

基于等级反应模型的规则空间方法*

田 伟 辛 涛

(北京师范大学发展心理研究所, 北京 100875)

摘 要 本研究基于 Tatsuoka 的规则空间方法, 对理想反应模式与异常反应指标进行了扩展, 推导了多级评分项目下规则空间方法的算法公式。在 4 种属性层级结构(发散型、收敛型、线型与无结构型)×4 种“失误”作答概率(2%、5%、10%与 15%)测验情境下, 以属性模式判准率、被试属性判准率、敏感性与特异性为指标, 检验了多级评分项目下规则空间方法的分类准确性。结果表明: (1) 基于多级评分项目构建的异常反应指标, 能有效地对被试进行分类与解释, 且 0-1 评分项目下异常反应指标及其性质都是多级评分下的特例; (2) 随着“失误”作答概率的增加, 4 种属性层级结构的分类准确性都会降低; (3) 线型和收敛型的分类准确性明显好于发散型与无结构型; (4) 纯规则点的分布对规则空间方法的分类准确性有显著影响。

关键词 多级评分项目; 规则空间方法; 等级反应模型; 异常反应指标; 理想反应模式

分类号 B841

1 前言

成就测验的测量对象是经过一段正式或非正式的教育经历之后, 被试在一个定义相对良好的领域之内对知识与技能的掌握情况, 它在教育中起着非常重要的作用。一般采用经典测量理论(Classical Test Theory, CTT)或者项目反应理论(Item Response Theory, IRT)估计被试的能力水平, 通过分数量尺转换后对被试进行评估与解释, 估计被试在群体中的位置, 进行个体间的比较以及预测未来学习活动中被试的表现。随着教育的不断深入, 一方面教师希望测验能成为课堂教学活动的一个重要部分, 助其达到最终的教学目标; 另一方面在很多大尺度评估中, 学校和教师也希望得到更多的教学结果信息, 帮助其提高教育质量。这就需要教育测验能够有效对被试的知识、技能或者其他认知特征进行评估。然而, 从传统成就测验分数中很难获得学生应该巩固或提高哪些具体知识技能以及教师应该补救哪些教学内容的诊断信息。20 世纪 60 年代以来, 随着认知心理学的不断发展, 人们对被试问题解决的心理过程有了更深的认识, 认知理论在教

育和心理测验中的重要作用已经受到很多心理测量专家的重视(例如 Embretson, 1985; Frederiksen, Glaser, Lesgold, & Shafto, 1990; Frederiksen, Mislevy, & Bejar, 1993; Irvine & Kyllonen, 2002; Nichols, Chipman, & Brennan, 1995; Snow & Lohman, 1989), 这就促进了认知诊断评估的不断发展。认知诊断评估是 20 世纪 80 年代中期认知心理学与心理计量学结合的产物, 用来测量学生具体的知识结构与认知技能, 可以给学生提供认知优势与不足的剖面诊断。认知模型与心理计量学模型是认知诊断评估的两个基础理论, 首先是基于认知心理学构建学生解决问题时的认知模型, 以此编制认知诊断测验并对被试进行施测, 然后选择认知诊断模型对被试进行分类诊断。目前, 常用的认知诊断模型有规则空间方法(Rule Space Methodology, RSM)、属性层级方法(Attribute Hierarchy Methodology, AHM)、DINA 模型(Deterministic-Input, Noisy-And, DINA)等, 这些模型建立之初都是基于 0-1 评分项目的。然而, 如果研究者要从测验表现中最大限度地提取被试的技能掌握信息, 那么我们就从单个项目表现中提取尽可能多的信息, 在这

收稿日期: 2010-10-20

* 国家自然科学基金项目(30670718 & 11171029)和教育部新世纪优秀人才支持计划(NCET-07-0097)资助。

通讯作者: 辛涛, E-mail: xintao@bnu.edu.cn

方面等级评分项目具有一定的价值(Bolt, 2007)。例如, 在 0-1 评分的 RSM 与 AHM 中的一个基本假设是, 当且仅当被试掌握了一个项目所涉及到的所有属性时, 才能正确作答该项目。假如一个项目测量了 K 个属性, 根据这个假设掌握 $K-1, K-2, \dots, 1, 0$ 个属性的被试在这个项目上都是 0 分。由此可见, 0-1 评分项目的认知诊断没有最大限度地提取被试知识结构中的丰富信息, 而多级评分却可以最大限度地地区分被试的测验表现与知识结构。因此, 针对多级评分的认知诊断模型需要进行更多的研究。近年来, 研究者开始将基于 0-1 评分项目的认知诊断模型扩展到多级评分项目, 例如融合模型(Fusion Model with Partial Credit Items, Bolt & Fu, 2004)、属性层级方法(祝玉芳, 丁树良, 2009)、DINA 模型(涂冬波, 蔡艳, 戴海琦, 丁树良, 2010)。本研究是将基于 0-1 评分项目的规则空间方法扩展到多级评分项目, 扩大规则空间方法的应用范围, 最大限度地挖掘被试反应模式中的技能掌握信息。

规则空间方法是 Tatsuoka 在 20 世纪 80 年代提出的认知诊断方法。从理论上讲, 规则空间方法是认知心理学、IRT 与多元统计相结合的产物(辛涛, 焦丽亚, 2006), 它将被试在测验项目上的作答反应分类到某个与认知技能相联系的属性掌握模式(Attribute Mastery Pattern, AMP), 模型中的 Q 矩阵也被后续的认知诊断模型所采用, 为认知诊断理论模型的发展奠定了基础。 Q 矩阵(Tatsuoka, 1983)表达的是项目与属性之间的关联关系, 行表示属性, 列表示项目, 其元素 $q_{ki}=1$ 或 0 , 1 表示第 i 个项目考察了属性 k , 0 即为没有考察属性 k 。通过比较 Q 矩阵行与行之间的包含关系, 就可以得到反映属性之间先决关系的 A 矩阵(邻接矩阵)或 R 矩阵(可达矩阵), A 矩阵表示属性之间的直接关系, R 矩阵表示属性之间的直接与间接关系。在所有属性组合模式中, 根据 R 矩阵删除不符合属性层级关系的组合模式, 就得到了属性掌握模式, 即知识状态。再通过确定性规则就可以得到属性掌握模式在测验项目上对应的理想反应模式(Ideal Item-response Pattern, IRP)。将理想反应模式与被试的观察反应模式(Observed Response Pattern, ORP)合并, 利用参数估计软件(例如 BILOG、MULTILOG、PARSCALE 等)得到被试的能力参数值 θ 以及项目参数估计值, 结合异常反应指标 ζ (Caution Index) (2.2 节中进行详细介绍), 就可以将理想反应模式和观察反应模式映射到“规则空间”中, 所谓的“规则空间”即为以

(θ, ζ) 为坐标的二维笛卡尔空间, 理想反应模式所对应的点称为“纯规则点”, 表征所有可能的知识状态。最后计算观察反应模式属于每一个知识状态的后验概率, 将被试分入后验概率最大的知识状态。

在规则空间方法中, 理想反应模式与异常反应指标都与计分方式有关。在多级评分项目下, 规则空间方法的理想反应模式和异常反应指标的定义与计算将随之发生变化, 需要将这两个概念及其计算方法进行扩展。因此, 本研究将首先回顾在 0-1 评分项目下, 理想反应模式与异常反应指标的定义与计算方法, 指出计分方式变化带来的概念涵义改变; 然后, 在此基础上对理想反应模式与异常反应指标进行扩展, 给出在多级评分项目下, 理想反应模式和异常反应指标的定义与计算公式, 从理论上对规则空间方法的公式进行扩展; 最后进行模拟研究, 采用蒙特卡洛方法, 在 4 种属性层级结构测验的 4 种不同“失误”概率下的测验情景中, 基于 4 个统计指标, 检验多级评分项目下规则空间方法的分类准确性。希望通过本研究给出多级评分项目下合理的理想反应模式与异常反应指标的算法公式, 从而扩大规则空间方法的应用范围。

2 多级评分对理想反应模式与异常反应指标的影响

规则空间方法借助于 Q 矩阵理论和统计模式识别, 最终将观察反应模式分到预先确定的组别(Fukunaga, 1990; Ripley, 1996)。在认知诊断中, 属性掌握模式表示的被试知识状态就是这里预先确定的组别。假设一个测验测量了 K 个属性, 那么第 j 个被试的属性掌握模式可以表示为 $\alpha_j=(\alpha_{j1}, \alpha_{j2}, \dots, \alpha_{jK})$, α_{jk} 表示第 j 个被试在第 k 个属性上的掌握情况, $\alpha_{jk}=1$ 表示被试掌握了第 k 个属性, $\alpha_{jk}=0$, 表示没有掌握。然而属性掌握模式是不可直接观察的, 于是 Tatsuoka (1991)引入了布尔描述函数(Boolean Description Function, BDF)求得理想反应模式, 由于理想反应模式由知识状态和 Q 矩阵决定, 即有 $KS(Knowledge State) \underline{Q} IRP$, 而 $ORP \underline{RSM} IRP$ 。如果 KS 与 IRP 一一对应, 即 $KS \underline{Q} IRP$, 则有 $ORP \underline{RSM} IRP \underline{Q} KS$, 诊断中的分类问题就可以解决(丁树良, 汪文义, 杨淑群, 2009)。

基于规则空间方法的认知诊断即是将被试的观察反应模式分类到与之最近的理想反应模式中, 即 $ORP \underline{RSM} IRP \underline{Q} KS$ 。Tatsuoka (1983)等人通过降维方式将理想反应模式与观察反应模式压缩成了

一个具有语义的异常反应指标 ζ , 它表示的是每一个项目上的观察得分与其期望得分之间差异的加权之和, 再结合 IRT 中的能力参数 θ , 就构建了一个以 (θ, ζ) 为坐标的二维笛卡尔空间, 即“规则空间”。最后在这个“规则空间”中借助于统计模式识别对被试进行分类。由于理想反应模式与评分方式有关, 而异常反应指标又是反应模式的函数, 评分方式的变化也随之使理想反应模式与异常反应指标的定义与计算方法发生改变。因此这里将主要介绍规则空间方法中的理想反应模式和异常反应指标的定义与计算, 在此基础上指出规则空间方法扩展中的研究问题。

2.1 0-1 评分的理想反应模式及其多级变化

在 0-1 评分项目下, Tatsuoka 假定“当且仅当被试掌握了一个项目所涉及到的所有属性时, 被试才能正确作答该项目”, 并利用布尔代数运算求得理想反应模式。布尔算子的定义是, 对于加法有 $1+1=1$, $1+0=0+1=1$, $0+0=0$, 对于乘法有 $1 \times 1=1$, $0 \times 1=1 \times 0=0$, $0 \times 0=0$ 。例如, 在一个考察 4 个属性的诊断测验中, 共有 7 题。假设一个属性掌握模式为 $\alpha=(1100)$ 的理想被试, 7 个题目考察的属性分别为 $i_1=(1000)'$, $i_2=(1100)'$, $i_3=(1110)'$, $i_4=(1110)'$, $i_5=(1001)'$, $i_6=(1101)'$, $i_7=(1111)'$ 。根据布尔代数定义, 在 i_1 的 4 个属性上分别有 $1 \times 1=1$, $1 \times 0=0$, $0 \times 0=0$, $0 \times 0=0$, 就得到了向量 (1000) , 由于 $(1000) \subseteq (1100)$, 即项目所考察的属性都包含在被试掌握的属性之中, 所以这个属性掌握模式在第 1 题上的反应为 1, 以此类推可以得到这个属性掌握模式的理想反应模式为 (1100000) 。而在多级评分项目下, 假设一个测验考察了 K 个属性, 共有 n 个项目, 每一个项目的最高分分别为 $(m_1, m_2, \dots, m_n)$, $m_i > 1$, $i=1, 2, \dots, n$, 被试在任何一个项目上的反应在 $[0, m_i]$, $i=1, 2, \dots, n$ 。可见, 评分不是正确或者错误两类, 而是等级计分, 理想反应模式的定义及其计算也将随之发生变化。

2.2 0-1 评分的异常反应指标及其多级变化

在 0-1 评分项目下, 异常反应模式是基于 Guttman 量表定义的, 当被试的反应模式中有难题做对而简单题做错时, 它就是一个异常反应, 假设一个测验有 n 个项目, 则可以用 $f(X) = \langle P(\theta) - X, P(\theta) - T(\theta) \rangle$ 来刻画被试反应模式的异常程度, 这里 $P(\theta) = [P_1(\theta), P_2(\theta), \dots, P_n(\theta)]$ 是由 1PLM、2PLM 或 3PLM 表示的正确作答概率向量, $X = [x_1, x_2, \dots, x_n]$ 是由 0 或者 1 组成的观察反应向量, $T(\theta) =$

$[\sum_{i=1}^n P_i(\theta)]/n$ 。 $f(X)$ 是异常反应指标, 将 $f(X)$ 标准化之后就得到了 ζ , ζ 越大, 表示多是难题做对而简单题做错, 反应模式异常。Tatsuoka (1985) 证明了对于任意一个给定的能力水平 θ , 所有反应模式对应的 $f(X)$ 的期望为 0, 方差为 $\sum_{i=1}^n P_i(\theta)Q_i(\theta)[P_i(\theta)-T(\theta)]^2$, 且与能力水平正交。

在多级评分项目下, Guttman 量表是不存在的, 每个题目的分值大小也不同, 不能直接比较得分高低判断异常反应模式。例如假定有难度不同的两个项目, i_1 是简单题, 最高分是 5 分, i_2 是难题, 最高分是 20 分, 如果一个被试的反应模式为 $(4, 8)$, 如果按照难题得分高而简单题得分低, 它是一个异常反应, 但由于两个题目的量纲不同, 我们不能简单地据此而判定它是一个异常反应模式。那么异常反应模式如何定义? 它的量化公式——异常反应指标怎样构建? 异常反应指标的期望、方差及其与能力水平之间的协方差矩阵如何计算? 异常反应指标的语义是否与 0-1 评分下的具有相同的含义? 这些都是规则空间方法扩展中需要解决的问题。

3 多级评分项目的理想反应模式与异常反应指标扩展

3.1 多级评分项目的理想反应模式扩展

实际考试评分时一般将多级评分项目的解题步骤分解成一些“采分点”, 这些“采分点”对应于规则空间方法中的属性。当被试在一个采分点上正确作答时, 就可以得到 1 分, 否则得 0 分, 这样所有采分点上的得分之和是被试在这个项目上的总分。假定满分为 m_i 的项目含有 m_i 个属性, 每个属性赋值为 1, 按照属性评分, 被试正确作答一个属性则累计 1 分 (Tatsuoka, 1995; 祝玉芳, 丁树良, 2009), 根据这个特点, 多级评分项目下规则空间方法中的理想反应模式就定义为, 在没有失误与猜测的情况下, “被试能够正确作答的项目的属性个数”。根据这个定义就可以给出多级评分项目下理想反应模式的两种计算方法:

(1) 矩阵乘法: 一般地, 假设一个测验有 n 个项目, 测量了 K 个属性。根据属性之间的层级关系, 得到一个表征属性之间直接与间接关系的 R 矩阵 (K 阶方阵); 再根据 R 矩阵从所有属性组合中删除不符合属性层级关系的 $\alpha_j = (\alpha_{j1}, \alpha_{j2}, \dots, \alpha_{jK})$, 就得到了属性掌握模式 (AMP) $_{m \times K}$, m 表示所有可能的知识状态个数。 Q 是一个 $K \times n$ 矩阵, n 表示项目个数。将 $AMP \times Q$ 就得到了多级评分项目下的理想反应模式 IRP 。

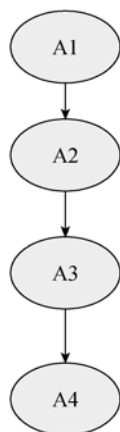


图 1 4个属性的层级关系

表 1 根据图 1 得到的项目属性关联矩阵 Q

属性	题目			
	i_1	i_2	i_3	i_4
A_1	1	1	1	1
A_2	0	1	1	1
A_3	0	0	1	1
A_4	0	0	0	1

注： A_1, A_2, A_3, A_4 表示测验测量的 4 个属性； i_1, i_2, i_3, i_4 表示测验由 4 个项目组成

表 2 图 1 得到的属性掌握模式 AMP 及其对应的理想反应模式 IRP

编号	属性掌握模式(AMP)				理想反应模式(IRP)			
	A_1	A_2	A_3	A_4	i_1	i_2	i_3	i_4
1	1	0	0	0	1	0	0	0
2	1	1	0	0	1	2	2	2
3	1	1	1	0	1	2	3	3
4	1	1	1	1	1	2	3	4

以图 1 所示的 4 个属性之间的层级关系为例，直接可以得到 R 矩阵；再构造一个包含 2^4-1 行 4 列的所有属性组合矩阵，从中删除不符合属性层级关系的属性组合模式，就得到了属性掌握模式 AMP 如表 2 所示；再将 AMP 转置后作为测验 Q 矩阵如表 1 所示；将 $AMP \times Q$ 就得到了 IRP 如表 2 所示。例如 $\alpha_3=(1\ 1\ 1\ 0)$ ，它在第 3 个项目上的理想反应是 α_3 与 Q 的第三列对应元素乘积之和，即 $3=1 \times 1 + 1 \times 1 + 1 \times 1 + 0 \times 0$ 。由此可见，只有 1×1 对项目总分产生贡献，它表示被试掌握了项目所涉及到的对应属性，根据多级评分项目下理想反应模式的定义，这个属性上的得分为 1；而 0×1 表示被试没有掌握项目所涉及到的对应属性，因此这个属性上的得分为 0；而 1×0 表示被试掌握了对应属性，而

项目没有涉及到这个属性。可见，矩阵乘法体现的是“属性计分”。

(2) 向量加法： AMP 中的每一行与 Q 的每一列对应元素相加，向量中元素 2 的个数为理想被试的项目得分。例如 AMP 的第一行是 $(1\ 0\ 0\ 0)$ ， Q 的第一列是 $(1\ 0\ 0\ 0)'$ ，对应元素相加可得到向量 $(2\ 0\ 0\ 0)$ ，元素 2 的个数为 1，因此 IRP 的第一行第一列为 1，即第一个理想被试在第一道题目上的得分为 1，以此类推就可得到 IRP 。

3.2 多级评分项目的异常反应模式定义及其度量——异常反应指标的扩展

在 0-1 评分项目下，异常反应模式通过 Guttman 量表定义。而在多级评分项目下，虽然不存在 Guttman 量表，但是每个项目都有难度高低之分。假设 N 个被试参加了一个长度为 n 的测验，根据 CTT 中难度定义 $p_i = \bar{u}_i / m_i$ ， $\bar{u}_i$ 表示 N 个被试在项目 i 上的平均得分， m_i 是项目 i 的最高分。将 n 个项目的难度 $p_1, p_2, \dots, p_n$ 按照难度从高到低或者从低到高排列后，再将每个被试在每个项目上的得分转化为得分比率，即项目得分除以最高分。若是难题得分比率高，而简单题得分比率低，就是异常反应。

3.2.1 多级评分项目下异常反应指标的构建

在多级评分项目下，被试反应模式异常程度的度量仍根据项目残差的思想来构建异常反应指标 ζ 。假设测验长度为 n ，能力为 θ 的被试在 n 个项目上的观察反应模式为 $U = [u_1, u_2, \dots, u_n]$ ， $u_i \in [0, m_i]$ ， u_i 表示能力为 θ 的被试在第 i 个项目上的观察得分， m_i 表示第 i 个项目的最高得分。又设 $P_{ij}(\theta)$ 表示能力为 θ 的被试在第 i 个项目上得分等级为 j 的项目反应函数，于是可以得到能力为 θ 的被试在第 i 个项目上的期望得分，即项目特征函数为：

$$P_i(\theta) = \sum_{j=1}^{m_i} j * P_{ij}(\theta)$$

因此可以得到能力为 θ 的被试在 n 个项目上的期望反应模式为：

$$P(\theta) = [P_1(\theta), P_2(\theta), \dots, P_n(\theta)]$$

为了消除每个项目量纲不同对异常反应模式解释造成的影响，再将能力为 θ 的被试在第 i 个项目上的期望得分转化为期望得分比率，即：

$$R_i(\theta) = P_i(\theta) / m_i$$

进而就得到了能力为 θ 的被试在 n 个项目上的期望得分比率向量为：

$$R(\theta) = [R_1(\theta), R_2(\theta), \dots, R_n(\theta)]$$

最后将能力为 θ 的被试在 n 个项目上期望得分比率求均值, 就得到了测验期望得分比率向量为:

$$T(\theta) = [t(\theta), t(\theta), \dots, t(\theta)], \quad t(\theta) = \left[\sum_{i=1}^n R_i(\theta) \right] / n.$$

同样为了消除量纲的影响, 再将每一个项目上的观察得分也转化为观察得分比率, 即:

$$v_i = u_i / m_i$$

因此观察得分比率向量为:

$$V = [v_1, v_2, \dots, v_n], v_i \in [0, 1]$$

最后将 $R(\theta) - V$ 与 $R(\theta) - T(\theta)$ 做内积来构建多级评分项目下的异常反应指标:

$$f(V) = \langle R(\theta) - V, R(\theta) - T(\theta) \rangle$$

将这个式子展开可以得到:

$$f(V) = \sum_{i=1}^n [R_i(\theta) - t(\theta)] * [R_i(\theta) - v_i]$$

3.2.2 多级评分项目下异常反应指标的意义 对于能力为 θ 的被试, 将 $f(V) = \langle R(\theta) - V, R(\theta) - T(\theta) \rangle$ 展开可得:

$$f_\theta(V) = [R(\theta) - T(\theta)] R'(\theta) - [R(\theta) - T(\theta)] V'$$

对于任意一个能力 θ , 当项目参数已知时, 第一项是一个常数, $f_\theta(V)$ 会随着观察得分比率向量 V 的变化而变化。将第二项写成求和形式为 $-\sum_{i=1}^n [R_i(\theta) - t(\theta)] v_i$ 。若项目按照难度从易到难排列后, 则有:

$$R_1(\theta) > R_2(\theta) > \dots > R_n(\theta),$$

由于 $t(\theta) = (1/n) \sum_{i=1}^n R_i(\theta)$, 因此可得 $R_1(\theta) > R_2(\theta) > \dots > R_k(\theta) > t(\theta) > R_{k+1}(\theta) \dots > R_n(\theta)$ 。

对于前面 k 个项目, $-\sum_{i=1}^n [R_i(\theta) - t(\theta)] v_i$ 中 v_i 前面的系数为负; 后面 $n-k$ 个项目的系数为正。若是正常反应模式, 多是简单题得分比率高, 难题得分比率低, 被试在前面 k 个项目上的得分比率 v_i 都较大, 后面 $n-k$ 个项目上的得分比率 v_i 都较小, $-\sum_{i=1}^n [R_i(\theta) - t(\theta)] v_i$ 较小(负值), $f_\theta(V)$ 也较小; 若是异常反应模式, 多是简单题得分比率低, 难题得分比率高, 被试在前面 k 个项目上的得分比率 v_i 较小, 后面 $n-k$ 个项目上的得分比率 v_i 都较大, $-\sum_{i=1}^n [R_i(\theta) - t(\theta)] v_i$ 较大(正值), $f_\theta(V)$ 也较大。由此可见 $f_\theta(V)$ 越小, 表明被试反应模式越正常; $f_\theta(V)$ 越大, 表明被试反应模式越异常, $f_\theta(V)$ 可以作为衡量被试反应模式异常程度的指标。

3.2.3 多级评分项目下异常反应指标的性质

(1) 异常反应指标的期望与方差: 对于任意一个给定的能力水平 θ , 所有反应模式对应的 $f(V)$ 期望为 0; 方

差为 $\sum_{i=1}^n [R_i(\theta) - t(\theta)]^2 * \left[\sum_{j=1}^{m_i} v_j^2 P_{ij}(\theta) - R_i^2(\theta) \right]$;

(2) 对于任意一个给定的能力水平 θ , 能力水平 θ 与异常反应指标正交(证明基于等级反应模型)。

由于在项目反应理论中, $\text{var}(\theta) = 1/I(\theta)$ ($I(\theta)$ 是测验信息函数), 再根据上述 $f(V)$ 的基本性质就可以得到能力水平 θ 与异常反应指标 $f(V)$ 的协方差矩阵。

$$\begin{pmatrix} \text{var}(\theta) & \text{cov}(\theta, f(V)) \\ \text{cov}(\theta, f(V)) & \text{var}(f(V)) \end{pmatrix} = \begin{pmatrix} 1/I(\theta) & 0 \\ 0 & \sum_{i=1}^n [R_i(\theta) - t(\theta)]^2 * \left[\sum_{j=1}^{m_i} v_j^2 P_{ij}(\theta) - R_i^2(\theta) \right] \end{pmatrix}$$

将 $f(V)$ 标准化之后就得到了 ζ , θ 与 ζ 的协方差矩阵就可以简化为:

$$\begin{pmatrix} \text{var}(\theta) & \text{cov}(\theta, \zeta) \\ \text{cov}(\theta, \zeta) & \text{var}(\zeta) \end{pmatrix} = \begin{pmatrix} 1/I(\theta) & 0 \\ 0 & 1 \end{pmatrix}$$

根据 θ 与 ζ 的协方差矩阵, 就可以在规则空间中, 进一步计算观察反应模式所对应的点到每一个纯规则点之间的马氏距离与贝叶斯后验概率, 将被试分类到后验概率最大的知识状态, 从而诊断出被试在解决问题过程中所使用的知识结构与解题策略。

4 模拟研究

4.1 研究目的

模拟研究的主要目的是检验多级评分项目下的规则空间方法对被试分类的准确性。一般来说, 实际测验测量的属性之间具有一定的层级关系, 是由线型、发散型、收敛型与无结构型结合形成的复杂网络(Kim, 2001)。另一方面, 被试在实际测验过程中, 由于环境、应试动机、情绪、态度等随机因素的影响, 被试在项目作答中会产生猜测或者失误。为了方便起见, 参照以往研究(Leighton, Gierl, & Hunka, 2004; 祝玉芳, 丁树良, 2009), 本研究将考察 4 种属性层级结构与 4 种“失误”水平对多级评分项目下规则空间方法分类准确性的影响, 希望能对测验设计与模型应用提供参考。

4.2 研究方法

4.2.1 研究设计 研究采用 Leighton 等人(2004)的 4 种属性层级结构, 分别为发散型、收敛型、线型与无结构型, 具体的属性层级关系见图 2; 4 种“失误”水平为 2%、5%、10%与 15%(祝玉芳, 丁树良, 2009)。模拟研究共有 $4 \times 4 = 16$ 种测验情景, 每个测

验情景都重复进行 30 次以减少误差, 模拟研究的整体设计见表 3。

4.2.2 观察反应数据生成方法 首先确定测验 Q 矩阵, 即根据 4 种属性层级结构可以依次得到 A 矩阵、 R 矩阵、属性掌握模式 AMP , 将计算得到的属性掌握模式转置后作为测验 Q 矩阵。从理论上说, 对于测量 7 个属性的测验, 所有属性组合得到的潜在项目集中共有 $2^7-1=127$ 道题目, 然而当属性之间存在层级关系时, 潜在项目集中的题目可以简化 (Leighton et al., 2004)。属性之间的层级关系越多, 潜在项目集就越小。这里以线型结构测验为例来介绍潜在项目集简化。理论上测量 1 个属性的题目个数有 C_7^1 个, 如表 4 所示。然而, 属性之间具有层级关系, 例如, 第 2 题测量了第 2 个属性 A_2 , 属性层级关系表明, 属性 A_2 需要它的先决属性 A_1 , 因此项目 2 必须通过(1 1 0 0 0 0 0)表征, 这个表征又使项

目 2 与项目 8 相同, 鉴于此可以将项目 2 删除。按照相同的方法, 就可以得到线型结构的测量项目共有 7 个。同样的方法, 发散型共有 25 题, 收敛型共有 8 题, 无结构型共有 64 题。

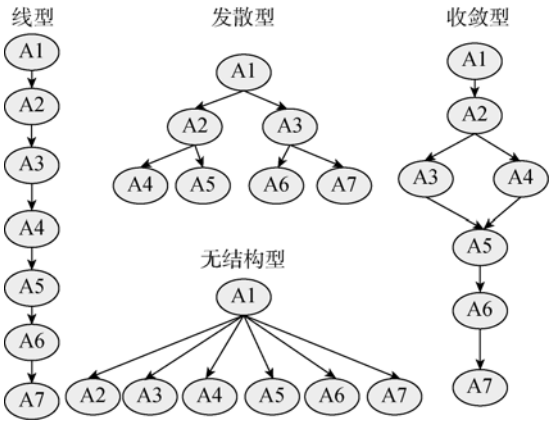


图 2 七个属性之间的四种层级结构

表 3 模拟研究的整体设计表

模拟次数	线型/发散型/无结构型/收敛型			
	2%	5%	10%	15%
1	①根据属性层级关系图生成测验 Q 矩阵与理想反应模式; ②生成观察反应模式 1: 使理想反应模式产生随机“失误”; ③分类诊断: 被试能力与项目参数估计 & 规则空间方法分类; ④计算评价分类结果的统计指标: S_1 。			
...				
30	①根据属性层级关系图生成测验 Q 矩阵与理想反应模式; ②生成观察反应模式 30: 使理想反应模式产生随机“失误”; ③分类诊断: 被试能力与项目参数估计 & 规则空间方法分类; ④计算评价分类结果的统计指标: S_{30} 。			
平均值	将 30 次的统计指标结果求平均, 减少实验误差。			

表 4 线型结构潜在项目集中的部分项目

属性	题目												
	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}
A_1	1	0	0	0	0	0	0	1	1	1	1	1	1
A_2	0	1	0	0	0	0	0	1	1	1	1	1	1
A_3	0	0	1	0	0	0	0	0	1	1	1	1	1
A_4	0	0	0	1	0	0	0	0	0	1	1	1	1
A_5	0	0	0	0	1	0	0	0	0	0	1	1	1
A_6	0	0	0	0	0	1	0	0	0	0	0	1	1
A_7	0	0	0	0	0	0	1	0	0	0	0	0	1

其次计算理想反应模式, 即根据属性掌握模式与测验 Q 矩阵, 采用矩阵乘法计算理想反应模式。

最后生成观察反应模式, 即将每个理想反应模式的分量加上随机误差, 造成“失误” (失误或猜测) 后所得到的反应向量作为观察反应模式。具体步骤是, 首先将理想反应模式按照总分从低到高排序,

然后使具有这些知识状态的观察被试人数服从标准正态分布, 得分若是相同的理想反应模式, 平均分配人数。然后为了得到产生“失误”之后的观察反应模式, 研究所用的模拟方法是将理想反应模式分量加上随机误差。例如要生成每个被试在每个项目上有 2% 的概率发生“失误”的情况。先产生一个服

从 $U(0,1)$ 的随机数 r , 如果 $r > 0.99$ 且理想反应模式分量不是满分, 则理想反应模式分量增加 1 分, 当理想反应模式分量是满分时, 则理想反应模式分量减 1 分; 如果 $r < 0.01$ 且理想反应模式分量不是 0 分, 则理想反应模式分量减 1 分, 当理想反应模式分量

为 0 分时, 增加 1 分; 如果 $0.01 \leq r \leq 0.99$, 则理想反应模式分量保持不变, 如表 5。这样就可以模拟生成“失误”概率为 2% 的观察反应模式。同样的方法可以得到 5%、10% 和 15% “失误”概率时的观察反应模式。

表 5 随机“失误”作答概览表

随机数范围	非满分	满分	非 0 分	0 分	保持不变
$r > 0.99$	↑1	↓1			
$r < 0.01$			↓1	↑1	
$0.01 \leq r \leq 0.99$					—

4.2.3 统计评价指标 假设一个测验测量了 K 个属性, 参加测验的被试有 N 个, $\alpha_j = (\alpha_{j1}, \alpha_{j2}, \dots, \alpha_{jK})$ 表示第 j 个被试的真实属性掌握模式, $j=1, 2, \dots, N$, $\alpha_{jk}=1$ 表示第 j 个被试掌握了第 k 个属性, $\alpha_{jk}=0$ 表示没有掌握, 掌握属性有 n_{j1} 个, 没有掌握属性有 n_{j2} 个。将第 j 个 ($j=1, 2, \dots, N$) 被试被判归的属性掌握模式与其真实属性掌握模式比较, 基于以下 4 个统计指标评价规则空间方法的分类准确性。

(1) 模式判准率, 即所有属性都判准的被试比例。如果两个属性掌握模式完全一致, 则 $f_j=1$, 否则 $f_j=0$ 。那么模式判准率可以表示为 $\sum_{j=1}^N f_j / N$, 这个值越大, 表明分类准确性越高。

(2) 被试属性判准率, 即每个被试被判准的属性比例。如果两个属性掌握模式在某个属性上一致, 则 $h_{jk}=1$, 否则 $h_{jk}=0$ 。那么一个被试的属性判准率可以表示为 $\sum_{k=1}^K h_{jk} / K$ 。

(3) 敏感性(Sensitivity): 敏感性指标常见于医学领域, 它表示当一个来诊者真实患病时, 被检测出来的比例。在规则空间方法中, 它表示每个被试未掌握属性被判为未掌握的比例, 如果一个属性上真实状态是未掌握, 而被判为未掌握, 则 $r_{jk}=1$, 否则 $r_{jk}=0$ 。那么一个被试的敏感性可以表示为 $\sum_{k=1}^K r_{jk} / n_{j2}$ 。

(4) 特异性(Specificity): 也是医学领域的一个指标, 它表示当一个来诊者真实没有患病时, 被检测出来没有患病的比例。在规则空间方法中, 它表示每个被试掌握属性被判为掌握的比例。如果在一个属性上真实状态是掌握的, 而被判为掌握, 则 $u_{jk}=1$, 否则 $u_{jk}=0$ 。那么一个被试的特异性可以表示为 $\sum_{k=1}^K u_{jk} / n_{j1}$ 。

在判别一个被试的知识状态时, 期望敏感性与特异性很高, 这意味着规则空间方法能敏感地检测出被试没有掌握的知识, 同时也不会把那些已经掌

握的知识诊断为没有掌握。两者任何一个指标过低, 都可能出现将被试已经掌握的知识诊断为没有掌握, 没有掌握的知识诊断为掌握, 这会降低对被试进行补救教学的指导作用。

4.2.4 数据分析

(1) 参数估计

本研究采用 MULTILOG 7.03 估计被试能力参数与项目参数。首先将理想反应模式与观察反应模式合并进行参数估计, 得到被试的能力水平及其标准误、项目参数估计值。

(2) 多级评分规则空间方法的分类

研究基于规则空间方法, 采用 MATLAB 程序对被试分类。首先根据 Q 矩阵计算可达到矩阵 R ; 然后根据排列组合原理得到属性之间的所有组合, 结合 R 矩阵将不满足属性层级关系的属性组合删除(减法算法), 得到属性掌握模式; 再次根据矩阵乘法计算理想反应模式, 将理想反应模式与观察反应模式合并, 借助于 MULTILOG 7.03 估计项目参数与被试能力参数, 计算多级评分项目下的异常反应指标, 将所有的反应模式都映射到规则空间中; 最后利用马氏距离与贝叶斯后验概率对被试进行分类, 并计算各个统计指标以评价分类准确性。

4.3 研究结果

表 6 是 16 种测验情景下规则空间方法的分类准确性统计表, 每一个单元格中的统计量值都是 30 次模拟结果的平均, 以此降低误差。

4.3.1 模式判准率比较 从图 3 与表 6 可以看到, 发散型、收敛型、线型与无结构型的模式判准率都会随着“失误”概率的增加而降低。当“失误”概率达到 10% 与 15% 时, 模式判准率都有降低, 例如收敛型与线型的判准率降低到了 0.65 左右, 发散型的模式判准率降低到了 0.25 左右, 无结构型的模式判准率只有 0.15 左右; 在同一“失误”概率下, 四种属性

层级结构的模式判准率之间的差异是,收敛型与线型明显好于发散型与无结构型,例如收敛型与线型在“失误”概率为 2%时,模式判准率分别为 0.927 与

0.957,而发散型与无结构型只有 0.461 与 0.404;同时还可以看到,收敛型与线型之间的准确性接近,发散型与无结构型之间有差异。

表 6 规则空间方法统计指标结果

属性层级结构	分类指标	失误概率			
		2%	5%	10%	15%
发散型(25 题)	模式判准率	0.461	0.361	0.286	0.244
	被试属性判准率	0.815	0.785	0.762	0.751
	敏感性	0.731	0.649	0.592	0.590
	特异性	0.794	0.747	0.704	0.696
收敛型(8 题)	模式判准率	0.927	0.844	0.681	0.547
	被试属性判准率	0.985	0.968	0.937	0.892
	敏感性	0.926	0.865	0.783	0.700
	特异性	0.942	0.883	0.812	0.804
线型(7 题)	模式判准率	0.957	0.882	0.789	0.644
	被试属性判准率	0.990	0.974	0.953	0.922
	敏感性	0.923	0.857	0.792	0.742
	特异性	0.931	0.869	0.831	0.795
无结构型(64 题)	模式判准率	0.404	0.211	0.177	0.159
	被试属性判准率	0.766	0.689	0.675	0.666
	敏感性	0.646	0.517	0.473	0.446
	特异性	0.749	0.652	0.606	0.579

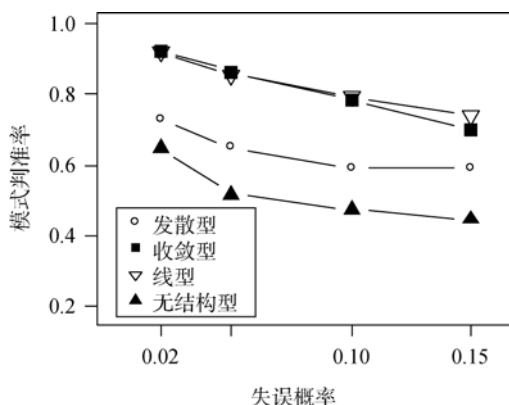


图 3 4 种结构模式判准率比较

4.3.2 被试属性判准率比较 从图 4 与表 6 可以看到,发散型、收敛型、线型与无结构型的被试属性判准率都会随着“失误”概率的增加而降低,但下降趋势不明显;在同一“失误”概率下,四种属性层级结构的被试属性判准率有较大差异,收敛型与线型的结果明显好于发散型与无结构型,例如,收敛型与线型在“失误”概率为 2%时,被试属性判准率分别为 0.985 与 0.990,而发散型与无结构型却只有 0.815 与 0.766,而且收敛型与线型之间的被试属性判准率接近,发散型与无结构型之间有一些差异;虽然发散型与无结构型的模式判准率不高,但被试属性判准率相对较高,例如发散型在四种“失误”概

率下的被试属性判准率分别是 0.815、0.785、0.762 与 0.751,无结构型的分别是 0.766、0.689、0.675 与 0.666。被试属性判准率表明,就每一个被试而言,仍然可以有 70%-80%的属性诊断准确。无结构型的被试属性判准率最低。

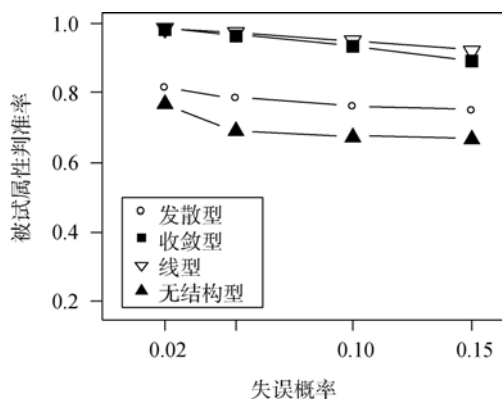


图 4 4 种结构被试属性判准率比较

4.3.3 敏感性与特异性比较 敏感性反映了对“未掌握属性”决策时的敏感程度,特异性反映了对“掌握属性”决策时的特异程度。从图 5 与图 6 可以看到,敏感性与特异性都随着“失误”概率的增加而降低;同一“失误”概率下收敛型与线型的敏感性与特异性高于发散型与无结构型;4 种属性层级结构下,4 种“失误”概率对应的特异性指标高于敏感性指标,

表明规则空间对高能力被试的诊断准确性高于低能力被试。

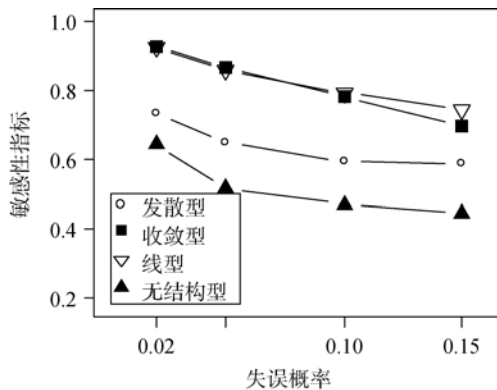


图 5 4 种结构敏感性指标比较

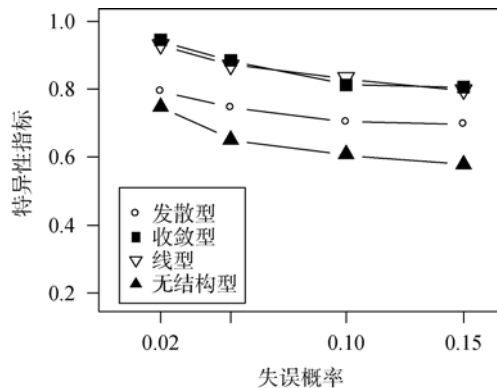


图 6 4 种结构特异性指标比较

综上所述, 线型分类准确性最好, 其次是收敛型, 再次是发散型, 最后是无结构型。而且线型与收敛型的分类准确性明显好于发散型与无结构型。

5 讨论

5.1 纯规则点分布对分类准确性的影响

从模拟研究结果可以看到, 线型与收敛型的模式判准率明显好于发散型与无结构型, 而且线型与收敛型之间没有明显差异, 发散型与无结构型之间有明显差异, 这与规则空间方法的分类思想有关。

规则空间方法是借助于马氏距离对被试进行分类的几何方法。在对被试进行分类之前, 必须有一些界定良好且相互排斥的组别。规则空间方法中的组别就是理想反应模式对应的“纯规则点”。如果纯规则点之间距离接近, 那么不同纯规则点周围所形成的椭圆错误分布交集扩大, 类别与类别之间不能满足互相排斥的条件。这样恰好处在交集范围内的观察反应模式距离多个纯规则点之间的距离可能是近似的, 导致分类准确率下降。

图 7~图 10 分别是发散型、收敛型、线型与无结构型四种属性层级结构所对应的纯规则点分布, 图中每一个数字代表一个知识状态。可以看到线型结构所对应的纯规则点之间完全没有重叠, 收敛型结构中只有纯规则点 3 与 4 之间有重叠, 发散型结构纯规则点的重叠就明显增多, 而无结构型纯规则点的重叠现象就非常严重, 大多数纯规则点簇拥在一起, 难以分辨, 给统计模式分类带来了困难。

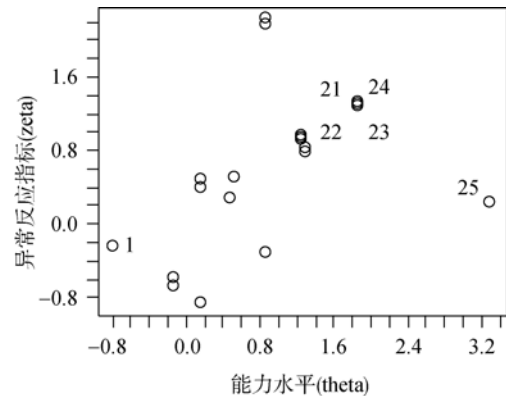


图 7 发散型纯规则点分布

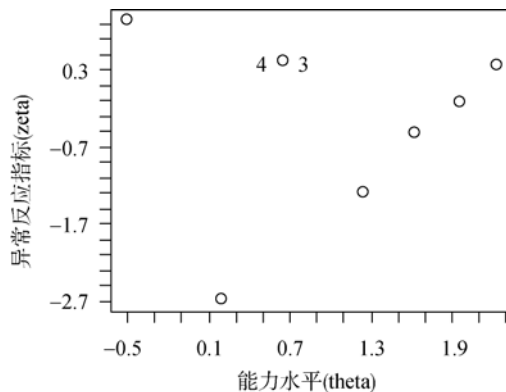


图 8 收敛型纯规则点分布

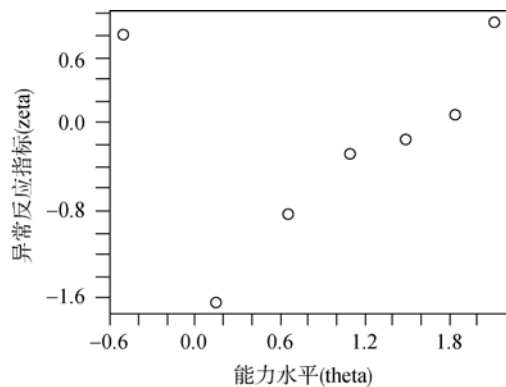


图 9 线型纯规则点分布

例如, 发散型结构的 KS21、KS22、KS23 与 KS24

四个纯规则点之间重叠严重(如图 7 所示)。从图 11~图 14 可以看到,在统计模式分类时,根据这 4 个理想反应模式所产生的观察反应模式的误判比例较大,多数被试都被分到了邻近的知识状态。相反 KS1 与 KS25 是完全独立的,从图 15 与图 16 可以看到,当纯规则点之间没有重叠时,在分类时几乎不会产生错误分类。

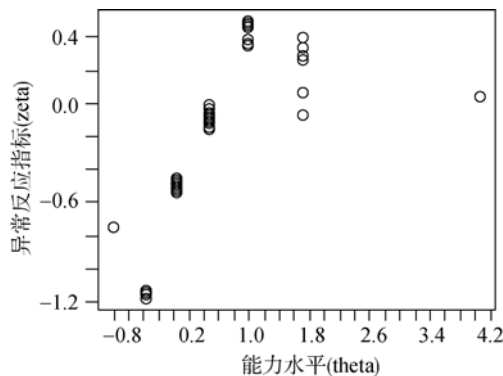


图 10 无结构型纯规则点分布

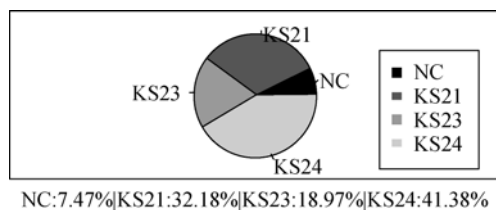


图 11 发散型结构 KS21 分类图

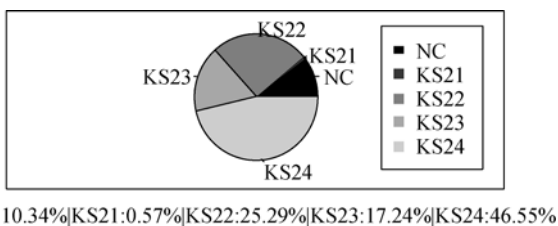


图 12 发散型结构 KS22 分类图

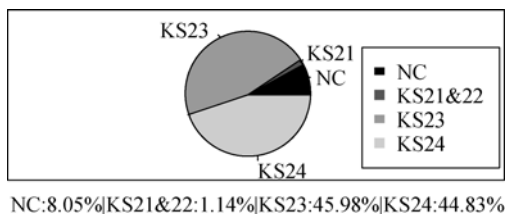


图 13 发散型结构 KS23 分类图

Varadi 与 Tatsuoka (1989, 1992)将测验分成不同的子项目集,计算被试在每一个子项目集上的异常反应指标 $\zeta_1, \zeta_2, \dots, \zeta_k$,在多维规则空间 $(\theta, \zeta_1, \zeta_2, \dots, \zeta_k)$ 中对被试进行分类,可以拉开纯规

则点之间的距离(Tatsuoka, 1995, 2009)。

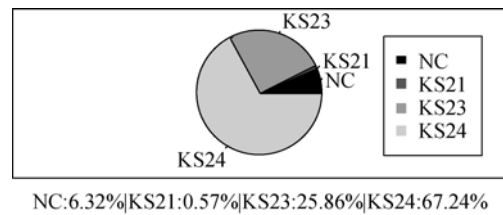


图 14 发散型结构 KS24 分类图

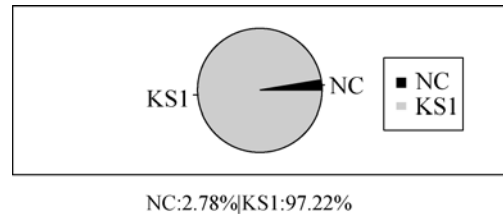


图 15 发散型结构 KS1 分类图

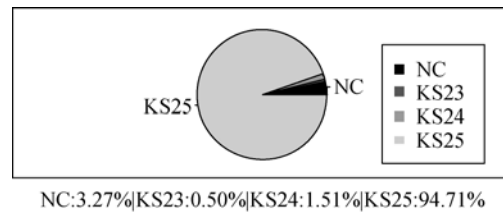


图 16 发散型结构 KS25 分类图

5.2 0-1 评分与多级评分项目的异常反应指标对照

在 0-1 评分项目下,所谓的异常反应就是被试在难题上答对,而在简单题上答错。反应模式异常程度的大小采用异常反应指标 $f(X) = \langle P(\theta) - X, P(\theta) - T(\theta) \rangle$ 来刻画,将 $f(X)$ 标准化之后就得到了 ζ , ζ 越大,表示多是难题做对而简单题做错,反应模式异常。对一个给定的 θ 来说, $f(X)$ 的期望为 0, 方差为 $\sum_{i=1}^n P_i(\theta) Q_i(\theta) [P_i(\theta) - T(\theta)]^2$, 与能力水平的协方差为 0 (Tatsuoka, 1985)。在多级评分项目下,异常反应指标为 $f(V) = \langle R(\theta) - V, R(\theta) - T(\theta) \rangle$, 同样采用项目残差的思想构建公式。可以证明,对于任意一个能力水平 θ , 所有反应模式对应的 $f(X)$ 的期望为 0, $f(V)$ 的方差为 $\sum_{i=1}^n [R_i(\theta) - t(\theta)]^2 * [\sum_{j=1}^{m_i} v_j^2 P_{ij}(\theta) - R_i^2(\theta)]$, 与能力水平的协方差为 0。通过比较我们可以发现,0-1 评分项目下的异常反应指标 $f(X)$ 是多级评分项目 $f(V)$ 的特例。因为,在 0-1 评分项目下,被试的反应只有 0 或 1, 每一个项目的最高分是 1, 此时, $R(\theta) = P(\theta)$, $V = X$, $T(\theta) = T(\theta)$, $\sum_{i=1}^n [R_i(\theta) - t(\theta)]^2 * [\sum_{j=1}^{m_i} v_j^2 P_{ij}(\theta) - R_i^2(\theta)] = \sum_{i=1}^n P_i(\theta) Q_i(\theta) [P_i(\theta) - T(\theta)]^2$,

5.4 多级规则空间方法的意义以及模拟研究的局限性和未来研究

随着教育标准和教育问责 (educational accountability) 的推行, 美国教育部 2001 年出台的“不让一个孩子落后”法案的实施细则里明确规定, 各州的学业评估系统必须包括对学生个体的解释、描述和诊断报告等, 让家长、老师和校长都能理解并致力于满足具体的学业标准 (Stout, 2002)。在这样的政策背景下, 辅助教学和学习过程的形成性评估 (formative assessment) 开始出现在教育评估的实践中, 帮助教育者和政策制定者确定需要关注的学生群体、学校甚至学生个体, 促进教育质量的提高。Bennett 和 Gitomer (2008) 提出了在 K-12 评估中整合教育问责制、形成性评估和专业辅导的模型框架, 充分发挥了认知诊断评估在帮助学生进步中的作用。随着我国教育改革的深化, 国家中长期教育改革和发展规划纲要 (2010~2020) 提出, 要着眼于促进教育公平, 提高教育质量, 认知诊断理论的发展与完善也将会为我国的教育改革提供支持。

认知模型与心理计量学模型是认知诊断评估的两个基本要素。进行认知诊断评估最理想的情况是根据认知模型开发一个认知诊断测验, 然而, 教育测量中最普遍的情况是, 有很多开发之初并非用于诊断目的的学业成就测验, 这些测验也是通过测量知识与技能来考察学生的能力, 如果能够充分挖掘学业成就测验中的信息, 对于学生学习与教师教学都有重要意义, 很多对现有成就测验的诊断分析都是基于规则空间方法的思想 (如余嘉元, 1995; 戴海崎, 张青华, 2004; 李峰, 余娜, 辛涛, 2009; Buck & Tatsuoka, 1998; Tatsuoka, Corter, & Tatsuoka, 2004)。

目前, 规则空间方法的认知诊断分析都限于 0-1 评分项目, 通过本研究规则空间方法的理论拓展, 以后可以将规则空间方法应用到多级评分测验, 甚至是 0-1 评分与多级评分项目都有的混合测验, 扩大了规则空间方法的应用范围, 使得对综合分析题、应用题、作文题等考核认知能力层次相对较高 (如分析、综合、应用、评价) 的多级评分题型的诊断分析成为可能, 充分挖掘测验中所包含的信息 (涂冬波等, 2010)。特别是对于混合测验, 我们首先对测验 Q 矩阵进行分块计算理想反应模式, 然后应用拓展的异常反应指标就可以同时将 0-1 与多级评分项目反应同时映射到规则空间中, 从而利用马氏距离与贝叶斯后验概率对被试进行分类。

当然, 本研究的模拟方法也存在一些局限。模拟研究数据是在理想反应模式的基础上随机发生“失误”作答之后得到的观察反应模式。这样有可能使得被试的知识结构产生改变, 从一种理想反应模式改变到另外一种理想反应模式, 从而降低规则空间方法的分类准确性。例如, 表 2 中的属性掌握模式 $\alpha_3 = (1 \ 1 \ 1 \ 0)$, 它的理想反应模式 IRP 为 $(1 \ 2 \ 3 \ 3)$, 如果第 4 题发生猜测作答, 观察反应模式为 $(1 \ 2 \ 3 \ 4)$, 那么从 α_3 得到的观察反应模式将被分类到 $\alpha_4 = (1 \ 1 \ 1 \ 1)$, 同理, 如果第 3 题和第 4 题都发生失误作答, 那么观察反应模式为 $(1 \ 2 \ 2 \ 2)$, 那么从 α_3 得到的观察反应模式将被分类到 $\alpha_2 = (1 \ 1 \ 0 \ 0)$ 。另外, 被试在一个题目上的作答, 有时不仅在 1 个属性上产生随机“失误”, 有可能是两个或多个属性上产生作答“失误”, 后续的模拟研究设计应该加以完善以更贴合测验中学生的作答反应。

由于规则空间方法是建立在 IRT 模型基础之上, IRT 参数估计的精确性会影响 θ 与 $I(\theta)$, 从而对规则空间方法的分类准确性造成一定的影响。模型与数据之间的拟合检验是保证参数估计精确性的一个重要前提, 限于本研究的数据分析工具 MULTILOG 7.03 在结果中没有给出相应的 χ^2 检验, 我们没有对 IRT 模型的拟合进行评估, 希望在后续的研究中能弥补这一不足之处。其实, IRT 模型的拟合检验是一个综合过程, 包括维度检验、局部独立性检验、模型与数据拟合检验, 限于本研究仅是一个模拟研究, 我们将在后续研究中通过一个实证数据, 对 IRT 模型拟合以及规则空间方法分类的信效度进行综合检验。

6 研究结论

本研究基于有序难度量表定义了多级评分项目下的异常反应模式, 并构建了异常反应指标公式, 将规则空间方法从 0-1 评分项目扩展到了多级评分项目, 并通过模拟研究检验了多级评分项目下, 规则空间方法的分类准确性。本研究的结论主要有:

(1) 多级评分项目下的异常反应指标 $f(V)$ 是基于项目残差建立, 对于任意一个给定的能力水平 θ , 所有反应模式对应的 $f(V)$ 期望为 0, 方差为 $\sum_{i=1}^n [R_i(\theta) - t(\theta)]^2 * \left[\sum_{j=1}^{m_i} v_j^2 P_{ij}(\theta) - R_i^2(\theta) \right]$, 且与能力水平 θ 的协方差为 0, $f(V)$ 越大, 表示被试的反应模式越异常, 多是难题得分比率高而简单题得分比率低;

(2) 多级评分项目下, 规则空间方法的分类准确性随着“失误”概率的增大而降低, 线型和收敛型的分类准确性好于发散型与无结构型;

(3) 纯规则点的分布对规则空间方法的分类准确性有显著影响。当纯规则点之间的距离分散时, 分类准确性高, 随着纯规则点之间重叠度的增加, 分类准确性也将降低。

参 考 文 献

- Bennett, R. E., & Gitomer, D. H. (2008). *Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support (research memorandum)*. Princeton, NJ: Educational Testing Service.
- Bolt, D. (2007). The present and future of IRT-Based cognitive diagnostic models (ICDMs) and related methods. *Journal of Educational Measurement*, 44(4), 377-383.
- Bolt, D., & Fu, J. B. (2004). A polytomous extension of the fusion model and its bayesian parameter estimation. Paper presented at NCME, San Diego, USA.
- Buck, G., & Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, 15(2), 119-157.
- Dai, H. Q., & Zhang, Q. H. (2001). An applied research of the rule space model in the identification of statistic study pattern. *Psychological Science*, 27(4), 949-951.
- [戴海崎, 张青华. (2001). 规则空间模型在描述统计学习模式识别中的应用研究. *心理科学*, 27(4), 949-951.]
- Ding, S. L., Wang, W. Y., & Yang, S. Q. (2009). A principle of construction of cognitive diagnostic test. <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*.]
- Ding, S. L., Zhu, Y. F., Lin, H. J., & Cai, Y. (2009). Modification of Tatsuoka's Q matrix theory. *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.]
- Embretson, S. E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics*. Orlando, FL: Academic Press.
- Frederiksen, N., Glaser, R. L., Lesgold, A. M., & Shafto, M. G. (1990). *Diagnostic monitoring of skills and knowledge acquisition*. Hillsdale, NJ: Erlbaum.
- Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (1993). *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.
- Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed.). New York: Academic Press.
- Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development*. Mahwah, NJ: Erlbaum.
- Kim, S. (2001). Towards a statistical foundation in combining structures of decomposable graphical models (Research Report No. 01-2). Yusong-Gu, Taejon: Korea Advanced Institute of Science and Technology, Division of Applied Mathematics.
- 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.
- Li, F., Yu, N., & Xin, T. (2009). Development of diagnostic math test for grade 4 and grade 5 based on the rule space model. *Psychological Development and Education*, 25(3), 113-118.
- [李峰, 余娜, 辛涛. (2009). 小学四、五年级数学诊断性测验的编制——基于规则空间模型的方法. *心理发展与教育*, 25(3), 113-118.]
- Luo, H., Ding, S. L., Wang, W. Y., Yu, X. F., & Cao, H. Y. (2010). Attribute hierarchy method based on graded response model with different scoring-weight for attributes. *Acta Psychologica Sinica*, 42(4), 528-538.
- [罗欢, 丁树良, 汪文义, 喻晓锋, 曹慧媛. (2010). 属性不等权重的多级评分属性层级方法. *心理学报*, 42(4), 528-538.]
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107-135.
- Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.
- Riply, B. D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.
- Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: American Council on Education, Macmillan.
- Stout, W. (2002). Psychometrics: from practice to theory and back. *Psychometrika*, 67(4), 485-518.
- Tatsuoka, K. K. (1983). Rule Space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4), 345-354.
- Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics*, 10(1), 55-73.
- Tatsuoka, K. K. (1991). *Boolean algebra applied to determination of the universal set of misconception states* (ONR-Technical Report No. RR-91-44). Princeton, NJ: Educational Testing Services.
- 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, NY: Routledge.
- Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, 41(4), 901-926.
- Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics*, 7(3), 215-231.
- Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement*, 20(3), 221-230.
- Tu, D. B., Cai Y., Dai, H. Q., & Ding, S. L. (2010). A polytomous cognitive diagnosis model: P-DINA model. *Acta Psychologica Sinica*, 42(10), 1011-1020.
- [涂冬波, 蔡艳, 戴海琦, 丁树良. (2010). 一种多级评分的认知诊断模型: P-DINA 模型的开发. *心理学报*, 42(10), 1011-1020.]
- Varadi, F., Tatsuoka, C., & Tatsuoka, K. K. (1992). *BUGLIB (Modified version)*. Unpublished computer program, Trenton, NJ.
- Varadi, F., & Tatsuoka, K. K. (1989). *BUGLIB*. Unpublished computer program, Trenton, NJ.
- Xin, T., & Jiao, L. Y. (2006). A new perspective for testing theory: The rule space model. *Journal of East China Normal University (Educational Sciences)*, 24(3), 50-56.
- [辛涛, 焦丽亚. (2006). 测量理论的新进展: 规则空间模型. *华东师范大学学报 (教育科学版)*, 24(3), 50-56.]
- Yu, J. Y. (1995). Application of rule space model for identifying cognitive errors in problem solving. *Acta Psychologica Sinica*, 27(2), 196-203.
- [余嘉元. (1995). 运用规则空间模型识别解题中的认知错误. *心理学报*, 27(2), 196-203.]
- Zhu, Y. F., & Ding, S. L. (2009). A polytomous extension of attribute hierarchy method based on graded response model. *Acta Psychologica Sinica*, 41(3), 267-275.
- [祝玉芳, 丁树良. (2009). 基于等级反应模型的属性层级方法. *心理学报*, 41(3), 267-275.]

A Polytomous Extension of Rule Space Method Based on Graded Response Model

TIAN Wei; XIN Tao

(Institute of Developmental Psychology, Beijing Normal University, Beijing 100875, China)

Abstract

Polytomous item has become an important item type in many large scale educational assessments. However, cognitive diagnostic models are initially proposed for dichotomous data, which limits the application and development of the Cognitive Diagnostic Assessment (CDA). For example, the rule space methodology (RSM) is still limited to the dichotomous item. The article thus extended the dichotomous RSM to that of polytomous item based on Graded Response Model, focusing on the definition and formula of ideal item-response pattern (IRP) and caution index in the polytomous item case.

In this paper, the IRP, defined as the number of attributes in an item that could be correctly responded by examinees without slippage or guessing, can be obtained by matrix multiplication. The degree of aberrant response, which indicated higher ratio score on hard items and lower ratio score on easy items, was described by the caution index, $f(V) = \langle R(\theta) - V, R(\theta) - T(\theta) \rangle$. The expectation of $f(V)$ was zero, its variance was

$\sum_{i=1}^n [R_i(\theta) - t(\theta)]^2 * \left[\sum_{j=1}^{m_i} v_j^2 P_{ij}(\theta) - R_i^2(\theta) \right]$, and it was orthogonal to the ability of θ at a given level of ability.

In the simulation study, Monte-Carlo experiment was used to test the classification accuracy under 4(attribute hierarchy)×4(slippage probability) conditions, indicated by four indexes, the pattern ratio, marginal ratio, sensitivity and specificity. Under each of 16 situations, the IRP was firstly ranked by raw score in ascending order, and the distribution of IRP in the population was made to approximate a normal distribution. A sample of 5000 observed response patterns was generated from the ideal item-response patterns, by randomly adding slips to each component of the ideal item-response patterns. For example, in order to generate 2% random errors, a random number was produced from the distribution of $U(0,1)$. For each item, if the random number was less than 0.01, and its ideal item score was not zero, then 1 was subtracted from it, otherwise 1 was added; if random number was higher than 0.99, and its ideal item score was not full, then the 1 was added to it, otherwise 1 was subtracted. MULTILOG 7.03 was applied to estimate the item and person parameters.

The results showed that: (1) the caution index for the polytomous item was constructed properly, and it could be interpreted with semantic meaning; the caution index for the dichotomous item was a special case for that of polytomous item; (2) the accuracy of classification under each of 4 hierarchy structures decreased as the slippage probability increased; (3) the classification accuracy associated with the linear and convergent hierarchy was obviously better than that of the divergent and unstructured hierarchy. (4) The distribution of the dots mapped from the ideal item-response patterns has an effect on the classification accuracy, that is, the overlap of the dots could reduce the classification accuracy rate.

Key words polytomous item; rule space methodology; graded response model; caution index; ideal item-response pattern