文章编号:1000 - 5862(2019)05 - 0441 - 07

0-1评分认知诊断测验设计

丁树良,罗 芬,汪文义,熊建华
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要:认知诊断测验设计实质上是测验 Q 矩阵设计,设计应最大限度覆盖诊断的构念,充分发挥代表认知模型的属性层级关系的重要作用. 主张测验充分表达被试知识结构,提倡测验 Q 矩阵和被试知识状态共享同一层级结构,以实现对被试更加准确的诊断. 对于非独立型层级结构,无法实现属性使用次数平衡而应采用题目属性向量使用平衡策略. 对 Liu Ren(2017) 的测验设计提出质疑.

关键词:认知诊断测验设计;Q 矩阵设计;属性层级关系;属性向量平衡
中图分类号:B 841 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2019.05.01

0 引言

认知诊断测验设计本质上是测验 Q 矩阵设计[1]. 测验 Q 矩阵是缩减 Q 矩阵[2-3]的子矩阵,缩减 Q 矩阵的每一列对应一种潜在题目属性向量,故又称为潜在 Q 矩阵[4]. 潜在 Q 矩阵获得的前提是属性及其层级关系确定,这也是测验 Q 矩阵设计的前提. 抽象地说,测验 Q 矩阵设计是按照某些目标,如何从潜在 Q 矩阵中选择某些列组成测验 Q 矩阵才能够达到目标. 在一般的 Q 矩阵设计中是假定属性及其层级关系是正确的,并且 Q 矩阵中元素标定是准确的(比如模拟研究满足这些假设). 可以通过 Q 矩阵行的逐对比较方法[2-3]获得潜在 Q 矩阵或者其子矩阵表达的属性层级关系,但是从其子矩阵挖掘出的属性层级关系不一定与潜在 Q 矩阵表达的层级关系一致[4]. 当然,属性及其层级关系和 Q 矩阵元素标定是否准确,是应该仔细考虑和探索的问题,限于本文主旨,对此不进行讨论.

研究者根据一些研究结果或者探索某些可能影响测量的信度、效度的因素(比如测验长度、对属性的测量次数是否平衡、对题目属性向量测量次数是否平衡、对不同知识状态对应的理想反应模式是否不同、题库中题目质量的高低等),决定从潜在 Q 矩阵中选取部分列构成测验 Q 矩阵,希望经过设计以后的测验 Q 矩阵能够准确测量被试的知识状态. 综上所述,认知诊断测验设计应该包含如下内容:对于给定的领域范围,存在多少种可供选择的题目类型,以及从这些题目类型中根据某些什么样的原则(准则/目标)选择题目组成一个 Q 矩阵,以达到预期的目标.

对于 0-1 评分认知诊断,在属性之间互相不为先决的条件下,即属性层级结构为独立型,此时,可达阵是单位阵. Chiu Chia-Yi 等[5]、L. T. DeCarlo[6]、M. J. Madisond 等[7]的研究结果,可以归结为一个准则,即单位阵作为测验 Q 矩阵的子矩阵,可提升诊断分类精度,反之,则分类精度不理想[8].

J. S. Gorin[9]不主张每一个题目至少包含 2 个属性,并且认为测验 Q 矩阵应该多包含潜在 Q 矩阵的列. K. K. Tatsuoka[2-3]认为不论属性层级关系是否是独立型结构,当测验 Q 矩阵是充分 Q 矩阵(sufficient Q matrix)(即测验 Q 矩阵蕴含可达阵)时可以提高测验构念效度(construct validity)[2-3]. 但是,充分 Q 矩阵对应的诊断分类精度可能低于非充分 Q 矩阵[10].

J. P. Leighton 等[11]认为应将潜在 Q 矩阵作为测验 Q 矩阵,但是当属性数目比较大且层级结构相当松散时(比如独立型或者无结构型),潜在 Q 矩阵

收稿日期:2019-04-22
基金项目:国家自然科学基金(61967009,31500909,31360237,31160203),国家社会科学基金(16BYY096),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(2016BAB212044),江西省社会科学规划课题(17JY10),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资助项目.
作者简介:丁树良(1949-),男,江西樟树人,教授,主要从事计算机辅助教学及教育和心理测量方面的研究. E-mail:ding06026@163.com

的列数较多,一般认知诊断测验无法容纳这么多项目. 罗欢等[12] 认为,对 $K=7$ 的独立型属性层级结构,测验 Q 矩阵中除考察 1 个属性的项目之外,还安排所有包含 2 个属性的项目,这是题目属性向量平衡的做法; 丁树良等[13-14] 认为,若属性之间不存在补偿作用并且采用 0-1 评分,则必要 Q 矩阵可以使得理想反应模式和知识状态一一对应,从而提高认知诊断分类的精度. 所谓必要 Q 矩阵就是可达阵作为子矩阵的 Q 矩阵[15] ,考虑到独立型结构对应的可达阵是单位阵,这推广了 Chiu Chia-Yi 等[5] 的相关结果. 彭亚风等[16] 讨论了认知诊断测验设计,给出可达阵等价类 R^* ,即对应项目可达阵列的相互置换后,将 R^* 作为测验 Q 矩阵的子矩阵,仍然可以使知识状态集合和理想反应模式集合一一对应,并且他们还发现诊断分类的精度不仅和测验长度相关,而且还和测验长度与属性数目的比例有关. 从必要 Q 矩阵一定可以挖掘正确的属性层级关系,即可以代表认知模型,这个事实或许是必要 Q 矩阵能够提高诊断精度的原因.

本文主题是认知诊断测验设计. 因为属性层级关系代表认知模型,认知诊断测验欲达到最大限度覆盖诊断的构念,就应该高度重视属性层级关系在测验设计中的作用. 测验 Q 矩阵对应的属性层级结构应该尽量和被试知识结构相同. 由扩张算法,可达阵能够生成潜在 Q 矩阵,可达阵和潜在 Q 矩阵中对应出的层级关系完全一致,故可达阵在认知诊断测验编制中占重要地位; 在考虑非统计约束的测验设计时,对独立型结构项目属性向量平衡蕴含属性平衡,反之不真; 可供选择的项目类实际上对应潜在 Q 矩阵的列的集合. 本文先讨论属性层级关系的“大小”,分析当命题专家认定的属性层级关系与被试群体知识状态集合对应的属性层级关系不一致时,可能出现的问题及应对策略,并对 Liu Ren 等[17-18] 结果进行评论,最后给出一些总结和讨论.

1 属性层级关系的“大小”

1.1 新的基本属性层级关系的划分

结合 Liu Ren 等[17] 和丁树良等[19] 的研究结果,J. P. Leighton 等[11] 定义的基本属性层级关系可以划分为独立型、根树型(包含线性型、发散型和无结构型) 和倒金字塔型(inverted pyramid type) . 其他属性层级关系可以由它们复合而成,比如收敛型结构可分解为根树型和倒金字塔型这 2 个更加简单的结构的复合.

Liu Ren[18] 对线性型认知结构采用独立型结构进行测验设计,对此,面临一个问题,即专家可从与认知结构相异的结构入手进行测验设计吗? 一般地,既然是“设计”,就应该贴近实际,但有可能对属性及其层级关系没有百分之百的把握,这时如何做才比较稳妥,或者说,比较“稳健(robust)”,下面仅对属性确定而层级关系存在争议的情况讨论. 本文先讨论基本属性层级关系.

1.2 属性层级关系的大小

定义 1(属性层级关系的大小) 固定属性数目 K,如果层级关系 H_i 对应的潜在 Q 矩阵分别为 Q_i $(i=1,2)$,并且 Q_2 是 Q_1 的子矩阵,则称 H_1 大于 H_2 (或者 H_1 包含 H_2) .

定义 1 考察属性层级关系的“包含关系”. 易知,当属性数相同时,独立型和线性型层级结构分别对应“最大”和“最小”的层级结构. 杨淑群等[20] 定义属性层级关系的紧密性. 可以看出,固定属性数 K,属性层级关系 H_1 越松散,对应的潜在 Q 矩阵的列数越多,越容易包含那些比 H_1 结构紧密的层级关系.

1.3 使用实证手段探查独立型结构

由定义 1 知独立型层级结构可以包含相同属性数的其他层级结构,从这个意义上说,在属性层级结构存在争议时,采用独立型结构比较稳妥; 但是对于诸如规则空间方法(RSM)[2-3] 的分类方法,可能会增加许多不必要的类别(纯规则点) ,对于命题专家也可能存在对应的潜在 Q 矩阵的列无相应的题目可出的困境,因此有必要讨论到底什么时候能够出现独立型层级结构,或者说给定属性及其层级关系以后,层级关系是否还要修改.

专家认定的属性层级结构是可以修正的,甚至是必须多次讨论修改,特别是命题以后或者被试测验结果获得以后,根据预测结果进行修正. 一方面收集被试信息,请命题专家或根据学生口语报告尽量补全或完整地标出在学生作答时可能应用的所有其他属性或相关先决属性,这有利于根据实证数据和学科专家信息分析学生作答背后的认知模型或属性层级结构,如喻晓锋等[21] 基于被试作答反应使用贝叶斯网获得属性层级关系.

收集命题专家信息: 首先,命题专家和测量专家应该给出属性及其层级关系以指导命题; 其次,对应这个属性集合和层级关系,设计测验 Q 矩阵(对应了 1 个属性层级关系,记为 T_0) ; 接下来命题专家应该努力命制包含属性最少的题目(纵使专家认为它

们太简单而不一定入选作为实际的测验项目)，特别地，如果命题专家能够命制单位矩阵各列对应的题目，这表示潜在 Q 矩阵包含单位阵，那么所感兴趣的领域中的属性一定是独立型结构; 反之，则不是独立型结构. 如果命题专家能够命制的包含属性尽量少的题目，并且对应的题目的属性向量不能够由测验 Q 矩阵表达，那么就应该修正这个测验 Q 矩阵，同时依据这个新的测验 Q 矩阵修正原来的属性层级结构，将新的层级结构记为 T_1; 最后，将 T_1 看成 T_0，重复上述步骤，直至 T_0 稳定为止.

比如，异分母加减法运算，有的专家认为这对应 4 个属性: 基础知识(A_1) 、同分母加减法(A_2) 、求最小公倍数或通分(A_3) 、异分母加减法(A_4) . 它们的层级结构为 A_1 是 A_2 和 A_3 的先决属性、A_2 和 A_3 是 A_4 的先决属性，其形状为 4 元素收敛型[11] (见图 1) . 但是有专家认为 A_3 实际上是求最小公倍数，可以单独命制只测最小公倍数或者通分的试题，实际上最小公倍数和分数运算彼此不存在先决关系. 根据这一点，属性层级关系就应该修正，相应的 Q 矩阵(特别是潜在 Q 矩阵和测验 Q 矩阵) 应该修改(见图 2) .

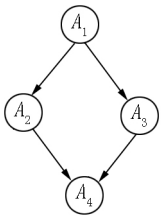

图 1　收敛型

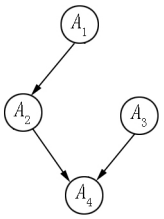

图 2　新层级模型

这个例子显示，当测验 Q 矩阵的初稿给出以后，请命题专家命制包含属性最少的题目很有必要. 它可以改变可能是固化在人们心目中的认知模型(潜在 Q 矩阵) ，而更加接近真实的认知模型，从而 Q 矩阵也更加合理，在老师(命题专家) 心目中的 Q 矩阵和在学生心目中的知识状态更加一致而不会不相容.

2　Q 矩阵设计的某些原则和标准

2.1　从划分层级关系类目的角度审视

设 $S=\{A_1,A_2,\cdots,A_K\}$ 为感兴趣的领域中的所有属性，为了便于阅读，使用大家熟悉的 J. P. Leighton 等[11] 对基本属性层级关系划分的术语，设 $G=$ { 线性型，收敛型，发散型，无结构型} ，即这些属性层级关系对应的可达阵一定不是单位矩阵，而独立型层级关系对应的关系矩阵是单位矩阵. 如果 S 中的属性对应 G 中某一个层级结构，这个层级关系中至少有 2 个属性，它们之间存在先决关系，有如下的一些结论.

定理 1　设所感兴趣的预诊断的领域中的属性集合为 $S=\{A_1,A_2,\cdots,A_K\}$，若 S 中的属性层级关系是独立型，则 S 对应的层级关系不属于 $G=$ { 线性型，收敛型，发散型，无结构型} ; 反之，若 S 中属性对应的属性层级关系属于 G，则一定不可能对应独立型层级结构.

证　根据独立型和 G 中这些属性层级结构的定义和先决关系的定义，以及在讨论同一个问题时，同一个属性标识符只能够代表一个属性的原则，则结论显而易见.

定理 1 可以导出一个重要推论.

推论 1　对于同一个属性集合 S，命题专家给出的 Q 矩阵对应的层级结构是独立型，则被试的知识结构不可能是 G 中层级结构.

这个推论意味命题专家若认为属性层级关系是独立型，则一定可以命制相应的潜在 Q 矩阵的列对应的题目，从而被试的知识状态也应该和潜在 Q 矩阵的列以及零向量对应，从而不可能对应 G 中层级结构.

有人认为，如果有 3 个属性 A_1、A_2、A_3，它们组成的线性型是指被试对属性的掌握过程，而认知诊断测验设计的任务是怎么测试才能够更有效率，因此题目的设计可以是只测 A_2，可以是同时测 A_1A_2，或者同时测 A_2A_3. 如果问 “(011) ”这样的题目是否符合线性型结构，关键是要区分线性型结构指的是人的掌握过程还是命题专家命制的题目，题目的设计可以不用考虑是什么结构，考虑更多的是内容(content) . 本文称这样的想法是 “人卷分离”. 定理 1 意味着人卷分离是难以实施的，本文主张 “人卷合一”，即被试的知识状态只能够是所有属性对应的属性层级结构(试卷应该充分表达这个结构) 的一部分，也就是命题专家和被试的知识状态是共享同一个层级结构，或者说命题专家是站在被试的角度看待问题的.

通常，属性层级关系是指属性之间的心理加工逻辑顺序，比如 K. K. Tatsuoka[3] 、M. J. Gierl 等[12] 将结构化属性称为属性层级结构，定义为这些属性求解测验问题所需要的属性的心理顺序(the psychological ordering) ，从而作为任务表现的认知模型(serve as the cognitive model of task performance) . 一般地，教学的顺序和学生掌握过程吻合，属性层级关系和掌握过程不能够等同，同一篇文章所说的概念应该包含同样的意义，否则无法让人理解. 在试题命制时所说的属性层级关系指的是认知的逻辑顺序，而不是学生对属性的掌握过程，这 2 个概念不应该

混淆.

当然,专家和生手的知识结构可以不同,一般地,问题解决时专家和生手使用的策略不同,但是如果规定使用相同的策略,那么属性及其层级结构就应该相同. 所以,专家的知识结构(属性层级关系)应该包含生手的知识结构.

推论 2 对于同一个属性集合,如果属性集合层级关系是独立型,则认知诊断测验设计可以既是独立设计,又是邻接设计或者可达设计;如果属性层级关系是根树型或者逆金字塔型结构,则不可能采用独立设计.

2.2 Liu Ren 的设计合理吗?

Liu Ren[18] 针对非独立属性层级结构,提出 3 种测验 Q 矩阵设计方案:独立设计、邻接设计和可达设计. 他认为这 3 种设计可以使得测验中属性使用数目平衡,并且认为采用邻接设计诊断分类结果最好. 邻接阵和可达阵中的"邻接"、"可达"和邻接设计、可达设计的"邻接"、"可达"不一样,之所以这样命名,只是一种"借用"而已. 因此不必过于纠结这 3 种设计的名字. 本文从先决关系定义的角度审视它们的合理性.

对线性型结构,Liu Ren[18] 给出具体的对应于独立设计、邻接设计和可达设计的测验 Q 矩阵,其中行对应属性. 对 5 个属性的线性型结构,这 3 种设计矩阵分别为

$$\begin{bmatrix}1000010000\\0100001000\\0010000100\\0001000010\\0000100001\end{bmatrix},\begin{bmatrix}1000010000\\1100011000\\0110001100\\0011000110\\0001100011\end{bmatrix},\begin{bmatrix}1111100001\\0111100011\\0011100111\\0001101111\\0000111111\end{bmatrix}.$$

如果前 5 列为第 1 部分,后 5 列为第 2 部分,则上述独立设计是 2 个单位矩阵对应的测验 Q 矩阵;邻接设计是 2 个下三角矩阵构成,这个下三角矩阵只有对角元和对角线邻接元为 1;可达设计第 1 部分是上三角 0-1 矩阵,对所有 $i\leqslant j, a_{ij}=1$;第 2 部分构成的方阵,次对角线及其下方的元素全部为 1,而次对角线上方的元素全部为 0.

Liu Ren[18] 的测验设计进行考察.

(i) Liu Ren[18] 使用的是模拟研究,故可假定属性及其层级关系、Q 矩阵标定都正确. 考察 Liu Ren[18] 给出的上述 3 个 Q 矩阵,使用"行的逐对比较方法"[24],"邻接设计"对应的 Q 矩阵可以挖掘出属性 2 是属性 1 的先决属性(这和文献[17],图 1 中属性 1 是属性 2 的先决恰恰相反),而从包含 10 列的独立设计和可达设计对应 Q 矩阵挖掘出来的层级关系都是独立型属性层级关系,这和线性型层级结构相差甚远(但是由推论 2,知独立型结构"包含"线性型结构).

(ii) 正如彭亚风等[16] 指出"Liu Ren 等[17] 的研究在模拟 Q 矩阵时会出现不合理的考核模式,例如直线型情况下使用独立方法生成的测验项目都是考察单个属性而忽略其先决属性,这违背了直线型的关系假设".

(iii) 邻接设计对应的项目属性向量也多数不是线性型潜在 Q 矩阵中的向量,比如(01100),(00110),(00011),显然都不符合线性型层级关系,它们均缺少公共的先决属性.

(iv) 可达设计中既然前半部分(左边 5 列)符合线性型层级结构,即属性层级关系是 A_i 是 A_{i+1} 的先决属性,$i=1,2,3,4$;而后半部分(右边 5 列)却将属性层级关系完全颠倒,变成为 A_{i+1} 是 A_i 的先决属性,$i=1,2,3,4$. 根据先决关系满足自反性得知,如果 A_i 与 A_{i+1} 互为先决,则 $A_i=A_{i+1}$,$i=1,2,3,4$. 于是 5 个属性等同于 1 个属性! 这是一个令人惊诧、难以接受的结论!

2.3 题目属性向量的平衡与属性平衡的关系

在具有诊断功能的认知诊断计算机和自适应测验(CD-CAT)选题策略的研究中,有研究认为考虑属性平衡非常重要[23-25].

CD-CAT 选题策略和 Q 矩阵设计的任务都是选择项目,所以 CD-CAT 选题策略的研究结果当然对于认知诊断测验编制有一定的借鉴意义. 由于认知诊断测验的"最小存在单位"是项目,就好像"分子"可以独立存在,项目属性向量是附着于项目的重要信息,故研究项目属性向量的平衡应该是有意义的;而且,除独立型结构之外,其他属性层级结构的某些属性存在先决关系(比如线性型,发散型,无结构型等),此时要达到属性使用次数平衡是做不到的,只能够退而求其次,实现题目属性向量使用的平衡.

属性平衡不一定可以导出题目属性向量平衡的结论. 比如包含 3 个属性的独立型结构,测验 Q 矩阵包含 5 列,前面 3 列对应单位矩阵,第 4 列仅仅包含属性 1、属性 2,第 5 列仅仅包含属性 3,则每一个属性安排测试 2 次,但是题目属性向量不平衡. 反之,对于独立型结构,若题目属性向量平衡,则属性平衡. 因为从 K 个属性取出 h 个的组合一共有 $C(K,h)$ 种,而所有组合中出现某个属性的次数是一个常数,等于 $hC(K,h)/K$,因此属性使用次数平衡. 请注意,这个结论对其他层级结构不一定成立.

2.4 测验 Q 矩阵设计的一个必要条件

笔者认为 Q 矩阵设计的最基本的要求是测验 Q 矩阵的列的布尔并中每一个元素均等于 1. 将它

列为一条定理,以便引用.

定理 2　测验 Q 矩阵的所有列的布尔并的每一个元素应该等于 1,否则至少存在 2 个不相同的知识状态对应的理想反应模式相同.

证　不失一般性,可以假设可达阵是对角元全部等于 1 的上三角布尔矩阵(即 0-1 矩阵). 注意到可达阵的对角元均为 1,如果测验 Q 矩阵的所有列的布尔并的某一个元素不等于 1,那么可达阵的某一列一定不在这个测验 Q 矩阵之中. 从而必有 2 个不同的知识状态对应相同的理想反应模式[26-27],因此这个测验 Q 矩阵难以区分这 2 个不同的知识状态.

3　结论与讨论

3.1　结论

在理想反应条件下,若属性之间不存在补偿作用且采用 0-1 评分,则可达阵(或者其列进行交换所得到的矩阵,即可达阵的等价类)可以使知识状态与理想反应模式一一对应[13-14];但是观察反应模式中带有随机误差,随机误差越大,则可达阵设计的方法的效果越差,因此寻找对带随机误差能够比较稳健的测验设计是十分重要的[28-29];如果所感兴趣的领域中的所有属性是独立的,那么所有被试的知识状态集合对应的属性层级结构不可能是线性型,此时 Liu Ren[18] 提出的 3 种设计不可能全部实现. 他提出的测验中属性出现的数目平衡及题目属性向量类型出现的次数平衡的原则对于独立型层级结构可以同时满足,而且题目属性向量的平衡可以导出属性平衡;而对于非独立型属性层级结构不可能同时满足.

对于测验 Q 矩阵设计,存在 2 种看法(做法),一种是"人题合一型",即测验 Q 矩阵的层级关系和被试的认知层级关系完全一致,比如都是线性型;另外一种是"人题分离型",即测验 Q 矩阵对应的层级关系比全体被试对应的认知层级关系要"大",比如测验 Q 矩阵对应的属性层级关系是独立型(可达阵是单位矩阵),而被试认知属性层级关系非独立型. 这种做法的优点是比较稳妥,对于属性层级关系存在争议的情况下,这不失为是一种稳妥的方案,不至于出现某种知识状态没有相匹配的题目属性向量,这也可以看成是许多模拟研究中,为什么认知诊断测验的设计往往采用独立型结构的理由;但是本文认为,命题时如果能够针对被试认知情况,命制和被试非零知识状态对应的题目(属性向量),是有好处的. 比如考察被试对分数运算掌握情况,通分是正确完成异分母加减运算的先决属性,但是考察通分这个属性可以看成是考察求 2 个正整数的最小公倍数这个属性,而求 2 个正整数的最小公倍数是和分数运算相互独立的属性,因此测验中可以命制仅仅考察求最小公倍数的项目,以考察被试对此的掌握情况.

3.2　讨论

3.2.1　**关于模拟研究**　模拟研究表明,Liu Ren[18] 提出的 3 项认知诊断测验设计中邻接设计效果最好. 问题是,对于线性型层级结构,现实中可以命制这样的题目吗? 难道这样的邻接结构或者可达结构满足线性型结构吗? 如果一个研究,既有理论,又有模拟研究,当然十分漂亮,但是模拟研究的条件应该尽量贴近实际情况,这正如其他产品开发,不考虑实际情况是行不通的. 比如研究者设计好一个认知诊断测验(或者说给出一个测验 Q 矩阵),请求命题专家编制相应的测验(哪怕是对应地编制样题);如果这个设计无法实现,那么就应该仔细检查,看一看哪一个环节出了问题,并加以改正. 测验设计者和命题专家反复磨合,才能获得成功.

3.2.2　**属性的先决关系和知识状态的偏序关系有区别**　注意,属性层级关系的表达一是采用图示的方法,其中哈斯图是图示的简洁表达;二是采用矩阵表达,其中邻接矩阵表示直接的邻接关系(直接先决关系),或者说是"父子关系",而可达阵表示直接和间接关系,即先决关系(prerequisite relation). 先决关系是定义在属性集合上的一种偏序(partial order relation). 至于知识状态之间的大小关系,虽然也是偏序关系,但是这种关系是根据 2 个知识状态(向量)的差向量来判断,如果这个差向量的每一个分量均非负,那么这 2 个知识状态可以比较,否则不可以比较. 向量之间这种偏序,称为 Lowner 偏序. 属性的先决关系和知识状态的 Lowner 偏序显然不是同一种偏序关系. 例如独立型结构的属性之间不可以相互比较(即没有先决关系),但是 Q 矩阵中的项目属性向量(或者知识状态)集合上可以定义 Lowner 偏序.

3.2.3　**如何表达熟手和生手的知识结构的不同**　专家(熟手)和学生(生手)的知识结构可以不同,甚至很多情景下是不同的,比如专家可以使用代数方法求解算术问题. 在某个学习阶段,要求被试掌握某些求解方法,专家命题是应该是"设身处地",要求被试学习过的方法求解,而不是超越他们的知识范围,所以尽管 2 者知识结构不同,面临的测验对应的属性应该相同;而确定的属性集合,属性之间的层级结构一般来说应该是确定的,不应该因人而异. 生手和熟手之间的知识结构的不同,是否意味着"解题"策略的不同? 或许熟手的"解题"策略对应简洁明

快甚至巧妙的方法,而生手的"解题"策略对应的是复杂、"笨拙"的方法? 所谓的"熟能生巧"恐怕说的是这么一回事. 但是,不同的策略对应的 Q 矩阵不同. 同一个 Q 矩阵是不是对应同一个"解题"策略? 同一个 Q 矩阵挖掘出来的属性层级结构是否也就确定? 特别是通过被试的"出声思考"以后获得的 Q 矩阵应该反映被试的知识结构. 所以认为一个测验的设计可以既是邻接设计,又是可达设计或者独立设计,恐怕不一定合理.

当然,专家考虑独立型层级结构似乎比其他属性层级结构更加"保险",至少不会漏掉可以考察的属性组合模式(题目属性向量),但是当我们的注意力是考察认知诊断测验设计的时候,"大"的属性层级结构对应的可达阵如果作为测验 Q 矩阵的子矩阵时,可达阵的某些列对应的题目就可能无法命制,这一点应该引起足够的重视.

3.2.4 认知诊断测验设计与 CD-CAT 选题策略的区别 尽管认知诊断测验设计可以借鉴 CD-CAT 选题策略,但是 CD-CAT 选题策略和测验设计选择的试题还是有所区别的: CD-CAT 选题策略一般基于被试反应做出的"动作",它是在被试提供了作答信息的基础上而且是针对"特定"的被试(这是自适应的应有之义);而测验设计是测验尚未发生之前的动作,它应该针对"所有"的被试;对于那些不是非参数选题策略,一般的参数化选题策略倾向于选择高质量或猜测和失误相当小的试题,而测验 Q 矩阵设计更多关注的是在项目参数未知情况下如何设计测验 Q 矩阵(在设计 Q 矩阵时也基本上不考虑项目参数的影响),以提高测验的分类准确率.

本文虽然对 Liu Ren[18] 的认知诊断测验设计方案提出不同看法,认为 Q 矩阵设计时应该鼓励命题专家尽可能命制包含最少属性的题目,并且对原先设计的 Q 矩阵进行修正,但是认知诊断处于"婴幼儿时期"[30],发展得不充分,还有许多问题值得讨论,比如 R. Henson 等[31] 提出的认知诊断测验测验编制(组卷)的指标 CDI 的应用,可以动态处理组卷约束问题,而组卷实质上也是测验设计. 如何动态实现认知诊断测验,这是 CD-CAT 研究的重要内容. 又比如,要进行 Q 矩阵设计,就必须明确属性及其层级关系,层级关系的划分,可粗可细,但是要整体把握,层级比较粗可能更容易一些,模拟实验的因素比较多,而层级关系作为其中一个因素,其水平数不能够太大时,层级结构划分比较粗可能更方便;如果仅仅考察层级关系,划分比较细或许考察更加深入.

本文在属性及其层级关系正确的条件下讨论认知诊断测验设计问题,如果这个前提不成立,即属性或者其层级关系不一定准确,那么测验设计需要有一定的容错能力,甘朝红等[32] 涉及这方面的内容,但愿对这个问题引起重视;据笔者所知,项目属性向量平衡的测验设计尚无研究结果,希望能够引起讨论.

本文讨论认知在测验设计,因此不涉及"翻新",'翻新"的做法得出的诊断分析的结果准确性比较低,但是在不得已的情况下,还是可以进行"翻新",这多少可以获得一点诊断信息.

4 参考文献

[1] 丁树良,罗芬,汪文义. Q 矩阵理论的扩展[J]. 心理学探新,2012,32(5):417-422.

[2] Tatsuoka K K. Architecture of knowledge structures and cognitive diagnosis: a statistical pattern classification approach[M]//Nichols P D, Chipman S F, Brennan R L. Cognitively Diagnostic Assessments. Hillsdale, NJ: Erlbaum, 1995: 327-359.

[3] Tatsuoka K K. Cognitive assessment: an introduction to the rule space method[M]. New York: Taylor and Francis Group, 2009.

[4] 丁树良,罗芬. 由偏序关系的可达阵导出 Hasse 图的有效算法:兼谈其在认知诊断中的作用[J]. 江西师范大学学报:自然科学版,2013,37(5):441-444.

[5] Chiu Chia-Yi, Douglas J A, Li Xiaodong. Cluster analysis for cognitive diagnosis: theory and applications[J]. Psychometrika, 2009, 74(4): 633-665.

[6] De Carlo L T. On the analysis of fraction subtraction data: the DINA model, classification, latent class sizes, and the Q-matrix[J]. Applied Psychological Measurement, 2011, 35(1): 8-26.

[7] Madison M J, Bradshaw L P. The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model[J]. Educational and Psychological Measurement, 2015, 75(3): 491-511.

[8] Liu Ren, Huggins-Manley A C, Bradshaw L. The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies[J]. Educational and Psychological Measurement, 2017, 77(2): 220-240.

[9] Gorin J S. Test construction and diagnostic testing[M]// Leighton J P, M J Gierl. Cognitive diagnostic assessment for education: theory and applications. New York: Cambridge University Press, 2007: 173-201.

[10] Ding Shuliang, Wang Wenyi, Luo Fen, et al. Irreplaceability of a reachability matrix[EB/OL]. [2019-02-19]. https://link.springer.com/chapter/10.1007%2F978-3-319-56294-0_21.

[11] Leighton J P, Gierl M J, Hunka S M. The attribute hierarchy method for cognitive assessment: a variation on tatsuoka's rule-space approach[J]. Journal of Educational Measurement, 2004, 41(3): 205-237.

[12] 罗欢，丁树良，汪文义，等. 属性不等权重的多级评分属性层级方法 [J]. 心理学报，2010，42(4)：528-538.
[13] 丁树良，汪文义，杨淑群. 认知诊断测验蓝图的设计 [J]. 心理科学，2011，34(2)：258-265.
[14] 丁树良，杨淑群，汪文义. 可达矩阵在认知诊断测验编制中的重要作用 [J]. 江西师范大学学报: 自然科学版，2010，34(5)：490-495.
[15] 丁树良，汪文义，罗芬，等. 可达阵功能的不可替代性 [J]. 江西师范大学学报: 自然科学版，2016，40(3)：290-294，298.
[16] 彭亚风，罗照盛，喻晓锋，等. 认知诊断评价中测验结构的优化设计 [J]. 心理学报，2016，48(12)：1600-1611.
[17] Liu Ren，Huggins-Manley A C. The specification of attribute structures and its effects on classification accuracy in diagnostic test design [M] //van der Ark L A，Bolt D M，Douglas J A，et al. Quantitative Psychology Research，Springer Proceedings in Mathematics and Statistics. New York: Springer，2016，167: 243-254.
[18] Liu Ren. Misspecification of attribute structure in diagnostic measurement [J]. Educational and Psychological Measurement，2017，78(4)：605-634.
[19] 丁树良，汪文义，罗芬. 多级评分认知诊断测验蓝图的设计: 根树型结构 [J]. 江西师范大学学报: 自然科学版，2014，38(2)：111-118.
[20] 杨淑群，蔡声镇，丁树良，等. 求解简化 Q 矩阵的扩张算法 [J]. 兰州大学学报: 自然科学版，2008，44(3)：87-91，96.
[21] 喻晓锋，丁树良，秦春影，等. 贝叶斯网在认知诊断属性层级结构确定中的应用 [J]. 心理学报，2011，43(3)：338-346.
[22] Gierl M J，Leighton J P，Hunka S M. Using the attribute hierarchy method to make diagnostic inferences about examinees´ cognitive skills [EB/OL]. [2019-02-11]. https://www.researchgate.net/publication/26498309_Using_the_Attribute_Hierarchy_Method_to_Make_Diagnostic_Inferences_about_Examinees´_Cognitive_Skills_in_Algebra_on_the_SATC.
[23] 刘舒畅，涂冬波，蔡 艳，等. 4 种新的基于属性平衡的 CD-CAT 选题策略开发研究 [J]. 心理科学，2018，41(4)：976-981.
[24] Cheng Ying. Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: the modified maximum global discrimination index method [J]. Educational and Psychological Measurement，2010，70(6)：902-913.
[25] Cheng Ying，Chang Huahua. The maximum priority index method for severely constrained item selection in computerized adaptive testing [J]. British Journal of Mathematical and Statistical Psychology，2019，62(2)：369-383.
[26] Cai Yan，Tu Dongbo，Ding Shuliang. Theorems and methods of a complete Q matrix with attribute hierarchies under restricted Q-matrix design [J]. Frontiers in Psychology，2018(9)：1-15.
[27] 丁树良，罗芬，汪文义，等. 知识状态的不同表达及其应用 [J]. 江西师范大学学报: 自然科学版，2017，41(3)：296-301.
[28] 汪文义，丁树良，宋丽红. 认知诊断中基于条件期望的距离判别方法 [J]. 心理学报，2015，47(12)：1499-1510.
[29] Wang Wenyi，Song Lihong，Chen Ping，et al. Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment [J]. Journal of Educational Measurement，2015，54(4)：457-476.
[30] Leighton J P，Gierl M J. Why cognitive diagnostic assessment? [M] //Leighton J P，Gierl M J. Cognitive diagnostic assessment for education: theory and applications. Cambridge，UK: Cambridge University Press，2007: 3-18.
[31] Henson R，Douglas J. Test construction for cognitive diagnosis [J]. Applied Psychological Measurement，2005，29(4)：262-277.
[32] 甘朝红，汪文义，丁树良. 项目属性标错时可达阵补救作用的研究 [J]. 江西师范大学学报: 自然科学版，2014，38(6)：600-604.

The Designing Cognitive Diagnostic Test with Dichotomous Scoring

DING Shuliang，LUO Fen，WANG Wenyi，XIONG Jianhua
(College of Computer Information Engineering，Jiangxi Normal University，Nanchang Jiangxi 330022，China)

Abstract: The essence of designing cognitive diagnostic test is how to design a test Q-matrix to attain the measurement goal. The design depends on the relationship on attribute hierarchy and if the test Q matrix coincides with the structure of all possible knowledge states，the accuracy and construct validity of the test result may be high. The balance of item attribute vectors is more effect than the balance of attributes in the test design when attribute hierarchy is dependent type. Some queries about the design proposed by Liu Ren(2017) are raised.

Key words: design of cognitive diagnostic test; design of Q matrix; attribute hierarchy; blance of item attribute vectors

(责任编辑: 冉小晓)