

教育认知诊断测验与认知模型一致性的评估^{*}

丁树良¹ 毛萌萌² 汪文义¹ 罗 芬¹ CUI Ying³

(¹ 江西师范大学计算机信息工程学院, 南昌 330022) (² 南昌大学教育学院教育系, 南昌 330031)

(³ Centre for Research in Applied Measurement and Evaluation (CRAME), University of Alberta, CA)

摘 要 构建正确的认知模型是成功进行认知诊断的关键之一, 如果认知诊断测验不能完整准确地代表这个认知模型, 这个测验的效度就存在问题。属性及其层级可以表示一个认知模型。在认知模型正确基础上, 给出了一个计量公式以衡量认知诊断测验能够多大程度上代表认知模型; 对于不止包含一个知识状态的等价类及其形成原因进行了分析, 对 Cui 等人的属性层级相合性指标(HCI)提出修改建议, 以更好地探查数据与专家给出的认知模型的一致性。

关键词 认知诊断; 认知模型; 认知诊断测验

分类号 B841

教育认知诊断中涉及到两类模型: 认知诊断模型(cognitive diagnostic model, CDM)和认知模型(cognitive model, CM)。认知诊断模型受到高度关注, 它是被试在项目上的反应的函数, 将被试可观察的反应模式(ORP)转化为被试知识状态(KS)的数学模型。大家比较熟悉的规则空间模型(RSM; Tatsuoaka, 1995, 2009)、属性层级模型(AHM; Leighton, Gierl, & Hunka, 2004)、确定性输入, 噪音“与”门模型(DINA; Junker & Sijtsma, 2001)等等, 都是 CDM。对 CDM 的研究包括对 CDM 的开发、CDM 的性质的探讨以及 CDM 的应用等等。而 CDM 又可依照不同标准进行分类, 如果从项目的评分方式来看, 可以分成 0-1 评分、多级评分、甚至连续评分的 CDM; 从认知模型的维度来看, 有单维和多维模型; 从认知属性对认知任务起的作用来看, 有补偿性、部分补偿性和非补偿性模型; 从认知策略的多寡来看, 有单策略与多策略模型; 从 Q 阵的完备性来看, 有要求 Q 阵完备和不要求 Q 阵完备模型等等 (Fu & Li, 2007; 汪文义, 2009; 陈平, 2011)。非补偿的含义是掌握的属性不能补偿没有掌握的属性, 非

补偿模型包括连接模型(conjunctive models)和非连接模型(disjunctive models)。连接的含义是掌握项目所有属性, 可以较高的概率正确作答, 而未掌握(即使只有一个属性没有掌握)也将大大降低答对概率。本文仅讨论连接的非补偿模型。

认知模型因为和具体的认知行为有关, 故更加多姿多彩。本文仅讨论教育测量中的认知模型。给定一个感兴趣的论域(domain), 教育测量中的认知模型是对问题解决的抽象, 选择合适的粒度(granularity)刻画任务, 即用合适的粒度(粒度的直观解释是考察点的细分或细化程度)对问题解决的简化描述, 以方便解释和预报被试表现, 包括他们在这个论域中认知上的长处和不足(Gierl, Leighton, & Hunka, 2007)。比如诊断小学生分数运算掌握情况, 就不能使用“分数运算”这么大的粒度对问题进行描述, 而要将这个属性进行分解, 用更小的粒度(比如相同分母的加减运算、求两个整数的最小公倍数、最大公因子、通分、异分母加减运算、约分, 等等)进行描述(许志勇, 丁树良, 汪文义, 涂冬波, 2010)。

认知模型提供了对测验反应进行解释的框架,

收稿日期: 2011-10-10

^{*} 国家自然科学基金(30860084, 31160203, 31100756)、国家社会科学基金(12BYY055)、国家教育科学规划项目(CCA110109)、教育部人文社科项目(09JJCX012, 10YJCX049, 11YJC190002)、江西省教育厅科技计划项目(GJJ11385, GJJ10238)、全国教育考试科研规划课题(2009JKS2009)和高等学校博士学科点专项科研基金(20113604110001)资助。

通讯作者: 丁树良, E-mail: ding06026@163.com

这个框架还可以用来指导项目开发,将测验观察结果和被试的知识、加工、策略等特殊认知推断联系起来,使认知原理和测量实践联系起来(Gierl et al., 2007)。

认知模型是认知诊断中最基础也是最关键的部分。如果没有认知模型,就好像进入一个大城市缺少一张导游图,只能在纵横交错的道路上摸索,而不可能对城市的布局了然于胸。同样,没有认知模型,认知诊断测验的设计就缺乏基础,也就很有可能无法探查某些被试的知识结构(这些知识结构就相应于地图中的一些道路)。之所以说规则空间模型对现存的测验数据进行认知诊断分析,即使用“翻新”(retrofitting)方法,会存在这样或那样的缺陷,就是因为现存的测验根本不是依据认知模型去设计的。当然,由于认知诊断测验编制的原理和方法的研究成果还不丰富,纵使有了认知模型,如何编制一个能够完整准确地反映这个认知模型的测验蓝图的问题,也还值得认真研究。

构建认知模型主要有专家讨论、文献调查、口语报告等方法。认知模型的构建是一个十分艰巨的工作,所以目前能直接应用到教育认知诊断的认知模型还很少(Gierl et al., 2007)。认知模型的构建也不是一蹴而就的,比如下文中例 5 就给出了对认知模型进行争论的例子。这些争论提供了一些竞争的认知模型,如何评价各自的优劣是一件很有意义的工作。

本文对教育诊断测验与认知模型一致性的评估进行讨论。首先将认知模型和属性及其层级(attributes and their hierarchy, A&H)对应,再将 A&H 用数学符号表示,给出充分必要 Q 矩阵概念及其性质;其次讨论 A&H 和观察反应模式(ORP)与期望反应模式(ERP)的关系,由此引出认知诊断测验的理论构想效度(theoretical construct validity)的概念,并且用两个例子分析、说明理论构想效度是如何受到影响的;接着考虑到不论先构建认知模型再编制测验的方式还是先开发测验,再对所得反应数据做认知诊断分析的“翻新”方式,都有必要考查测验数据是否能代表(符合)认知模型的问题:如果得到 Q 矩阵,将现有的层级相合性指标(HCI; Cui & Leighton, 2009)进行一点修正,获得修正的 HCI (MHCI),希望用 HCI (MHCI)考察观察反应模式与认知模型吻合的好坏;最后一节是总结与讨论。

1 认知模型与属性层级的对应

概括地讲,属性(attribute)是指正确求解项目

所需的基本认知加工(cognitive processes)或技能(skill)。属性层级(hierarchy)是属性之间的逻辑和(或)心理特征结构,不能与项目解决中要求的属性顺序关系相互混淆(Gierl, Leighton, & Hunka, 2000)。若不掌握属性 A 便不能掌握属性 B,则称 A 为 B 的先决(prerequisite)属性。比如,不掌握同分母加法(A)和求两个数的公倍数(B),就不能正确掌握异分母加法(C),于是 A 和 B 都是 C 的先决属性,而 A 和 B 之间没有先决关系。如果 A_1 是 A_2 的先决属性,并且不存在另外一个属性 D,使得 A_1 是 D 的先决属性而 D 又是 A_2 的先决属性,则称 A_1 是 A_2 的直接(immediate)先决属性,或者用偏序关系的语言来说, A_2 是 A_1 的覆盖(屈婉玲,耿素云,张立昂,2008)。

有专家认为对于所感兴趣的论域,属性及其层级可以构成认知模型(Gierl et al., 2007; Leighton et al., 2004; Tatsuoka, 1995, 2009)。Leighton 等人(2004)给出四种基本属性层级、线型、收敛型、发散型及无结构型(图 1),并且认为其他复杂层级关系可以由这四种基本层级结构组合而成。从图 1 可知 Leighton 等人(2004)的无结构型是一种特殊的发散型,他们这种分类显得有一点混乱;而 Tatsuoka (1995, 2009)给出一种独立型结构(图 1),即所有属性之间无先决关系。这种独立型结构,在实际应用中也是存在的,如有研究认为 GRE 牵涉的属性之间的关系符合独立型结构(可参见 Leighton et al., 2004)。在国外研究中经常使用这种独立型结构,由于 Leighton 等人强调属性之间的先决关系,而没有引入这种属性之间无先决关系的结构类型。Leighton 等人所给出的结构都是连通图,独立型是非连通图(屈婉玲等,2008),为了研究范围更加广泛而不冗余,我们主张用独立型代替无结构型。

给定 K 个属性及其层级,根据图论(graph theory),便可以给出一系列元素为 0 或 1 的矩阵,比如 K 阶邻接矩阵(adjacency matrix) $A = (a_{ij})$, $a_{ij} = 1$ 表示 i 是 j 的直接先决属性;否则 $a_{ij} = 0$,当 $i = j$ 时 $a_{ii} = 0$;由邻接矩阵 A 和 K 阶单位矩阵 I 的和 $A + I$ 可以获得可达矩阵 $R = (r_{ij})$; R 的第 i 行等于 1 的元素表示属性 i 是这些非零元素对应的属性的先决属性; R 第 j 列中非零元素表示这些非零元素对应的属性和第 j 个属性在同一“路径”上;故 R 第 j 列表示的是以第 j 个属性为终点的路径(杨淑群,丁树良,2011)。

由可达矩阵 R 通过 Tatsuoka (1995, 2009)的方法可以导出简化 Q 阵 Q_r ,或者通过扩张算法(Ding,

Luo, Cai, Lin, & Wang, 2008; 杨淑群, 蔡声镇, 丁树良, 林海菁, 丁秋林, 2008; 丁树良, 汪文义, 杨淑群, 2009; 丁树良, 祝玉芳, 林海菁, 蔡艳, 2009) 导出 Q_r 。我们将 Q_r 改写成 Q_p , Q_p 的列的集合表示所有潜在(potential)项目类的集合, 即 Q_p 中第 j 列代表一类项目, 这类项目所含属性恰好对应了 Q_p 第 j 列中非零元素; 在 Q_p 中加上一个全 0 列, 记为 Q_s , 表示所有被试的知识状态类。显然由 R 可以导出 Q_p , 由 Q_p 可以导出 Q_s ; 同样由 Q_s 立即可得 Q_p 。而由 Q_p 通过行的包含关系的比较(Tatsuoka, 1995,

2009), 或者扩张算法的逆算法(缩减算法)可以得到可达矩阵 R 。而使用“缩减算法”时, 要使用“布尔并”运算。所谓两个 0-1 列向量的“布尔并”运算, 即两个向量的对应分量中取最大元, 亦即若 x, y 为取值为 0、1 的元素, 则 x 与 y 的布尔并记为 $x \vee y = \max(x, y)$ 。而缩减法的要义是如果 Q 矩阵中一个列可以表示成为这个 Q 矩阵中其他两个或者两个以上列的布尔并, 那么这一列是这个 Q 矩阵中“非本质”的列, 可以删除, 不断地删除 Q 中“非本质”的列, 一直到所有“非本质”的列都删除为止。缩减算法放在附录中。

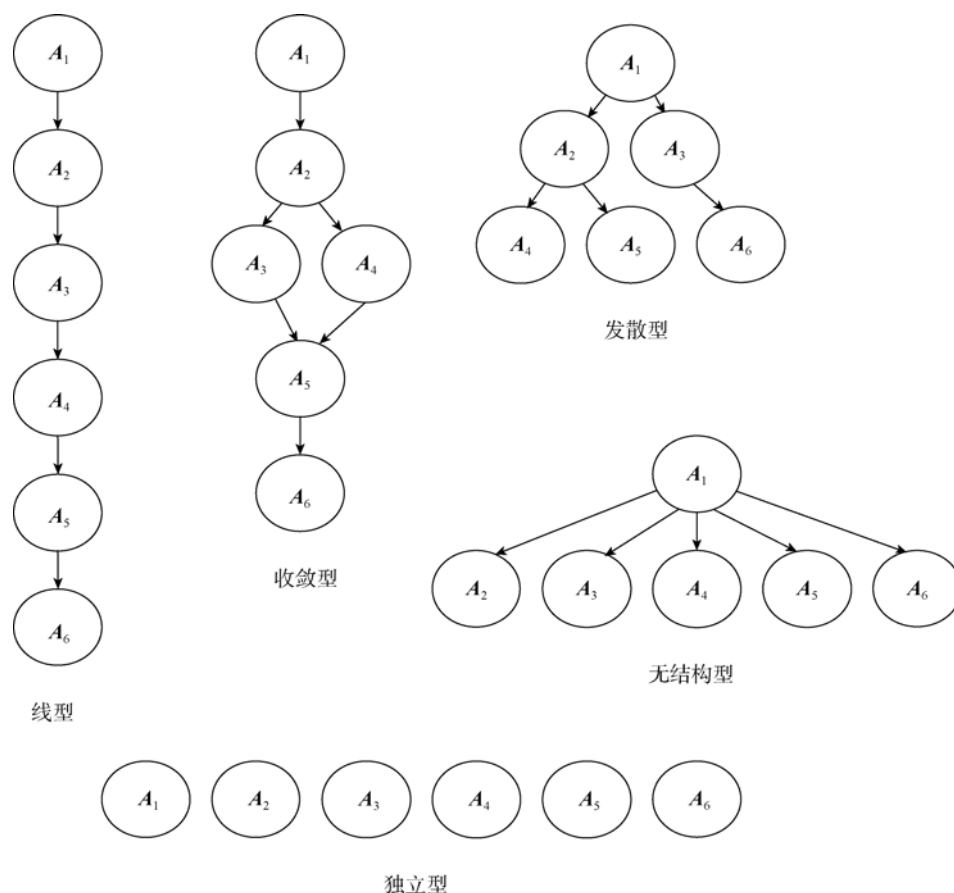


图 1 6 个属性对应的 5 种基本属性结构

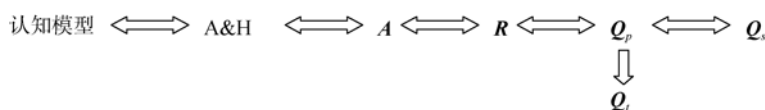


图 2 认知模型和几类矩阵之间的关系

图 2 中的矩阵 Q_i 表示测验蓝图, 即测验 Q 阵。与 Q_p 一样, 它的行表示属性, 列表示项目类, 同一类项目中所包含的属性相同, 由 Q_i 对应的列中的非零元素表示。 Q_i 是 Q_p 的子矩阵(sub-matrix), 即

Q_i 中每一列均取自 Q_p 。当然实际使用的测验蓝图中的列可以重复(重复测量只可提高信度, 而不影响效度), 而我们这里约定 Q_i 中相同的列仅仅保留一列。在图 2 中, 除了 Q_p 到 Q_i 之间用单向箭头表

示之外,其他均用双向箭头标识,表示箭头两边的对象是“等价”的,即相互可以推出,例如认知模型 $\longleftrightarrow A\&H$, 这是我们的基本假定;而 $A\&H \longleftrightarrow A$, 则由图论知识可得;其他也可以类推;只是可达矩阵 R 和邻接矩阵 A 的相互推导,要略加如下说明。可达矩阵 R 是偏序关系对应的一个关系矩阵。可达矩阵 R 通过“清洗算法”可以导出邻接矩阵 A (丁树良, 罗芬, 2005)。

由 Q_p 可以给出 Q_i , 而 Q_i 不一定可以给出 Q_p 。如果这两者可以互推,即由 Q_i 可以推出 Q_p , 从而可以推出认知模型,这样的测验蓝图在认知诊断中就显得十分重要。然而在什么条件下,由 Q_i 可以推导出 Q_p ?

其实 Tatsuoka (1995, 2009) 曾经也想解决这个问题,她的办法是给出充分 Q 阵(sufficient Q matrix)的概念。

设 Q_i 为 K 行 m 列 Q 矩阵,按列剖分成 $(q_1, q_2, \dots, q_m)$, 其中 $q_j = (q_{1j}, q_{2j}, \dots, q_{Kj})'$ 。如果 $q_i - q_j \geq 0$, 即 $q_i - q_j$ 的每个分量都不小于零,则表示项目 j 中的所有属性均包含在项目 i 之中,也可以称项目 i 是项目 j 的父项目, j 是 i 的子项目,又称项目 i 包含项目 j 。相仿,将 Q_i 按行划分, Q_i 的第 i 行记为 s_i , $s_i = (q_{i1}, q_{i2}, \dots, q_{im})$, $i = 1, 2, \dots, K$ 。如果 $s_i - s_j \geq 0$, 则称属性 i 是属性 j 的先决属性。这种 Q_i 中任意两行相减以寻找属性之间先决关系的方法,我们称其为 Q_i 中包含关系基础上的逐对比较方法,不致混淆时,简称为 Q_i 行的逐对比较方法。实际上 Tatsuoka (1995, 2009) 介绍过这种 Q_i 行逐对比较方法,她使用的是集合包含的运算,我们这里是用向量减法运算对其加以表达而已,这两种方法运算结果是等价的。

定义 1(充分 Q 阵, Tatsuoka, 1995, 2009) 假设 Q_i 的行表示属性,列表示项目,如果通过使用 Q_i 行的逐对比较方法以后,可以导出可达矩阵 R , 则称 Q_i 对所感兴趣领域的认知模型的表达是充分的(sufficient)。

例 1 设 $K=m=3$, 三个属性是独立结构, 则可

达矩阵 R 为三阶单位阵 I_3 。设 $Q_i = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$, 则由

Q_i 的行在包含关系的基础上进行逐对比较, 可得属性之间没有先决关系, 于是由 Q_i 可以导出可达阵为三阶单位阵 I_3 , 即 Q_i 是充分 Q 阵。

正如 Tatsuoka (1995, 2009) 所说的, 纵使 Q_i 是

充分的, 但它并不一定是合适的(appropriate) Q 阵, 因为充分 Q 阵仍可能使许多观察反应模式无法分类, 更不能保证对所有观察反应模式正确分类。事实上, 充分 Q 阵不能保证不同知识状态对应于不同的期望反应模式(ERP)。这里所说的 ERP 是指给出若干个项目, 被试若掌握了某个项目中所有的属性, 则必答对该项目(不失误), 否则必做出错误反应(无猜测), 这样导出的反应称为对该项目的理想反应, 对所有项目的理想反应称为理想反应模式(ERP)。对于规则空间模型, 每个 ERP 对应一个纯规则点(Tatsuoka, 1995, 2009), 而下例说明有的充分 Q 阵可能会将几个不同的知识状态对应于同一个纯规则点。

例 2 对于知识状态为(0 0 0)、(1 0 0)、(0 1 0)、(0 0 1)的四类被试, 在例 1 所示的 Q_i 下, ERP 都是(0 0 0), 而(1 1 0)、(0 1 1)、(1 0 1)、(1 1 1), 对应的 ERP 分别是(1 0 0)、(0 0 1)、(0 1 0)、(1 1 1)。于是在例 1 的 Q_i 之下, 由 8 类不同的知识状态仅能给出 5 类不同的 ERP, 也就只能给出 5 个纯规则点。可见充分 Q 阵不一定能提高规则空间模型(RSM)的分类准确率, 甚至使用充分 Q 阵还可能使一些被试不能归类或者错误归类。

定义 2 (充分必要 Q 阵, 简称为充要 Q 阵) (丁树良, 杨淑群, 汪文义, 2010; 丁树良, 汪文义, 杨淑群, 2011) 若可达矩阵 R 是测验蓝图 Q_i 的子矩阵, 则 Q_i 阵对所感兴趣的领域的认知模型的表达是充分必要的。

定义 2 中的子矩阵的概念是指, 如果矩阵 M_1 的所有列都是矩阵 M 的列, 则称 M_1 是 M 的子矩阵。这个定义的意思是若 Q_i 为充分必要 Q 阵, 则由认知模型可以导出 Q_i 阵; 反过来这个 Q_i 阵也可以导出认知模型。这一点可以由图 2 看出, 可达矩阵和认知模型是双方可以相互推导的, 而 Q_i 阵为充分必要 Q 阵, R 则是 Q_i 的一部分, 这当然就可以导出认知模型, 也就可以导出潜在 Q 阵 Q_p 。事实上我们有如下的定理:

定理 1 假设所讨论的认知属性对认知任务所起的作用是非补偿, 连接的。在 0-1 评分模式下, Q_i 是充分必要 Q 阵当且仅当知识状态集合与期望反应模式集合可以建立起一一对应关系(双射关系)(丁树良等, 2010, 2011)。

从定理 1 可知, Q_i 是充要 Q 阵, 则每个 KS 等价类中有且只有一个元素(知识状态)。

显然, 充要 Q 矩阵和 Tatsuoka (1995, 2009) 所

定义的充分 Q 矩阵是不同的。事实上, 充要 Q 矩阵一定是充分 Q 矩阵, 反之则不然, 如例 1 中 Q_t 是充分 Q 矩阵, 但不是充要 Q 矩阵。

例 3 在例 1 条件下, 设 $Q_t = I_3$, 则 Q_t 为充要 Q 矩阵, 对所讨论领域中任一个知识状态 α , 在测验蓝图 Q_t 下, 其期望反应模式仍然为 α 。这里知识状态集合中有 2^3 个元素, 而期望反应集合中也恰有 8 个元素, 这两个集合之间是一一对应的。

2 A&H 与 ORP, ERP 之间的关系

已知属性及其层级 A&H, 则可确定 Q_p 、 Q_s , 如果再给出测验蓝图 Q_t , 被试参加测验以后, 便可以得到观察反应模式(ORP), 对于非补偿, 连接模型可以计算出期望反应模式(ERP) (丁树良等, 2009b), 通常的认知诊断是通过 ORP, ERP 建立起 KS 的对应, 从而达到诊断目的。对 RSM 和 AHM 的分类原理进行分析, 便可以清楚地看到这一点; 对于 DINA 等模型, 项目反应函数中观察反应、理想反应和知识状态的关系比 RSM 和 AHM 相应的关系稍复杂, 仍是通过建立三者的联系进行诊断的。

我们用图 3、图 4 来表示 A&H, ORP, ERP 与 KS 之间的关系。

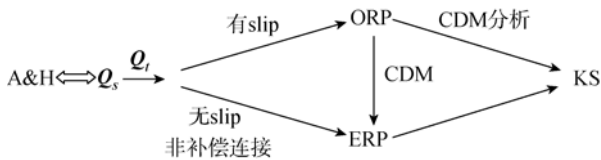


图 3 A&H 与知识状态之间的关系

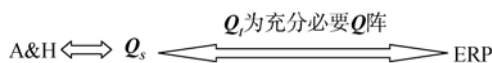


图 4 充分必要 Q 阵的作用

2.1 理论构想效度

将专家给出的 A&H 看作是理论上的 A&H。如果通过 Q_t 中行的逐对比较, 可以挖掘出 Q_t 对应的 A&H (记为 $A\&H(Q_t)$), 理论上的 A&H 与 $A\&H(Q_t)$ 的吻合程度可以度量 Q_t 的完备程度, 即 Q_t 有多大的资格代替理论上的 A&H? 我们给出理论构想效度(theoretic construct validity, TCV)的概念。

定义 3 (理论构想效度) 设专家给出的 A&H 对应 N_1 个知识状态 $\{\alpha_1, \alpha_2, \dots, \alpha_{N_1}\}$; 而测验蓝图 Q_t 可以导出 N_2 个知识状态 $\{\beta_1, \beta_2, \dots, \beta_{N_2}\}$, 又设 $\{\beta_1, \beta_2, \dots, \beta_{N_2}\} \cap \{\alpha_1, \alpha_2, \dots, \alpha_{N_1}\}$ 中含有 N_3 个元素,

则 Q_t 对应的理论构想效度

$$TCV = \begin{cases} \frac{N_3 + 1}{N_1}; & \text{若 } \{\beta_1, \beta_2, \dots, \beta_{N_2}\} \text{ 不包含零向量} \\ \frac{N_3}{N_1}; & \text{否则} \end{cases}$$

例 4 如例 3 中, 定义 3 中 $N_1=8$ (独立结构 Q_s 的列数), $N_3=8$, 这可形式上由 Q_t 通过扩张算法导出的矩阵列数加上 1 (加上全 0 列) 得到, 故例 3 中 Q_t 的 $TCV=8/8=1$ 。而例 1 中 3 个独立属性, A&H 可导出 8 个知识状态, 但

$$Q_t = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \xrightarrow{\text{扩张算法}} \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \xrightarrow{\text{加上全 0 列}} \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}, \text{ 它有 5 列, 故例 1 中 } Q_t \text{ 的 } TCV=5/8.$$

有 5 列, 故例 1 中 Q_t 的 $TCV=5/8$ 。

例 5 (引自 Sinharay & Almond, 2007) Tatsuoka, Linn, Tatsuoka 和 Yamamoto (1988) 的带分数减法认知诊断测验。理论上有 5 个属性, A_1 : 基本分数减法, A_2 : 分数或带分数化简, A_3 : 从分数中分离出整数, A_4 : 从带分数整数部分借 1, A_5 : 化整数为分数。专家认定的属性层级结构如图 5a, 测验 Q_t 中通过包含关系找出的层级结构如图 5b。图 5b 是依据下述方法导出的: 测验蓝图 Q_t 阵中相同的列只保留 1 列以后, 仍然记为 Q_t , 再依照 Q_t 行的逐对比较找出各属性之间的先决关系后整理而成图 5b。

图 5a 是理论上的层级结构, 可以导出 24 种知识状态, 由测验蓝图 Q_t 导出的图 5b, 却只能给出 9 种知识状态, 所以 Q_t 给出的 $TCV=9/24=0.375$, 这个 TCV 并不令人满意。而这又表明, 可能有若干个不同的知识状态都对应同一个理想反应模式, 如果使用 Tatsuoka (1995, 2009) “知识状态等价类”的术语来表述, 即某些 KS 等价类里含有多多个知识状态。这是知识状态集合 S 上由 Q_t 决定的等价类。给定 Q_t , 若 $\alpha, \beta \in S$, α 和 β 有相同的 ERP, 则称 α 和 β 等价, $\alpha \equiv \beta$ 。可以证明 (请见附录) S 上关系 $\equiv$ 是自反, 对称, 传递的, 从而是 S 上等价关系 (可以参见 屈婉玲等, 2008)。

下面给出理论构想效度的另一个定义, 它可能比定义 3 更广泛。

定义 4 (理论构想效度). 在定义 3 中, $\{\beta_1, \beta_2, \dots, \beta_{N_2}\} \cap \{\alpha_1, \alpha_2, \dots, \alpha_{N_1}\}$ 中所含 N_3 个元素记为 $\{\gamma_1, \gamma_2, \dots, \gamma_{N_3}\} \subseteq \{\alpha_1, \alpha_2, \dots, \alpha_{N_1}\}$, 若可以计算出每

$$B_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix},$$

如何考察 $Q_p^{(1)}$ 中哪一些知识状态对应同一个期望反应模式, 即构成同一个知识状态等价类? 能够用其它方法指出 Q_t 的缺陷吗? 第一个问题的回答可以应用丁树良等人(2009b)介绍的求期望反应的方法, 根据 Q_t 和 $Q_p^{(1)}$ 便可以计算出来, 其中 B_1 的 11 个列和零向量对应的期望反应模式相同, 所有分量都是 0; 而 B_2 中知识状态在 Q_t 下产生的期望反应模式分别和 $(1 \ 1 \ 0 \ 0 \ 0)^T$ 、 $(1 \ 0 \ 1 \ 0 \ 0)^T$ 、 $(1 \ 1 \ 1 \ 0 \ 0)^T$ 以及 $(1 \ 0 \ 0 \ 0 \ 0)^T$ 的 ERP 对应相等, 即这些知识状态等价类都含 2 个知识状态。

第二个问题除掉对照层级关系图, 写出可达矩阵, 直接对可达矩阵进行比对之外, 还可以使用“缩减算法”解决, 对 Q_{lost} 使用“缩减算法”, 得到

$$M = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

这表明 Q_t 的缺陷是缺少了 Q_{lost} 中本质的 5 列。当然由于 Q_t 本身的构造, M 中的前面 4 列才是 Q_t 本身缺少的, M 中的第 5 列是由 R_1 和 Q_t 中的相同列(第 1 列)和 R_1 中的第 5 列作布尔并得到的。而这缺少的 4 列正是图 5a 对应的可达矩阵 R_1 中的 2-5 列, 这进一步说明测验蓝图 Q 矩阵包含可达矩阵的所有列才可以使得期望反应模式和知识状态一一对应(丁树良等, 2010; 丁树良等, 2011)。

2.2.2 理论属性层级未知的情形 以上是对理论上的属性层级(即认知模型)已知条件下的分析, 然而目前往往有学者对已经存在的并不是为诊断而开发的测验进行认知诊断, 或者虽然是为诊断目的开发的测验, 但事实证明测验中项目属性标定不一定正确或者测验设计不一定合理; 这时如何去考察测验的理论构想效度并分析该测验导出的知识状态等价类?

结合 DeCarlo (2011)考察过的例子进行分析。DeCarlo 对 Tatsuoaka (1990)的分数减法数据重新分析。这个诊断测验包含 8 个属性, 20 个项目, 其中有 5 对项目含有相同的属性, 它们是项目 2 与 3、4

与 20、6 与 8、11 与 17、14 与 16, 具有相同属性的项目仅保留 1 个, 则这个测验可以用一个 8×15 的测验蓝图(Q_t 阵)来表示:

$$Q_t = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

根据 Q_t 行的逐对比较方法寻找属性之间先决关系, 知属性 A_7 是属性 A_1 、 A_4 、 A_6 、 A_8 、 A_5 的直接先决, 而 A_2 是 A_5 的直接先决, A_5 是 A_3 的直接先决, 由此可得到这 8 个属性的层级关系图 6, 由图 6 可导出如图 7 所示的可达矩阵 R 。

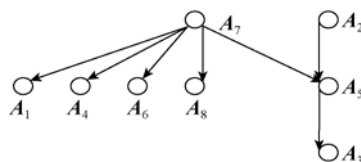


图 6 由 Q_t 挖掘出的层级关系

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

图 7 图 6 对应可达矩阵 R

对 Q_t “形式上”使用扩张算法, 可以导出除零向量外的 57 种知识状态, 而由 R 通过扩张算法可以导出除零向量外的 65 种知识状态。纵使 Q_t 挖掘出来的层级关系为真, 这时 Q_t 的理论构想效度也只有 $58/66=0.879$ 。这表示有 $66-58=8$ 种知识状态在给定的 Q_t 下的期望反应模式会和其他的知识状态产生的期望反应模式重合, 即有一些知识状态等价类中包含的元素不止 1 个。这里效度降低的原因在于 Q_t 中不可能通过布尔并运算获得可达矩阵 R 的第 6 列, 它包含且只包含属性 A_6 、 A_7 。由于缺乏这 1 列, 以及 A_6 、 A_7 这两个属性还可以和属性 A_1 、 A_2 、 A_4 、

A_8 中任一个搭配,考虑到 Q_i 中的第1列已经包含了 A_4 、 A_6 、 A_7 ,知属性 A_6 、 A_7 还可以从集合 $\{A_1$ 、 A_2 、 $A_8\}$ 中分别取0个、1个、2个、3个元素搭配组成新的项目类,于是缺少这8个项目类。由 Q_i 中的第3列和倒数第2列,知道不必再考虑 A_6 、 A_7 、 A_2 这3个属性分别和 A_5 、 A_3 搭配产生的项目类。这说明如果命题以后,项目属性均已标识,命题专家和测量专家仍需要对测验的“完整性”(即测验 Q 阵是否包含了 Q 挖掘出的可达矩阵 R)进行考察,如果 Q 阵缺乏 R 阵中的列,则应该修改测验,使之包含可达矩阵 R 中所有的列对应的项目。我们可以将这一过程小结为四步:

第一步:由测验 Q 阵(Q_i)挖掘出层级关系,导出可达矩阵 R ;

第二步:对 Q_i 形式上使用扩张算法,扩张出所有的知识状态 Q_s ;

第三步:由 R 扩张出所有知识状态 $Q_s(R)$;

第四步:比较 Q_s 和 $Q_s(R)$;如果 $Q_s(R)$ 的列数多于 Q_s 的列数,则应修改 Q_i ,使之包含 R 中所有列。

另外,还可以借用前文提到的“缩减算法”来完成,具体步骤如下:

第一步:由测验 Q 阵(Q_i)挖掘出层级关系,导出可达矩阵 R ;

第二步:对 Q_i 使用“缩减算法”,得到 $R(Q_i)$;

第三步:比较 R 和 $R(Q_i)$;如果两个矩阵 R 与 $R(Q_i)$ 不相等,则应修改 Q_i ,使之包含 R 中所有列。

2.3 ORP 与 ERP 不符合的其他原因初探

以上是专家认定的A&H与被试的知识结构基本一致的假定之下,讨论测验蓝图如何表达A&H,以及表达得是否充分的度量办法。然而专家认定的A&H与被试的知识结构并不一定是一致的,事实上一组专家或几组专家之间的认识都不能保证一致,比如Cheng (2009, p.627)就讲到对同一个测验不同专家组给出了两个不同 Q_i 阵。这一点McGlohen (2004)博士论文也提到过;更有说服力的例子是DeCarlo (2011)称有五批专家分析过Tatsuoka等人(1990)分数减法的数据,有些专家(比如de la Torre & Douglas, 2004)就修正了Tatsuoka (1990)给出的 Q 阵。既然专家之间对 Q 阵的认定都有分歧,谁又能保证被试所采用的策略与 Q 阵规定的一致?

被试与专家的这些不一致表现在观察反应模式(与 Q_i 有关)和理论上的A&H不太吻合。更详细一点说,如果被试的ORP中既无猜测也无失误,则

ORP的集合可能只是ERP的真子集,即两者可能不一致。前文已经分析了这两者不一致的原因之一,即 Q_i 不是充要 Q 矩阵;另外一个原因是可能 Q 矩阵标定有误而引起系统误差。Rupp和Templin (2008)研究过这个问题,他们认为, Q_i 中如果某个列(表示一类项目)中本应为0的元素却误指为1,则该项目猜测参数变大;比如一个项目仅含属性 A_1 ,却在 A_2 处也标记为1,本来不掌握属性 A_2 而只是掌握属性 A_1 的被试就应该答对该项目,但由于错误标定,从而认为这些被试是“猜对”该项目,使该项目的猜测参数变大。反之本应为1的元素却误指为0,则该项目失误参数变大;比如一个项目本应含属性 A_1 和 A_2 ,却在 A_2 处标记为0,则本来仅掌握属性 A_1 的被试答不对该项目是正常的,现在却认为是失误造成的,从而使该项目的失误参数变大;但是还有可能出现某些ORP不是ERP集合中的元素,比如3个属性的线性型结构, A_1 是 A_2 的先决属性, A_2 是 A_3 的先决属性,但是出现(101)这样不符合层级关系的ORP。这种不一致,不是因为 Q_i 不是充分必要 Q 矩阵引起的。这种不一致可能是由于失误引起的,也可能是猜测引起。失误和猜测是一种随机误差。

本文关注的不仅仅是由于测验蓝图(Q_i)的误指等原因引起的诊断结果的系统误差,而且还关心由于被试本身的失误(slip)引起的随机误差。也就是说图3中有一个问题,即如果ORP存在严重干扰,即slip过大,则诊断准确率必定下降,如何判断干扰的严重程度?可以直接由ORP对认知模型进行评估吗?

3 由ORP估出KS可信吗?

如果作答时不含失误或猜测,则ORP即是ERP。而如果再满足 Q_i 是充要 Q 阵的条件,则ERP与KS一一对应,这时由ORP估出KS是简单的,也是可信的;然而被试的ORP或多或少存在失误或猜测,ORP和ERP的对应就存在误差。

如果 Q_i 不是充要 Q 阵,则根据ORP估计出来的KS的集合能否真正代表被试群体也可能难以置信。比如例1中KS分别为(0 0 0)、(1 0 0)、(0 1 0)、(0 0 1)的被试,在给定的测验蓝图之下,所得的ORP都可能是(0 0 0)。而ORP为(0 0 0),则估出来的KS也应该是(0 0 0),这当然造成了估计误差,这种估计误差是由于设计上的缺陷造成的,下文我们假定专家给出A&H,在此基础上给出的 Q_i 也是充

分必要 Q 阵。但可能专家给出的认知模型不能代表被试的认知模型(比如被试采用的策略与专家在 Q 阵中规定的策略不相同), 这会造成 ORP 与 ERP 的系统误差; 而被试的 ORP 中必定掺杂了随机误差, 如果这些误差比较大, 估计出的 KS 也就不一定准确, 所以下面讨论如何描述和度量 ORP 偏离期望反应的程度。由于假定 Q_i 为充要 Q 阵, 故由本文定理 1 知 ORP 与 ERP 的偏离程度, 也就是 ORP 与 A&H 的偏离程度。

3.1 层级相合指标

如果有了 Q 矩阵, Cui 和 Leighton (2009) 给出了层级相合性指标(HCI)以度量每个被试的观察反应与期望反应接近的程度, 更重要的是有人认为个人拟合指数的均值可以作为评价认知模型和数据拟合程度的指标(Wang & Gierl, 2007)。其原理是: 设采用 0-1 评分且 j, g 为项目, 假定项目 j, g 中所测量的属性的集合分别为 AS_j 和 AS_g , 且 AS_g 包含在 AS_j 中, 则称项目 g 为项目 j 的子项目, j 为 g 的父项目; 那么在不失误的条件下, 如果被试 i 能对项目 j 正确反应, 则必可对项目 g 正确反应。如果用 MS_i 表示被试 i 在其正确作答的项目的所有子项目上错误作答(失误)的次数之和, 则此时 $MS_i = \sum_{j \in SC_i} \sum_{g \in S_j} x_{ij}(1-x_{ig}) = 0$, x_{ij}, x_{ig} 为被试 i 在项目 j, g 上的得分。式中 $SC_i = \{j|j \text{ 为项目, 且 } x_{ij}=1\}$, 它表示被试 i 正确反应的项目集合; $S_j = \{g|g \text{ 为项目, } AS_g \subseteq AS_j, g \neq j\}$ 即项目 j 的子项目集合。用 $\#S_j$ 表示集合 S_j 中元素个数, $N_{ci} = \sum_{j: x_{ij}=1} \#S_j$, 即 N_{ci} 表示被试 i 其答对项目上所有子项目的个数之和, 称为所有答对项目的比较数。在不失误的条件下, $MS_i = 0$, 而对于许多实际作答, $MS_i \neq 0$, 即 ORP 与 ERP 会出现不一致的情况, MS_i / N_{ci} 正是 ORP 与 ERP 不一致的一个度量, 从而 $1 - MS_i / N_{ci}$ 是被试 i 的 ORP 和 ERP 一致性的度量, 为了使这个度量值可以从 -1 变到 1, 将其修改为 $HCI_i = 1 - 2MS_i / N_{ci}$ 。HCI_{*i*} 越大表示 ORP 和 ERP 越一致, 特别地, 在不失误情况下, $HCI_i = 1$ 。

3.2 修正的层级相合指标

Cui 和 Leighton (2009) 引入的 HCI 指标的定义中, 其分母是 $N_{ci} = \sum_{j: x_{ij}=1} \#S_j$, 而且只计算 $S_j = \{g|g \text{ 为项目, } AS_g \subseteq AS_j, g \neq j\}$ 。即项目 g 与项目 j 不相等时才计算入内, 这时可能使 $N_{ci} = 0$, 即 HCI 的分母为零。比如项目 i 仅仅包含一个属性 k 且同

类试题仅此一道, 被试 j 只在该试题上正确作答, 在这种情况下, HCI 的分母等于零。为了避免这种错误我们将 S_j 做如下修改: $S_j = \{g|g \text{ 为项目, } AS_g \subseteq AS_j\}$ 。修改以后的 HCI 记为 MHCI。

我们可以使用 MHCI 对所获得的 ORP 数据进行分析, 考察被试反应与专家命制项目所蕴涵的层级结构是否基本符合。如果基本符合, 可以使用认知诊断模型做进一步诊断分析, 否则应该寻找两者不相符的原因。因此, 在对数据进行认知诊断分析时, 建议分成三步走。

第一步, 在施测之前, 对测验 Q 阵计算理论构想效度(TCV)。如太低应该检查测验蓝图, 替换项目中的试题, 修改测验 Q 阵后再计算 TCV, 直到满意为止。如果不太低, 则转第二步;

第二步, 在施测之后, 考察每个被试的反应是否与属性层级相容。然后计算全体被试的 MHCI 的均值, 如果太低则可能 A&H 与被试的知识结构不符, 应该对 Q_i 进行修正后再转第一步; 否则进行第三步;

第三步, 选择 CDM 对数据进行分析。

注意, 在第二步中, Cui (2007) 提出 HCI 低于 0.6, 表示拟合不好; 而 Wang 和 Gierl (2007) 认为 HCI 的均值 0.3 以下就可以认为模型设置得不好, HCI 均值 0.6 以上是好模型,

MHCI 的标准可以此作为参考。

4 结论与讨论

认知诊断是近年来的研究热点, 受到学者的广泛关注。但是关于认知模型和诊断测验之间关系的探讨却极为有限。本文指出认知诊断测验与认知模型之间应具有一致性, 否则测验的效度就不能保证。本文对教育测量中认知模型的评估进行讨论。为了评估的需要, 我们和 Gierl 等人(2007)、Leighton 等人(2004)以及 Tatsuoaka (1995, 2009)一样, 先将认知模型用 Q 矩阵这种数学工具加以表达, 但本文中采用的表达方式更细致一些, 比如使用专门的记号以区分各种不同的 Q 矩阵; 给出了充分必要 Q 阵的概念以及它在认知诊断中的重要作用; 其次, 在领域专家给出属性和层级(A&H)以后, 测验蓝图(Q_i)是潜在 Q 阵(Q_p)的子矩阵。我们引入理论构想效度评估 Q_i 在多大的程度上代表 A&H, 并且对本文引证的文献中的一些例子进行分析, 说明它们的理论构想效度都不太高, 由此想提醒测验开发者注意; 最后, 由于很可能在现有测验基础上进行认知

诊断(即“翻新”),这时通过对测验项目的分析得到 Q_i 阵,这个 Q 阵代表的 A&H 与专家认知的 A&H 吻合程度应该加以评估;而在进行诊断分析时,不论是先有 A&H 再进行测验编制,还是先进行测验再从现成测验中挖掘出诊断所需的 Q 阵,这两种情况下 Q 阵的标识都可能存在问题。如何从被试反应数据去找出存在的问题,这非常重要。Cui 和 Leighton (2009)提出 HCI,本文对其进行了修改,以辅助判断 Q 阵标识的正确性和测验是否适合当前被试,即是否可以正确描述当前被试的知识状态。另外,纵使 Q 阵中每一列标注正确,也还存在测验安排是否合理的问题,即是否考虑了知识状态与期望反应模式一一对应的问题。如何从测验 Q 阵分析产生不同知识状态对应相同的期望反应模式的原因,即 Q 阵中缺乏哪些列(项目类),这是一个需要细致分析的问题。本文分别结合目前国际测量学界讨论得比较多的认知诊断测验的例子,就理论属性层级已知和未知的情形进行讨论,并且给出了一些相应的操作方法。

我们认为属性和属性层级关系对于认知诊断十分重要,至少对于认知诊断测验的编制有指导作用(丁树良等, 2010, 2011)。然而属性层级关系和问题解决中要求的属性顺序关系是不能混为一谈的。比如我们知道,要掌握异分母加减运算(A),至少先应该掌握两个整数的公倍数(最小公倍数)的计算(B)、分数的基本性质(分子、分母同时扩大或者缩小若干倍,其值不变)(C)、还要掌握同分母分数加减(D),这三者缺一不可;而且 B、C、D 是 A 的先决属性。至于运算的结果是否是最简分数,还要考察学生是否掌握求两个整数的最大公因子及分数的基本性质。在计算 $(1/2-1/3)/(1/5+1/6)$ 的时候,即问题解决中要求的属性顺序关系是属性 B、C、D 到 A,再是属性 B、C、D 到 A;当然最后还要用到异分母相除这个属性。可见,属性层级关系和问题解决中要求的属性顺序关系是有区别的。

Cui 和 Leighton (2009)提出 HCI 要衡量被试的观察反应模式和期望反应模式失拟的程度。HCI 关注被试失误的反应,即“父项目做对而子项目做错”(简称为“父对子错”)的情形,这可以解释为“失误”,它相当于属性层级模型中的 B 方法,然而“父对子错”也可以是由于“猜测”而造成的。我们认为这两者都考虑(相当于属性层级模型中的 A 方法)才比较全面揭示被试反应对期望反应或属性层级的偏离。然而这样的修正以后,这个指标的统计分布如何,

修正的 HCI 到底小到什么程度这一批数据才不可用,这是一个值得探讨的问题。另外 HCI 是假设属性层级正确的前提下,探测被试的反应模式对属性层级的“失拟”程度;然而 Q 矩阵的正确指定远非易事,因此虽然 Cui 和 Leighton (2009)讨论过在某些情况下 Q 矩阵错误标定(属性层级结构错误指定, model misspecification)对 HCI 的影响,但是如何应用 HCI (修正的 HCI)的异常去发现 Q 矩阵标定的错误,为 Q 矩阵的修正提供线索,这个问题值得探讨。当然 HCI 或 MHCI 指标只能评估 Q_i ,而不能修正 Q_i 。然而当评估结果表明 Q_i 存在问题时,就应该启动修改 Q_i 矩阵的程序。总之,我们认为认知诊断测验应该受到 A&H 的支持,并且应该尽可能代表 A&H。否则诊断信息就不一定准确、完备,测验的效度也不可能提高。

确立 A&H 是艰苦细致的工作,认知诊断测验编制是一个新的挑战,应该多加注意。当然,使用 HCI 必须了解理论上的 A&H,从而可以获得一系列的 Q 矩阵(比如前面列举的 Q_p, Q_t, Q_s)。在获取 Q 矩阵的时候,扩张算法起着重要作用。在上文中,有时候对 Q 矩阵使用扩张算法,这时应该加上“形式上”三个字,这是强调扩张算法是针对可达矩阵设计的,而对一般的测验 Q 矩阵使用扩张算法,只不过是对于 Q 矩阵的列以及扩张出来的列不断地使用“布尔并”运算而已。这时不能保证获得的矩阵是完整的潜在 Q 矩阵,只有在 Q 矩阵是充分必要 Q 矩阵(丁树良等, 2010, 2011)的条件下,这样做才能保证获得的矩阵是完整的潜在 Q 矩阵。比如对例 1 中的 Q 矩阵使用扩张算法,仅仅只能得到 4 个符合层级关系的非零列,而对由 Q 矩阵蕴涵的可达矩阵(3 阶单位矩阵)应用扩张算法,却可以导出 7 个符合层级关系的非零列。值得注意的是,由于 Q_i 阵不一定充要,由 Q_i 阵蕴涵的可达矩阵不一定符合理论的预期。同时,这时候计算得到的理论构想效度只能作为参考,因为由 Q_i 阵扩张出来的 Q_s 阵通常只是理论上的 A&H 导出的 Q_s 阵的子矩阵,其列数一般会比较少,从而 TCV 可能会虚假的升高。

本文只提出了一些提高认知诊断测验效度的初步设想,真正到实用阶段或许还有很长一段路要走。这里提出一些初步设想是想抛砖引玉,希冀引起相关讨论,使得认知诊断工作少走弯路。

参 考 文 献

Chen, P. (2011). *Item replenishing in cognitive diagnostic*

- computerized adaptive testing-based on DINA model. Unpublished doctoral thesis, Beijing Normal University.
- [陈平. (2011). 认知诊断计算机化自适应测验的项目增补-以 DINA 模型为例. 博士学位论文, 北京师范大学.]
- Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing. *Psychometrika*, 74, 619-632.
- Cui, Y. (2007). *The hierarchy consistency index: Development and analysis*. Unpublished doctoral dissertation, University of Alberta, Edmonton, Alberta, Canada.
- Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, 46(4), 429-449.
- DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, 35(1), 18-26.
- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.
- Ding, S. L., & Luo, F. (2005). Algorithm: From poset to Hasse diagram. *Journal of Jiangxi Normal University (Natural Sciences Edition)*, 29(2), 150-152.
- [丁树良, 罗芬. (2005). 求偏序关系 Hasse 图的算法. *江西师范大学学报(自然科学版)*, 29(2), 150-152.]
- Ding, S. L., Luo, F., Cai, Y., Lin, H. J., & Wang, X. B. (2008). Complement to Tatsuoka's Q matrix theory. In K. Shigemasa, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New Trends in Psychometrics* (pp. 417-423). Tokyo: Universal Academy.
- Ding, S. L., Wang, W. Y., & Yang, S. Q. (2009). Principle of the design of cognitive diagnostic test. Retrieved April 3, 2011, from <http://www.paper.edu.cn>
- [丁树良, 汪文义, 杨淑群. (2009). 认知诊断测验编制的原则. 中国科技论文在线. 2011-4-3 取自 <http://www.paper.edu.cn>.]
- Ding, S. L., Wang, W. Y., & Yang, S. Q. (2011). The design of cognitive diagnostic test blueprints. *Journal of Psychological Science*, 34(2), 258-265.
- [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.]
- Ding, S. L., Yang, S. Q., & Wang, W. Y. (2010). The importance of reachability matrix in constructing cognitively diagnostic testing. *Journal of Jiangxi Normal University*, 34(5), 490-494.
- [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.]
- Ding, S. L., Zhu, Y. F., Lin, H. J., & Cai, Y. (2009). Modification of Tatsuoka's Q Matrix theory (in Chinese). *Acta Psychologica Sinica*, 41(2), 175-181.
- [丁树良, 祝玉芳, 林海菁, 蔡艳. (2009). Tatsuoka Q 矩阵理论的修正. *心理学报*, 41(2), 175-181.]
- Fu, J., & Li, Y. (2007). *Cognitively Diagnostic Psychometric Models: An Integrative Review*. Paper presented at the the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
- Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice*, 19(3), 34-44.
- Gierl, M. J., Leighton, J. P., & Hunka, S. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitively Diagnostic Assessment for Education: Theory and Applications* (pp. 242-274). New York: Cambridge University Press.
- 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.
- Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule space approach. *Journal of Educational Measurement*, 41(3), 205-237.
- McGlohen, M. K. (2004). *The application of cognitive diagnosis and computerized adaptive testing to a large-scale assessment*. Unpublished doctoral dissertation, University of Texas at Austin.
- Qu, W. L., Geng, S. Y., & Zhang, L. A. (2008). *Discrete Mathematics*. Beijing: High Educational Press.
- [屈婉玲, 耿素云, 张立昂. (2008). *离散数学*. 北京: 高等教育出版社.]
- Rupp, A. A., & Templin, J. (2008). The effects of Q-Matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, 68(1), 78-96.
- Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models A case study. *Educational and Psychological Measurement*, 67(2), 239-257.
- Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 458-488). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Tatsuoka, K. K. (1995). Architecture of knowledge structure and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 327-361). Hillsdale, NJ: Erlbaum.
- Tatsuoka, K. K. (2009). *Cognitive Assessment: An introduction to the Rule Space Method*. New York: Routledge, Taylor & Francis Group.
- Tatsuoka, K. K., Linn, R. L., Tatsuoka, M. M., & Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. *Journal of Educational Measurement*, 25(4), 301-319.
- Wang, C. J., & Gierl, M. J. (2007). *Investigating the Cognitive Attributes Underlying Student Performance on the SAT Critical Reading Subtest: An Application of the Attribute Hierarchy Method*. Paper presented at the 2007 annual meeting of the National Council on Measurement in Education.
- Wang, W. Y. (2009). *Research on item selection strategies of computerized adaptive testing-Based on GRM and DINA*. Unpublished master's thesis, Jiangxi Normal University.
- [汪文义. (2009). 计算机化自适应测验选题策略研究-以 GRM 和 DINA 模型为例. 硕士学位论文, 江西师范大学.]
- Xu, Z. Y., Ding, S. L., Wang, W. Y., & Tu, D. B. (2010). Explore the test development to improve learning. *China Examinations*, (7), 3-11.
- [许志勇, 丁树良, 汪文义, 涂冬波. (2010). 促进学生学习测验开发的探讨. *中国考试*, (7), 3-11.]
- Yang, S. Q., Cai, Z. S., Ding, S. L., Lin, H. Q., & Ding, Q. L.

(2008). Augment algorithm for reduced Q-matrix. *Journal of Lanzhou University (Natural Sciences)*, 44(3), 87–91.
 [杨淑群, 蔡声镇, 丁树良, 林海菁, 丁秋林. (2008). 求解简化 Q 矩阵的扩张算法. *兰州大学学报(自然科学版)*, 44, (3), 87–91, 96.]

Yang, S. Q., & Ding, S. L. (2011). Theory and method for predicating of valid objects. *Journal of Jiangxi Normal University*, 35(1), 1–4.
 [杨淑群, 丁树良. (2011). 有效对象的判定理论与方法. *江西师范大学学报*, 35(1), 1–4.]

附录 1 属性的先决关系是偏序关系

由于 R 中对角元素均为 1, 即表示任一个属性自己可以到达自己, 即自反性, 而属性 A 是属性 B 的先决属性, 则 B 一定不是 A 的先决属性, 这表明只要 $i \neq j$ 则 R 中第 (i, j) 与 (j, i) 元素不可能同时为 1, 这表示先决关系的反对称性; 而如果有属性 A 是属性 B 的先决属性, 且属性 B 是属性 C 的先决属性, 则属性 A 为属性 C 的先决属性。这称为传递性, 即属性之间的先决关系是偏序关系。

附录 2 相同 ERP 导出知识状态等价关系的证明

设 S 为知识状态的集合, 给定测验蓝图 Q , 任取 $\alpha \in S$, 则 α 对应一个 ERP:

- 1) 由于任一个 S 中元素 α , α 与 α 有相同的期望反应模式, 故 $\alpha \cong \alpha$ (自反性);
- 2) 若 $\alpha, \beta \in S$, $\alpha \cong \beta$ 即 α 和 β 有相同的 ERP, 则 β 和 α 有相同的 ERP, 即必有 $\beta \cong \alpha$ (对称性);
- 3) 若 $\alpha, \beta, \gamma \in S$, $\alpha \cong \beta$ 且 $\beta \cong \gamma$, 即 α 和 β 有相同的 ERP, β 和 γ 有相同的 ERP, 则 $\alpha \cong \gamma$ (传递性)。

附录 3 缩减算法

设 Q 为 $k \times m$ 的 0-1 矩阵。

1) 求 Q 阵中各列之和, 且按列和大小, 列和大的列排在列和小的列的右边, 列重排以后的 Q 阵, 仍记为 $Q = (q_1, q_2, \dots, q_m)$;

2) 从第 m 列起, 若 q_m 由 Q 中其他列的布尔并表示出来, 则 $Q \leftarrow Q - \{q_m\}$; 且 $m \leftarrow m - 1$ 。3) 重复第 2) 步, 直到矩阵 Q 中所有的列均不能表示成其他列的布尔并。

Evaluating the Consistency of Test Items Relative to the Cognitive Model for Educational Cognitive Diagnosis

DING Shu-Liang¹; MAO Meng-Meng²; WANG Wen-Yi¹; LUO Fen¹; CUI Ying³

⁽¹⁾ School of Computer and Engineering, Jiangxi Normal University, Nanchang 330022, China

⁽²⁾ College of Education, NanChang University, Nanchang 330031, China

⁽³⁾ Centre for Research in Applied Measurement and Evaluation (CRAME), University of Alberta, CA

Abstract

Attributes and their hierarchy may present a cognitive model. Building a cognitive model is one of key steps for cognitive diagnosis as it is directly related to the validity and usefulness of test results.

It is very important to detect whether the test specification coincides with the cognitive model before administering the test. In this paper, an explicit index is given to measure the extent to which the cognitive model is represented by the test items of the diagnostic test. We call this index as the theoretical construct validity (TCV). In terms of TCV, the test reported by Tatsuoaka and her colleagues (1988) is reanalyzed, and the TCV of the test is only 9/24. That is, 24 knowledge states are obtained from the theoretic cognitive model but the test specification could distinguish only 9 knowledge states.

Cui and her colleagues established a person fit index named hierarchy consistency index (HCI) to detect the fitness of an examinee's observed response pattern (ORP) to an expected response pattern (ERP). HCI is not defined well when an examinee mastered one attribute only and there is only one item measuring the same attribute in the test, and the examinee responses correctly to the item. The original HCI could not compute for the number of comparisons being zero, hence the denominator being zero. In addition, the HCI includes the slipping only. Combining the slipping and guessing in the response to an item, we proposed a new index called modified HCI (MHCI). MHCI could detect the consistency of the response data and the cognitive model that is obtained by experts.

Key words cognitive diagnosis; cognitive model; cognitive diagnostic test