = \max_{k=1}^{K} \sum_{\alpha} P(\alpha) E_a (| \alpha_k - \hat{\alpha}_k |), \quad (14)$$

$$F_3 = \sum_{k=1}^{K} | e_k - \varepsilon_k |, \quad (15)$$

其中 $e_k = \sum_{\alpha} P(\alpha) E_a (| \alpha_k - \hat{\alpha}_k |)$,$| \alpha_k - \hat{\alpha}_k |$ 表示真实属性 k 与估计属性 k 之间的差异,当两者完全一致时为 0,不一致时为 1,F_1 与 F_2 值越小表示诊断测验的分类准确率越高,构建的测验越好。$| e_k - \varepsilon_k |$ 为属性 k 的实际分类准确率与设定的分类准确率之间的绝对差值,F_3 值越小表示构建的测验越符合测验所规定的目标分类准确率。

2.2.2　蚁群算法组卷

蚁群算法(ant colony optimization,ACO) 与 GA 类似,均属于求取目标函数的启发式算法。Lin 等人(2017) 将 ACO 用于 CD - ATA,提出基于蚁群算法的测验构建方法(test construction method based on ant colony optimization,ACO - TC),该方法将 CD - ATA 视为一种路径优化问题,题库中每一种测验项目的组合均被认为是一条路径,通过建立目标函数,在所有路径中寻求接近最优解的路径。

ACO - TC 过程大体上可分为三步:局部组卷、局部信息量更新与全局信息素更新。局部组卷时,单

个蚂蚁($a = 1,2,\cdots,A$)从剩余题库中选择满足条件约束的项目j的后验概率可为:

$$P_a(j) = \frac{\tau_j[\eta_a(j) + \gamma_a(j)]}{\sum_{t\in T}\tau_t[\eta_a(t) + \gamma_a(t)]}, \quad (16)$$

其中T为剩余题库中满足约束的项目集合,τ_j为项目的信息素浓度(初始组卷时设置$\tau_0 = 1$),η_a与γ_a分别为项目信息量指标与项目满足测验约束程度的权重,为提高组卷过程中的适应性,可设置η_a为多种项目信息量指标的组合。当蚂蚁a完成组卷后对其所选中的项目进行局部信息量更新:

$$\tau_j = (1-\rho)\tau_j + \rho\tau_0, \quad (17)$$

公式(17)中的$\rho \in (0,1)$表示信息素蒸发速率。当所有蚂蚁均完成组卷后,可设置公式(13)、公式(14)、公式(15)为目标函数,评估所有蚂蚁的组卷结果,最优项目组的目标函数可记为f_{best},最差组记为f_{worst}。后对f_{best}中的项目进行全局信息量更新:

$$\tau_j = (1-\rho)\tau_j + \rho\Delta\tau, \quad (18)$$

上式的$\Delta\tau = \exp\left[1 - \frac{f_{best}}{f_{worst}}\right]$,表示对$f_{best}$项目的信息素增量,增加$f_{best}$项目在下次迭代时被选中的概率。当某些项目从未被选中进f_{best}中,计算公式(17)可保持其τ_j始终为1,同样具有被选中的可能性。

2.2.3 基于作答模拟的组卷方法评价

作答模拟组卷方法依靠自身不断的循环迭代,每一次的组卷结果都建立在上一次组卷结果的基础之上,寻求更优于上一次组卷结果的题目组合,当组卷结果不再变化时,则表示寻得当前组卷方法下的最优题目组合。这种循环迭代的组卷方式,提高了找到全局最优解的可能性。但由于其需要大量的迭代计算,需要耗费的组卷时间也相对更长。

2.3 基于项目多信息的组卷方法

在实际测验中,测验的项目构成、测验形式以及测验的时限要求等都是测验开发者应当考虑的问题。为使组卷结果与实际测验要求更加一致,研究者进一步考虑更多可利用的项目信息,开发得到基于项目多信息的组卷方法。

2.3.1 基于多选项项目的组卷方法

现有研究对诊断数据的处理往往采用二分法(正确作答与错误作答两类),多项选择认知诊断模型(multiple choice CDM,MC - CDM)认为错误选项同样包含着属性的分类信息,这些信息同样可被可用于KS判别(Henson et al.,2018)。Henson等人(2018)将DINA模型下的区分度指标:$1 - s_j - g_j$,用于MC - CDM,提出一种广义的区分度指标(discrimination index,DI):

$$DI_{jk} = \frac{1}{2}\max_{\alpha}\left(\sum_{h=1}^{H_j} \left| P(X_j = h \mid \alpha) - P_{jh}(X_j = h \mid \alpha^{-k})\right|\right), \quad (19)$$

H_j表示项目j的选项数量,$P(X_j = h \mid \alpha)$表示α的被试选择选项h的概率,$P_{jh}(X_j = h \mid \alpha^{-k})$表示与$\alpha$仅在第$k$个属性上存在差异的KS选择选项$h$的概率。$DI_{jk}$定义了单个项目对属性$k$的区分能力。在使用DI组卷时,采用与CDI相同的指标线性求和的方式,测验水平的DI为:

$$\overline{DI} = \frac{1}{J}\sum_{j=1}^{J}\max_k(DI_{jk}). \quad (20)$$

2.3.2 基于反应时的组卷方法

Finkelman等人(2020)认为,尽管当前CD - ATA已能够获得丰富的信息,但还要保证被试所花的时间是可接受的,许多测验也含有一定的时限要求,因此其将反应时信息融入CD - ATA,作为测验组卷的约束条件,提出反应时组卷(response time assembly,RTA)。基于van der Linden(2006)提出的项目反应时模型:

$$f(t_j;\tau,A_j,\beta_j) = \frac{A_j}{t_j\sqrt{2\pi}}\exp\left\{-\frac{1}{2}[A_j(\ln t_j - (\beta_j - \tau))]^2\right\}, \quad (21)$$

τ,A_j,β_j分别表示被试速度参数,项目区分度与项目强度参数。基于该模型,对反应时t_j取对数后,其服从均值为$\beta_j - \tau$,标准差为A_j^{-1}的正态分布。定义项目j的反应时参数的函数,$q_j = \exp(\beta_j + A_j^{-2}/2)$;$r_j = \exp(2\beta_j + A_j^{-2})\{\exp(A_j^{-2}) - 1\}$;$s_j = \exp(3\beta_j + 3A_j^{-2}/2)\{\exp(3A_j^{-2}) - 3\exp(A_j^{-2}) + 2\}$。在假设给定速度参数$\tau$的情况下,被试在不同项目上的反应时间是相互独立的,因此不同项目之间的q_j,r_j与s_j之间便具备独立可加性。组卷时设置反应时条件为:

$$\zeta_q - \delta_q \leqslant \sum_{j=1}^{J} q_j \leqslant \zeta_q + \delta_q, \quad (22)$$

$$\zeta_r - \delta_r \leqslant \sum_{j=1}^{J} r_j \leqslant \zeta_r + \delta_r, \quad (23)$$

$$\zeta_s - \delta_s \leqslant \sum_{j=1}^{J} s_j \leqslant \zeta_s + \delta_s, \quad (24)$$

其中ζ_q,ζ_r与ζ_s分别表示q,r与s的目标条件,δ_s,δ_s与δ_s分别表示q,r与s的可容忍残差。在CD - ATA组卷时,RTA方法与BP相同,组卷时将反应时信息作为一种额外的约束条件,使用LP求解器进

行求解。

2.3.3　基于项目多信息的组卷方法评价

基于项目多信息的组卷方法在测验形式、测验要求等方面上更加贴合于实际情况，在组卷时考虑更多对测验结果可能产生影响的因素，并将其纳入组卷过程。但其对项目本身的要求更高，如多选项项目组卷方法需知道选择错误选项的概率，反应时组卷方法需知道作答项目的时间分布情况。

3　组卷方法比较与选用

3.1　组卷方法比较

文章已对现有的十多种 CD－ATA 方法进行介绍。接下来进一步对不同组卷方法进行比较，为实际使用者以及后续研究者在选用方法与开发新方法提供思路。表1详细呈现了不同组卷方法的分类情况、方法特征及优缺点。

表1　不同 CD－ATA 方法对比

组卷方法	具体方法	方法特征	缺点	优点
信息量指标组卷	直接指标 间接指标	信息量指标线性求和，逐项组卷	忽略项目 q 向量复杂的交互作用，缺乏灵活性	组卷速度快，组卷过程清晰
	优化算法	信息量指标线性求和，整体组卷		符合测验约束程度高
作答模拟组卷	遗传算法 蚁群算法	多次迭代组卷，概率性组卷	组卷时间长，算法复杂难以理解	组卷灵活，组卷精度高
项目多信息组卷	多选项组卷 反应时组卷	考虑更多信息，贴合实际测验	对项目本身信息了解程度要求高	更符合实际测验情形

从方法的大类上可以看出：①信息量指标组卷方法沿用 IRT－ATA 使用 FI 线性和的组卷思想，根据属性离散的特点，在 CDA 中寻找 Fisher 信息量的替代品。在组卷时通常设置满足约束条件的最大测验信息量项目组合，为确定性组卷方法。然而，该类组卷方法忽视了 CD－ATA 与 IRT－ATA 的不同，未考虑项目 q 向量之间复杂的交互作用，缺乏灵活性。②与信息量指标组卷不同的是，作答模拟组卷方法选择项目时是非确定性的，题库中的每个项目都有被选入测验的概率，为概率性组卷方法。通过不断地迭代更新，每次迭代后的结果均优于上一次迭代，最终得到最优项目组合。相较于信息量指标组卷方法，模拟作答组卷在组卷时尝试的项目组合类型更多（信息量指标组卷仅尝试一种项目组合）。但由于其算法复杂，计算量大，导致其组卷效率较低。③项目多信息组卷对项目信息了解程度要求高，且在组卷时部分依赖指标组卷的方法，因此也部分具有指标组卷存在的缺点。

3.2　组卷方法选用

通过对不同方法的比较，文章从组卷精度与组卷效率两种角度，为实际使用者在选用组卷方法上提供建议。

（1）组卷精度，诊断测验的首要目的是为获得较高的诊断精度（Rupp et al.，2010），尽管不同组卷方法存在一定的精度差异，但相较于随机组卷，本文所提及的组卷方法在属性数量较少的情况下均能够获得较高的判准精度。但属性数量较多时，指标组卷方法的判准率将迅速下降（Henson & Douglas，2005；唐小娟 等，2013），此时应当选用模拟组卷方法。另外，当组卷的目的是为了获得特定属性精度的测验时（Finkelman et al.，2009；Lin et al.，2017），指标组卷方法将无法适用，此时仅能通过模拟组卷。

（2）组卷效率，除组卷精度外，组卷效率也是施测人员需要考虑的问题（Finkelman et al.，2009；Lin et al.，2017）。模拟组卷因其在组卷时需不断地迭代更新项目组合，计算要求高，组卷时间长，组卷效率低。其他方法仅需在前期计算项目信息量指标时耗费一定的时间（郭磊 等，2016），实际组卷的时间较短，而且由于指标组卷均属于确定性算法，因此仅需计算一次项目信息量，即可多次运用。因此，如希望在短时间内得到组卷结果，可选择基于指标组卷的方式。

4　研究展望

尽管现有的 CD－ATA 方法已达十余种，但面对实际测验的多样性，有关组卷方法的研究与应用均有待进一步拓展，文章在已有方法基础上从理论性研究和实际应用角度出发提出几点展望。

融合测验设计，基于信息量指标的组卷方法仅关注于单个项目的 q 向量与项目参数，未考虑诊断测验的整体性，忽略测验 Q 向量在诊断测验中起到的重要作用。目前已有部分关于测验构建策略的研究（唐小娟 等，2022），而仅有少数组卷研究探讨过将信息量指标组卷方法与测验构建策略进行融合，融合测验构建策略后的结果也表明，信息量指标组

卷方法的组卷精度可获得大幅增长(Kuo et al.,2016;Su & Chu,2021;Zeng et al.,2010)。未来可进一步探讨将更多诊断测验设计与信息量指标组卷方法相互融合,在保证信息量指标组卷效率的基础上,进一步提高其组卷精度。

非参数组卷,当前 CD-ATA 方法均是在假定项目参数已知的情况下进行,而实际情况中,项目的实际参数是难以获得的。尤其是对于一些具有较复杂的诊断模型而言,准确的项目参数估计依赖于大量被试的作答反应。而当项目参数稳健性难以保证的情况下(Veldkamp et al.,2013),使用非参数组卷方法则势在必行,未来可开发更多非参数组卷方法。

平行测验组卷,平行测验(parallel test)是一种常用的实际测验形式,而文章所介绍组卷方法均只针对于构建单份测验。在查阅文献后,发现当前有关认知诊断平行试卷的构建方法仅有少数研究者(Li et al.,2021;Lin et al.,2019)有过相关探讨。未来也可开发同时能构建多份平行测验的 CD-ATA 方法。

开发组卷软件,尽管当前已开发了多种 CD-ATA 方法,但这些方法并不适用于没有编程基础的使用者,这也在一定程度上阻碍了组卷方法的实际应用。目前,有关研究者已将 IRT-ATA 组卷方法开发为相应的软件与开源 R 包(Becker et al.,2021;Shao et al.,2020),使用者仅需少量操作便可进行组卷,极大的简化了组卷过程,而 CD-ATA 中目前仅可通过使用 R 中的 CDM 包计算 CDI 与 DI 指标(George et al.,2016;Shi et al.,2021),尚未见完整的组卷 *R* 包或专业组卷软件,未来可开发相应诊断组卷软件。

开展实证研究,当前 CD-ATA 的实证研究相对较少。这一方面是由于国内外诊断测验的研究尚处于起步阶段,缺少系统性的测验开发、题库建设的过程,这在一定程度上阻碍了 CD-ATA 的实际应用。考虑到 CDA 在教学评估过程中的优良特性、未来可开发系统性的诊断测验题库,开展 CD-ATA 的实证研究。

参考文献

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,郑禅金,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.(2012).*项目反应理论基础*.北京师范大学出版社.

唐小娟,丁树良,毛萌萌,俞宗火.(2013).基于属性层级结构的认知诊断测验的组卷.*心理学探新*,*33*(3),252-259.

唐小娟,丁树良,俞宗火.(2022).题目属性向量平衡策略的认知诊断测验设计.*心理科学*,*45*(6),1466-1474.

汪文义,宋丽红,丁树良.(2018).分类视角下认知诊断测验项目区分度指标及应用.*心理科学*,*41*(2),475-483.

Becker,B.,Debeer,D.,Sachse,K. A.,& Weirich,S.(2021). Automated test assembly in R:The eatATA package. *Psych*,*3*(2),96-112.

Chang,H. H.,& Ying,Z. L.(1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*,*20*(3),213-229.

de la Torre,J.(2011). The generalized DINA model framework. *Psychometrika*,*76*(2),179-199.

Finkelman,M.,Kim,W.,& Roussos,L. A.(2009). Automated test assembly for cognitive diagnosis models using a genetic algorithm. *Journal of Educational Measurement*,*46*(3),273-292.

Finkelman,M. D.,de la Torre,J.,& Karp,J. A.(2020). Cognitive diagnosis models and automated test assembly:An approach incorporating response times. *International Journal of Testing*,*20*(4),299-320.

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*(5),310-326.

George,A. C.,Robitzsch,A.,Kiefer,T.,Groß,J.,& Ünlü,A.(2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*,*74*(2),1-24.

Hartz,S. M.(2002). *A Bayesian framework for the unified model for assessing cognitive abilities:Blending theory with practicality*.(Unpublished doctoral disseration). University of Illinois at Urbana-Champign.

Henson,R.,& Douglas,J.(2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*,*29*(4),262-277.

Henson,R.,Roussos,L. A.,Douglas,J.,& He,X. M.(2008). Cognitive diagnosis attribute-level discrimination indices. *Applied Psychological Measurement*,*32*(4),275-288.

Henson,R.,DiBello,L.,& Stout,B.(2018). A generalized approach to defining item discrimination for DCMs. *Measurement:Interdisciplinary Research and Perspectives*,*16*(1),18-29.

Kim,S.(2004). *An automated test assembly for unidimensional IRT tests containing cognitive diagnostic elements*.(Unpub-

lished doctoral disseration). The University of Texas at Austin.

Kuo, B. C., Pai, H. S., & de la Torre. (2016). Modified cognitive diagnosis index and modified attribute – level discrimination index for test construction. *Applied Psychological Measurement*, *40*(5), 315 – 330.

Li, G. Y., Cai, Y., Gao, X. L., Wang, D. X., & Tu, D. B. (2021). Automated test assembly for multistage testing with cognitive diagnosis. *Frontiers in Psychology*, *12*, 509844.

Lin, Y., Gong, Y. J., & Zhang, J. (2017). An adaptive ant colony optimization algorithm for constructing cognitive diagnosis tests. *Applied Soft Computing*, *52*, 1 – 13.

Lin, Y., Jiang, Y. S., Gong, Y. J., Zhan, Z. H., & Zhang, J. (2019). A discrete multiobjective particle swarm optimizer for automated assembly of parallel cognitive diagnosis tests. *IEEE Transactions on Cybernetics*, *49*(7), 2792 – 2805.

Rupp, A. A., Templin, J., & Henson, R. J. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

Shao, C., Liu, S., Yang, H., & Tsai, T. H. (2020). Automated test assembly using SAS operations research software in a medical licensing examination. *Applied Psychological Measurement*, *44*(3), 219 – 233.

Shi, Q. Z., Ma, W. C., Robitzsch, A., Sorrel, M. A., & Man, K. W. (2021). Cognitively diagnosis analysis using the G – DINA model in R. *Psych*, *3*(4), 812 – 835.

Song, L. H., & Wang, W. Y. (2019). An attribute – specific item discrimination index in cognitive diagnosis. *Quantitative Psychology*, 169 – 181.

Su, Y. H., & Chu, K. H. (2021). Improving the measurement efficiency in test construction related to cognitive diagnosis models. *Quantitative Psychology*, 243 – 251.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*(2), 181 – 204.

Veldkamp, B. P., Matteucci, M., & de Jong, M. G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement*, *37*(2), 123 – 139.

von Davier, M., & Lee, Y. S. (2019). *Handbook diagnostic classification models*. New York: Springer.

Wang, W. Y., Song, L. H., Chen, P., & Ding, S. (2019). An item – level expected classification accuracy and its applications in cognitive diagnosis assessment. *Journal of Educational Measurement*, *56*(1), 51 – 75.

Wang, W. Y., Zheng, J. J., Song, L. H., Tu, Y. K., & Gao, P. (2021). Test assembly for cognitive diagnosis using mixed – integer linear programming. *Frontiers in Psychology*, *12*, 623077.

Zeng, L. Y., Ding, S. L., & Gan, D. W. (2010). Test construction for cognitive diagnosis. In 2010 *Asia – Pacific conference on wearable computing systems* (pp. 12 – 15). Shenzhen: IEEE.

Research on Automated Test Assembly Method for Cognitive Diagnosis Tests

Ma Dafu[1,2], Qin Chunying[1,3], Yang Jianqin[3], Xu Xinai[3], Yu Xiaofeng[1]

(1. School of Psychology, Jiangxi Normal University, Nanchang 330022;
2. Jinan Institute for Education and Teaching Research, Jinan 250002;
3. School of Mathematics and Statistics, Nanchang Normal University, Nanchang 330032)

Abstract: Constructing diagnosis tests that meet test constraints (statistical constraints and non – statistical constraints) is an important part of cognitive diagnosis evaluation, which is helpful to realize the diagnosis evaluation of participants' attributes. The cognitive diagnosis automated test assembly (CD – ATA) is a commonly used diagnosis test assembly method. This paper first classifies the existing CD – ATA methods, and then discusses the test assembly ideas, test assembly processes, and connections between different methods. Secondly, it compares different methods to provide users with reference when selecting methods. Finally, on the basis of the existing methods, it is proposed that future research can be carried out from five aspects: fusion test design, non – parametric test assembly, parallel test assembly, developing test assembly software, and application scenario expansion.

Key words: cognitive diagnosis tests; automated test assembly; assembly accuracy; assembly efficiency