文章编号:1000-5862(2016)01-0047-09

基于属性层级关系的 rRUM 模型优化
——模型解释力及判准率的提升视角

蔡 艳,涂冬波
(江西师范大学心理学院,江西省心理与认知科学重点实验室,江西 南昌 330022)

摘要: 以提高认知诊断模型判准率及对数据的解释力为视角,对当前应用较广泛的 rRUM 模型进行优化(优化后的模型简记为 rRUM-AH),并采用 Monte Carlo 模拟研究及实证研究相结合的范式,比较分析了传统的 rRUM 模型和 rRUM-AH 模型的诊断正确率及诊断结果的解释力. 研究结果表明:当属性间存在层级关系时,不论在何种实验设计条件下,优化后的 rRUM-AH 模型属性诊断正确率远远高于传统的 rRUM 模型;当属性间存在层级关系时,rRUM 模型的模式判准率平均不到 80%(而 rRUM-AH 模型平均高达 90% 以上),难于满足实际需求,此时实际应用者选用该研究新开发的模型是一个较好的选择.

关键词: 认知诊断模型;属性层级关系;rRUM 模型;属性判准率
中图分类号: B 841 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2016.01.09

0 引言

认知诊断技术是认知心理学与心理计量学(Psychometrics)相结合的产物,它不仅要对个体心理特质水平进行宏观层次评价,还对个体心理内部加工过程进行诊断,揭示个体的认知加工特点[1]. 然而,要实现对人的内部心理加工过程的测量、诊断、评估并不是易事,因为人们无法直接观察到个体大脑中的思维过程,而只能得到他们对于测验项目的解答结果. 为此心理测量学家和认知心理学家以心理测量学和认知心理学为基础,对认知诊断做了大量艰苦并具创造性的尝试和探索,将认知心理学研究成果直接纳入心理计量模型中,开发出具有诊断功能的心理计量模型,即认知诊断模型(Cognitive Diagnosis Models,CDMs),从而实现对被试内部心理加工过程的测量,进而提供认知诊断信息.

在认知诊断实践中,这种充分融入了认知变量的认知诊断模型(CDMs)是实现认知诊断的关键环节,但同时又是认知诊断工作中的难点所在. 认知诊断模型的好坏直接决定着诊断结果的可解释性、准确性及诊断的效率. 为此,国际上心理测量学者们开发了 60 多种认知诊断模型[2]. 比较有名的认知诊断模型有 RSM(rules space model)[3]、AHM(attribute hierarchy model)[4]、DINA(deterministic input noisy 'and' gate)[5] 模型、rRUM 模型(reduced reparameterized unified model)[6]、G-DINA(the generalized DINA)[7] 模型等. 每种模型各具特点,其中 rRUM 是目前被研究较多的认知诊断模型之一[8-11];该模型设置了基于 Q 矩阵的项目难度参数和属性区分度参数,即惩罚参数(the penalty parameter). 惩罚参数详细描述了被试缺乏项目测量某一属性对被试正确反应概率的影响,因而与其它模型(如 DINA 模型)相比,rRUM 能更深入细致探讨被试问题解决的加工过程[12];同时有研究表明该模型具有较高的诊断正确率[8]. 当前,rRUM 模型被大多学者和实践运用者所推崇[8-11]. 但在认知诊断实践中,由于 rRUM 模型并未充分考虑认知属性间的层级关系(Attribute Hierarchy Structure)[4],因而时常估计出不符合属性逻辑关系的知识状态(knowledge states,KSs),即属性掌握模式,从而进一步影响了该模型的可解释性及诊断正确率. 那么如何才能克服这种现象呢? 如何才能进一步优化 rRUM 模型呢? 这是本研究拟重点探讨的问题.

收稿日期:2015-09-17
基金项目:国家自然科学基金(31100756,31300876,31160203,31360237),江西省社会科学规划重点项目(13JY01),江西省教育科学规划(12YB088,13YB029),高等院校博士点基金(20123604120001)和江西师范大学青年英才培育资助计划资助项目.
作者简介:蔡 艳(1979-),女,江西宜春人,副教授,博士,主要从事心理统计与测量的研究.

本研究拟结合 J. P. Leighton 等[4]的属性层级关系,对当前 rRUM 模型进行优化(为描述方便,优化后的模型简记为 rRUM-AH 模型),以进一步提高 rRUM 模型的诊断结果的解释力并提高模型的诊断正确率,从而为更好地推动认知诊断为实践服务提供方法学支持.

1 rRUM 模型及其优化

1.1 rRUM 模型

rRUM 模型通过项目难度参数 π_i^* 及属性区分度参数 r_{ik}^* 来构建被试对项目的作答概率,其项目反应函数为

$$P(x_{ij}=1\mid\boldsymbol{\alpha}_i)=\pi_j^*\prod_{k=1}^{K}r_{jk}^{*\,q_{jk}(1-\alpha_{ik})},$$

其中 $\pi_i^*=\prod_{k=1}^{K}P(Y_{ijk}=1\mid\alpha_{jk}=1)^{q_{ik}}$ 为被试正确应用项目 i 所有属性的概率,被称为以 Q 矩阵为基础的项目难度参数,其值界于 0~1 之间,π_i^* 越大说明项目越容易,一个项目只有一个难度参数; $r_{ik}^*=P(Y_{ijk}=1\mid\alpha_{jk}=0)/P(Y_{ijk}=1\mid\alpha_{jk}=1)$ 为被试缺乏属性 k 与掌握属性 k 但都答对项目的概率比,它能反应属性 k 的重要性,若其值为 0.25,则说明被试掌握属性 k 答对该题的概率是未掌握属性 k 也答对该题的概率的 4 倍,即掌握属性 k 对答对该题很重要. r_{ik}^* 的值越小说明属性 k 越重要. 它被称为项目 i 属性 k 的区分度参数,其值界于 0~1 之间; r_{ik}^* 越小说明项目 i 的属性 k 在正确答对项目 i 上越重要,也即该属性越能区分开答对与答错该题的被试,属性 k 有高的区分度. 一个项目若有 K 个属性,则该项目有 K 个区分度参数. r_{ik}^* 参数也被称为惩罚参数(the penalty parameter),即被试缺乏一个属性则其答对概率需乘以一个 0~1 的小数,从而使答对概率更小,每缺乏一个属性则进行一次相应的惩罚.

在 rRUM 模中,每个项目有一个难度参数(π_i^*)和 K 个区分度参数(r_{ik}^*),对于一个好的项目而言,它应该是高 π_i^* 值和低 r_{ik}^* 值.

1.2 属性层级关系

J. P. Leighton 等根据大量认知心理学研究成果[13-14],认为认知属性不是独立操作,而是从属于一相互关联的网络,认知属性间可能存在一定的心理顺序、逻辑顺序或层级关系,由此提出属性层级模型(AHM),并用属性层级关系(attribute hierarchy)图来表征相关任务的认知加工模型. J. P. Leighton 等指出属性层级关系有 4 种基本类型,分别为线型、收敛型、分支型和无结构型,且这 4 种基本类型又可组合成更为复杂的网络层级关系(complex networks of hierarchies). 由于无结构型是分支型的一种特例,因此本文用独立型代替无结构型,见图 1.

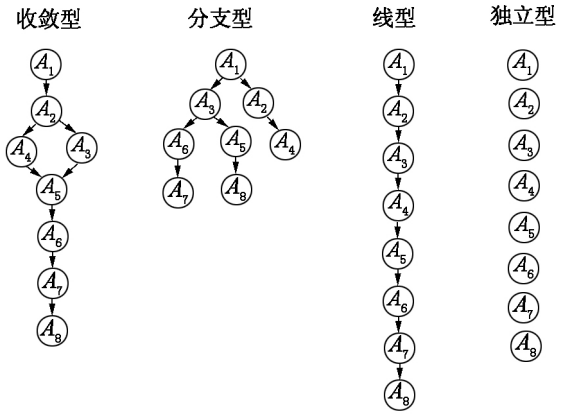

图 1 属性层级关系类型(8 个属性)

图 1 中,对于线性层级关系而言: 属性 A_1,A_2 和 A_3 是属性 A_4 的先决条件. 具体地说,属性 A_1 是属性 A_2 的先决条件,即被试只有先掌握了属性 A_1 才有可能掌握属性 A_2; 如果被试没有掌握属性 A_1,其它的属性是不可能掌握的. 对于有收敛的层级: 属性 A_2 是属性 A_3 和 A_4 的先决条件,但 A_3 和 A_4 同是属性 A_5 的先决条件(即只有同时掌握了 A_3 和 A_4 才能掌握 A_5). 而对于独立型,它是一种非结构化的层级,它呈现了另一个可能层级结构的极端; 在独立型关系中,属性 A_1 至 A_8 均无先决条件、没有次序,呈并列独立关系. 属性层级关系往往能被用来描述某一具体领域问题解决所需要的认知加工的整个次序及过程[4].

1.3 基于属性层级关系的 rRUM 模型优化: rRUM-AH 模型的开发

根据属性层级关系及 K. K. Tatsuka 的 Q 矩阵理论[3],可以得出所有可能的项目测量模式,这个对于认知诊断测验的项目设计及测验编制具有重要的指导意义,这一优点已被大多数研究者及实践者运用.

同时,根据属性层级关系也可以得到所有可能的符合逻辑的知识状态(KSs),即被试属性掌握模式. 可惜的是,一些知名的认知诊断模型(如 DINA 模型、rRUM 模型、G-DINA 模型等)在参数估计时并未充分利用这一信息,从而使这些模型估计的知识状态有可能不符合属性间的逻辑关系(如线性层级

关系出中估计出了“11111101”的知识状态),这大大降低了诊断结果的可解释性及诊断正确率; 反之,若在参数估计过程中能够充分利用属性阶层关系信息,则一方面可以大大缩减知识状态的全集空间以减少参数估计的复杂度,另一方面也有望提高模型的诊断结果的可解释性及诊断正确率. 以图 1 分支型层级关系为例(8 个属性): 如果在参数估计过程中忽略属性间的这种层级关系,那么所有可能的知识状态共有 $2^8 = 256$ 种; 若利用这种属性层级关系,则所有可能知识状态仅有 31 种,这就大大缩减了分类空间全集,简化了参数估计的复杂性.

根据以上分析,本研究拟结合属性层级关系来优化 rRUM 模型,具体优化的思路是根据属性间的层级关系及 Q 矩阵理论导致所有可能的知识状态,即属性掌握模式,然后将这一信息加入 rRUM 模型参数估计过程中: 若采用 MCMC 算法估计项目参数,则需将不符合属性层级关系的知识状态的先验概率设定为 0; 若采用 EM 算法或最大似然估计算法,则在所有知识状态中删除不符合属性层级关系的知识状态. 其余如项目反应函数、似然函数等均与 rRUM 完全一致. 这种优化的核心思想是通过属性层级关系来缩减诊断分类空间全集、简化参数估计的复杂性,从而进一步提高诊断结果的解释力及诊断正确率,当然本研究会通过大量的 Monte Carlo 模拟研究及实证研究来证实这一效果.

2　研究设计

2.1　研究方法及内容

为了进一步探讨优化后的 rRUM 模型(rRUM-AH)的效果及性能,并采用 Monte Carlo 模研究范式,比较分析传统的 rRUM 模型和 rRUM-AH 模型的诊断正确率及诊断结果的解释力. 为此本文具体开展了 2 项研究,涉及下列 3 个实验.

实验 1　不同属性阶层关系下,固定样本容量、测验长度和属性数的条件下,rRUM-AH 模型与 rRUM 模型的比较.

实验 2　不同属性阶层关系下,不同样本容量和测验长度条件下 rRUM-AH 模型与 rRUM 模型的比较.

实验 3　不同属性阶层关系下,不同属性数条件下 rRUM-AH 模型与 rRUM 模型的比较.

2.2　参数估计算法及收敛判断指标

目前国际上已开发 Arpeggio 软件程序[15] 专门用来估计 rRUM 模型参数,该软件采用马尔科夫链蒙特卡罗(Markov Chain Monte Carlo,MCMC)算法实现参数估计; 而对于本研究新开发的 rRUM-AH 模型,由于 Arpeggio 程序未考虑属性层级关系因此无法实现 rRUM-AH 模型的参数估计,因此本研究采用 Matlab 2012 语言自编程序实现 rRUM-AH 模型的参数估计. 同时为了保证结果的可比性,rRUM-AH 模型的参数估计同样采用 MCMC 算法. 且 Arpeggio 程序与自编程序的 MCMC 设置均为: 生成 3 条马尔科夫链,链长为 30 000,内燃值(burn-in)为 15 000; MCMC 参数估计收敛判断标准采用 A. Gelman 等[16] 的 $\hat{R}$ 统计量以及 Hartz[6] 采用的参数后验估计图(estimated posterior plot)和参数估计时间序列图(time series plots of parameters). 关于这些统计量的详细介绍感兴趣的读者可以参考文献[6]. 对于本研究中涉及的收敛指标,$\hat{R}$ 统计量均小于 1.2[16],同时参数后验估计图以及参数估计时间序列图均表明本研究中涉及的 MCMC 算法基本收敛,限于篇幅,本文未报告相关图,感兴趣读者可向笔者索要.

2.3　比较指标

采用属性边际判准率(Average Attribute Match Ratio,*AAMR*)、模式判准率(Pattern Match Ration,*PMR*)以及平均绝对离差(mean absolute bias,*MAB*)3 个比较指标. *AAMR* 和 *PMR* 指标用于评价模型的属性判准率,*MAB* 指标用于评价模型的项目参数估计精度:

$$PMR = \sum_{i=1}^{N} N_{i\_correct} / (N \times K),$$

$$AAMR = \sum_{i=1}^{N} \sum_{k=1}^{K} N_{ik\_correct} / (N \times K),$$

$$MAB(\tau) = \sum_{j=1}^{m} |\tau_j - \hat{\tau}_j| / m,$$

其中 N 为被试总数,$N_{i\_correct}$ 表示被试 i 的整个属性掌握模式是否判对,判对为 1,判错为 0; K 为属性个数,$N_{ik\_correct}$ 表示被试 i 的属性 k 是否判对,判对为 1,判错为 0. τ 和 $\hat{\tau}$ 为分别为项目参数的真值和估计值,m 为项目数.

3　Monte Carlo 模拟研究

3.1　实验 1

实验 1 主要是在不同属性层级关系下,固定样本容量、测验长度和属性数的实验条件比较 rRUM-

AH 模型与 rRUM 模型参数估计精度及诊断正确率.

3.1.1 实验 1 Monte Carlo 模拟过程 在实验 1 中,样本容量、测验长度和属性数分别固定为 1 000 人,30 题和 6 属性;实验 1 涉及的属性层级关系有线型、分支型、收敛型和独立型 4 种,即图 1 中的 4 种属性层级关系,但认知属性数只有前 6 个(即删除 A_7 到 A_8 属性),其 Monte Carlo 模拟过程如下文所述.

(i) 测验 Q 矩阵模拟. 根据 Q 矩阵理论,可知图 1 中(仅含前 6 个属性)独立型、分支型、收敛型及线型 4 种属性层级关系对应的所有可能的项目测量模式分别有 $2^6-1=63$、15、7 和 6 种. 为了模拟 30 道试题的测验 Q 矩阵,因此有些属性测量模式需要删除(如独立型),而有些属性测量模型需要重复模拟(如分支型、收敛型及线性),为此实验 1 中 4 种属性层级关系下模拟的 30 题测验 Q 矩阵固定如下所述,Q 矩阵设置中均考虑到测验 Q 矩阵中包含可达矩阵 $\boldsymbol{R}$ 阵:

$$\boldsymbol{Q}'_i=\begin{pmatrix}100000111110000000000111100000\\010000100001111000000100011100\\001000010001000111000010010011\\000100001000100100110001001010\\000010000100010010101000100101\\000001000010001001011111111111\end{pmatrix},$$

$$\boldsymbol{Q}'_d=\begin{pmatrix}111111111111111111111111111111\\010100111101111010100111101111\\001011111111111001011111111111\\000100000101101000100000101101\\000010010011011000010010011011\\000001001010111000001001010111\end{pmatrix},$$

$$\boldsymbol{Q}'_c=\begin{pmatrix}111111111111111111111111111111\\011111101111110111111011111101\\001011100101110010111001011100\\000111100011110001111000111100\\000011000001100000110000011000\\000001000000100000010000001000\end{pmatrix},$$

$$\boldsymbol{Q}'_l=\begin{pmatrix}111111111111111111111111111111\\011111011111011111011111011111\\001111001111001111001111001111\\000111000111000111000111000111\\000011000011000011000011000011\\000001000001000001000001000001\end{pmatrix},$$

其中 $\boldsymbol{Q}'_i$,$\boldsymbol{Q}'_d$,$\boldsymbol{Q}'_c$ 和 $\boldsymbol{Q}'_l$ 分别对应独立型、分支型、收敛型和线性的测验 Q 矩阵.

(ii) 被试知识状态(KSs)真值的模拟. 所有可能的知识状态则是在所有可能的项目测量模式中增加一种全为 0 的模式,即增加模式(000000),因此独立型、分支型、收敛型及线型 4 种属性层级关系对应的所有可能的知识状态分别有 64、16、8 和 7 种. 1 000名被试的知识状态真值则分别从 4 种属性层级关系对应所有可能的知识状态中随机产生.

(iii) 项目参数真值模拟. 项目参数真值从如下分布中随机产生: $\pi_j^* \sim U(0.85,0.98)$,$r_{jk}^* \sim U(0.01,0.3)$.

(iv) 被试作答反应矩阵模拟. 根据(i) ~(iii)步模拟的真值及 rRUM 模型项目反应函数计算被试答对项目概率 p,再产生一随机数 r,若 $p<r$ 则判被试得 0 分,否则得 1 分,从而模拟所有被试的得分矩阵.

(v) 对于模拟的数据分别采用 rRUM 模型和 rRUM-AH 模型进行参数估计,同时计算两模型属性判准率及项目参数估计精度.

(vi) 重复以上实验 30 次,以减少实验误差.

3.1.2 实验 1 结果 实验结果见表 1. 从表 1 可以看出,当属性层级关系为独立型时,rRUM 与 rRUM-AH 模型项目参数估计精度及属性诊断正确率相关无差,这个与先前的预期基本一致. 因为当属性间无层级关系时 rRUM-AH 模型并不能在 rRUM 模型的基础上简化知识状态分类空间全集,两者的知识状态分类空间全集均为 2^K,这时 rRUM-AH 模型等价于 rRUM-AH 模型.

当属性层级关系为分支型、收敛型和线型时,由于属性间存在一定的层级关系,知识状态全集不再是 2^K,这时 rRUM-AH 模型在参数估计过程中不同于 rRUM-AH 模型,前者充分考虑了属性间的层级关系. 实验结果也表明,当属性间存在层级关系时(如分支型、收敛型和线型),优化后的 rRUM-AH 模型的属性诊断正确率较传统的 rRUM 模型有大幅提升,尤其是模式判准率(PMR),其平均增幅高达 26.9%,最高增幅为 37.4%. 因此建议,当属性间存在层级关系时,研究者采用认知诊断模型进行诊断分析时,应充分利用属性间的这种关系,从而提升诊断结果的正确率,这也是本研究优化后的 rRUM-AH 模型能大幅度提高诊断正确率的重要原因. 而对于项目参数估计精度而言,两模型的表现无实质性差异,但两模型对项目参数估计的精度均较理想.

表 1　实验 1 结果

属性层级关系	评价指标	模型	
		rRUM	rRUM-AH
独立型	Mean *AAMR*	0.953	0.952
	Mean *PMR*	0.771	0.773
	Mean *MAB*(π_i^*)	0.018	0.018
	Mean *MAB*(r_{ik}^*)	0.027	0.026
分支型	Mean *AAMR*	0.956	0.980
	Mean *PMR*	0.781	0.892
	Mean *MAB*(π_i^*)	0.016	0.015
	Mean *MAB*(r_{ik}^*)	0.072	0.072
收敛型	Mean *AAMR*	0.916	0.991
	Mean *PMR*	0.575	0.949
	Mean *MAB*(π_i^*)	0.013	0.012
	Mean *MAB*(r_{ik}^*)	0.088	0.087
线型	Mean *AAMR*	0.925	0.993
	Mean *PMR*	0.639	0.960
	Mean *MAB*(π_i^*)	0.012	0.012
	Mean *MAB*(r_{ik}^*)	0.098	0.100

3.2　实验 2

实验 2 主要是基于不同属性阶层关系下,不同样本容量和测验长度条件下,比较 rRUM-AH 模型与 rRUM 模型诊断正确率.

3.2.1　实验 2Monte Carlo 模拟过程　由于实验 1 研究表明,独立型属性层级关系下 rRUM-AH 模型与 rRUM 模型无实质性差异,因此实验 2 中属性阶层关系只涉及分支型、收敛型和线型 3 种. 采用 3 × 3 × 2 实验设计,其中属性层级关系分别为分支型、收敛型和线型 3 个水平; 样本容量为 200 人、1 000 人和 3 000 人 3 个水平,分别代表小样本、中等样本和大样本; 测验长度为 30 题和 60 题 2 个水平. 但测验属性数与实验 1 一样固定为 6 个. 实验 2 的模拟过程与实验 1 基本一致. 当测验项目数为 30 题时,测验 Q 矩阵与实验 1 完全一致(即图 2) ; 当测验项目数为 60 题时,则是在实验 1 测验 Q 矩阵的基础上进行翻倍. 其余模拟过程与实验一样,但被试数不同.

3.2.2　实验 2 结果　表 2 是不同属性层级关系、不同样本容量及测验长度 3 种实验因素设计下,rRUM 与 rRUM-AH 2 个模型在各种实验处理中属性诊断正确率结果. 与实验 1 一样,实验 2 中两模型在项目参数估计精度中无实质性差异,因此为节省篇幅此处未详细列表报告(同理,实验 3 也未列出项目参数返真性结果) ,感兴趣者可向笔者索要.

表 2 说明不论在何处实验条件下,优化后的 rRUM-AH 模型的属性边际判准率(*AAMR*) 和属性模式判准率(*PMR*) 均明显优于 rRUM 模型. 尤其是属性模式判准率,分支型、收敛型和线性 3 种属性层级关系下,rRUM-AH 模型的模式判准率分别高出 rRUM 模型 12.5%、31.6% 和 30.4%,平均提高 25.0%,提升的幅度十分明显. 综合表 2 所有实验条件,rRUM 模型的模式判准率平均为 71.0%,而优化后的 rRUM-AH 模型的模式判准率平均高达 95.8%,这进一步说明当属性间存在层级关系时,传统的 rRUM 模型的属性诊断正确率不太理想,达不到实际要求(如模式判准率在 80% 以上) ,而本研究优化后的 rRUM-AH 模型属性诊断正确率则高达 95% 以上,较好地达到了实际要求,与 rRUM 相比能更好地为实际服务,因此这种新模型在实践应用中值得提倡及进一步推广.

表 2 实验 2 结果

属性层级关系	样本容量	测验长度	Mean *AAMR*		Mean *PMR*	
			rRUM	rRUM-AH	rRUM	rRUM-AH
分支型	200	30	0.962	0.977	0.808	0.879
		60	0.984	0.995	0.951	0.971
	1 000	30	0.956	0.980	0.781	0.892
		60	0.972	0.995	0.85	0.973
	3 000	30	0.938	0.979	0.718	0.890
		60	0.944	0.995	0.719	0.973
	Mean		0.959	0.987	0.805	0.930
收敛型	200	30	0.962	0.991	0.799	0.943
		60	0.989	0.999	0.932	0.993
	1 000	30	0.916	0.991	0.575	0.949
		60	0.962	0.998	0.795	0.991
	3 000	30	0.846	0.993	0.388	0.959
		60	0.877	0.999	0.438	0.992
	Mean		0.925	0.995	0.655	0.971
线型	200	30	0.96	0.99	0.786	0.940
		60	0.987	0.999	0.924	0.996
	1 000	30	0.925	0.993	0.639	0.96
		60	0.964	0.999	0.805	0.993
	3 000	30	0.821	0.994	0.409	0.962
		60	0.859	0.999	0.456	0.995
	Mean		0.919	0.996	0.670	0.974

3.3 实验 3

实验 3 主要是基于不同属性阶层关系下,不同属性数条件下,比较 rRUM-AH 模型与 rRUM 模型的诊断正确率.

3.3.1 实验 3Monte Carlo 模拟过程 实验 3 采用 3 × 4 双因素实验设计,涉及属性层级关系和认知属性数 2 个实验因素. 其中属性层级关系有 3 个水平(与实验 2 同),测验属性数有 4 个水平(为别 5、6、7 和 8 个属性). 与实验 1 类似,被试固定为 1 000 人,项目数固定为 30 题.

实验 3 与实验 1 的模拟过程基本似,但由于属性数不同,其涉及的测验 Q 矩阵不同. 但同样考虑 R 阵及不同属性层级关系下对应的所有可能的项目测量模式,实验 3 的测验 Q 矩阵为:

(i) 当 $K=5$ 时:

$$\boldsymbol{Q}'_l = \begin{pmatrix} 11111111111111111111111111111111 \\ 01111011110111101111011110111101111 \\ 00111001110011100111001110011100111 \\ 00011000110001100011000110001100011 \\ 00001000010000100000100001000010001 \end{pmatrix},$$

$$\boldsymbol{Q}'_d = \begin{pmatrix} 11111111111111111111111111111111 \\ 01010111101010111101010111101010 \\ 00101111100101111100101111100101 \\ 00010001100010001100010001100010 \\ 00001010100001010100001010100001 \end{pmatrix},$$

$$\boldsymbol{Q}'_c = \begin{pmatrix} 11111111111111111111111111111111 \\ 01111101111101111101111101111011 \\ 00101100101100101100101100101011 \\ 00011100011100011100011100011101 \\ 00000100000100000100000100000101 \end{pmatrix};$$

(ii) 当 $K=6$ 时,与实验 1 同;

(iii) 当 $K=7$ 时:

$$\boldsymbol{Q}'_{l}=\begin{pmatrix}1111111111111111111111111111111\\0111111011111101111110111111101\\0011111001111100111110011111100\\0001111000111100011110001111100\\0000111000011100001110000111100\\0000011000001100000110000011100\\0000001000000100000010000001000\end{pmatrix},$$

$$\boldsymbol{Q}'_{d}=\begin{pmatrix}1111111111111111111111111111111\\0101000111101110111110101000011\\0010111111111111111111001011111\\0001000000100110100110001000000\\0000100010010101011110000100001\\0000011001011011111111000001100\\0000001000001001101010000000100\end{pmatrix},$$

$$\boldsymbol{Q}'_{c}=\begin{pmatrix}1111111111111111111111111111111\\0111111011111110111111101111111\\0010111100101111001011110010111\\0001111100011111000111110001111\\0000111000001110000011100000111\\0000011000000110000001100000001\\0000001000000010000000100000000\end{pmatrix};$$

(iv) 当 $K=8$ 时:

$$\boldsymbol{Q}'_{l}=\begin{pmatrix}1111111111111111111111111111111\\0111111101111111101111111110111\\0010111110010111110010111111001\\0001111110001111110001111111000\\0000111100000111100000111100000\\0000011100000011100000011100000\\0000001100000001100000001100000\\0000000100000000100000000100000\end{pmatrix},$$

$$\boldsymbol{Q}'_{d}=\begin{pmatrix}1111111111111111111111111111111\\0101000011110111100110111111111\\0010111111111111111111111111111\\0001000000010001100110000101111\\0000100101001011011011111111111\\0000011000101100111101111111111\\0000001000000100010101001010101\\0000000100000001000101101001100\end{pmatrix},$$

$$\boldsymbol{Q}'_{c}=\begin{pmatrix}1111111111111111111111111111111\\0111111101111111101111111110111\\0010111110010111110010111111001\\0001111110001111110001111111000\\0000111100000111100000111100000\\0000011100000011100000011100000\\0000001100000001100000001100000\\0000000100000000100000000100000\end{pmatrix}.$$

3.3.2　实验 3 结果　表 3 是 3 种属性层级关系下,不同属性数下 rRUM 与 rRUM-AH 在各种实验处理中属性诊断正确率结果. 表 3 中,不论在何种实验条件下,优化后的 rRUM-AH 模型的属性边际判准率(*AAMR*)和属性模式判准率(*PMR*)仍均明显优于 rRUM 模型. 表 3 还表明,随着测验属性数的增加,rRUM 和 rRUM-AH 模型的诊断正确率均有下降趋势,这是因为属性数的增加势必增加待估计未知参数的个数,从而导致参数估计的复杂性及降低参数估计精度; 从表 3 还可以看出,当属性数增加到 8 个时,传统 rRUM 模型的模式判准率非常低,3 种属性层级关系下平均才 41.9%,远未达到实践中的需求(如 80% 以上); 而 8 个认知属性时,优化后的 rRUM-AH 模型的模式判准率平均高达仍 86.9%,高出 rRUM 模型 45%.

综上,实验 3 再次表明,当属性间存在层级关系时,即使当属性数较多时(如 8 个),优化后的 rRUM-AH 模型仍具有较理想的属性诊断正确率,能较好地满足实际需求,值得借鉴及推广.

4　研究总结与展望

本研究以提高认知诊断模型判准率及对数据的解释力为视角,对当前应用较广泛的 rRUM 模型进行优化,比较分析了优化后的 rRUM-AH 模型和传统的 rRUM 模型. 研究结果表明:

1) 当属性间存在层级关系时,不论在何种实验设计条件下(不同属性层级关系、不同样本容量、不同项目数和不同认知属性数),优化后的 rRUM-AH 模型属性诊断正确率远远高于传统的 rRUM 模型. 尤其是属性模式诊断正确率,优化后的 rRUM-AH 模型平均高出传统 rRUM 模型 28%,最大增幅高达 65% 以上(见表 3 收敛型 8 个认知属性实验处理). 与传统 rRUM 相比,本研究新开发的 rRUM-AH 模型具有更理想的诊断正确率.

2) 当属性间存在层级关系时,rRUM 模型的模式判准率平均不到 80%(优化后的 rRUM-AH 模型平均高达 90% 以上),难于满足实际需求,尤其当认知属性达到 8 个时,其模式判准率平均才 41.9%(优化后的 rRUM-AH 模型平均为 86.9%). 因此当属性间存在层级关系时,传统 rRUM 模型难于满足实际要求; 根据本研究结果,当属性间存在层级关系时,实际应用者选用本研究新开发的 rRUM-AH 模型是一个不错的选择.

表 3 实验 3 结果

属性层级关系	属性数	Mean *AAMR*		Mean *PMR*	
		rRUM	rRUM-AH	rRUM	rRUM-AH
分支型	5	0.942	0.988	0.742	0.944
	6	0.956	0.980	0.781	0.892
	7	0.949	0.979	0.721	0.871
	8	0.936	0.967	0.626	0.789
	Mean	0.946	0.979	0.718	0.874
收敛型	5	0.963	0.993	0.826	0.968
	6	0.916	0.991	0.575	0.949
	7	0.853	0.989	0.386	0.925
	8	0.786	0.987	0.244	0.898
	Mean	0.880	0.990	0.508	0.935
线型	5	0.944	0.995	0.759	0.977
	6	0.925	0.993	0.639	0.960
	7	0.890	0.992	0.470	0.945
	8	0.859	0.990	0.386	0.919
	Mean	0.905	0.993	0.564	0.950

总之,本研究在充分利用属性间层级关系的基础上,开发了一种新模型——rRUM-AH 模型,较好地克服了当前应用较广泛的 rRUM 模型的不足,在 rRUM 模型基础大大提高了属性诊断正确率及诊断结果的解释力,为更好地推动认知诊断在实践中的应用提供了新方法,值得借鉴和推广.

限于时间及精力本研究还有许多地方值得进一步深入:(i) 由于优化后的 rRUM-AH 模型充分利用了属性间层级关系,那么属性层级关系本身的合理性则会直接影响到 rRUM-AH 模型的诊断效果,而只有充分吸纳合理的属性层级关系时 rRUM-AH 模型的价值才能得到最大发挥. 因此,关于属性层级关系,未来有两方面问题值得解决探讨,一是属性层级关系不合理或有误时其对 rRUM-AH 模型的影响;另一是,在实际中有待开发出新的方法用于指导合理属性层级关系的构建. 虽然当前 Cui Ying 等[17] 开发的 HCI 指标(hierarchy consistency index) 可以用于属性层级关系合理性的检验,但这毕竟是事后分析,无法直接用于指导合理属性层级关系的构建. (ii) 有研究[18] 表明 Q 矩阵的好坏会直接影响到认知诊断的效果,未研究还可以进一步探讨 Q 矩阵的合理性对 rRUM-AH 模型的影响. (iii) 限于比较目的,本研究中 rRUM-AH 模型和 rRUM 模型的参数估计均是采用 MCMC 算法,未来还可进一步探讨 EM 算法来实现并比较两种模型的参数估计.

5 参考文献

[1] 涂冬波,蔡艳,丁树良. 认知诊断理论方法、方法与应用 [M]. 北京: 北京师范大学出版社,2012.

[2] Fu J,Li Y. Cognitively diagnostic psychometric models: an integrative review [R]. ETS Research Report,2008.

[3] Tatsuoka K K. Cognitive assessment: an introduction of the rule space method [M]. New York: Routledge,2009.

[4] Leighton J P,Gierl M J,Hunka S. The attribute hierarchy structure model: an approach for integrating cognitive theory with assessment practice [J]. Journal of Educational Measurement,2004,41: 205-236.

[5] Junker B W,Sijtsma K. Cognitive assessment models with few assumptions,and connections with nonparametric item response theory [J]. Applied Psychological Measurement,2001,25(3): 258-272.

[6] Hartz S M. A Bayesian framework for the unified model for assessing cognitive abilities: blending theory with practicality [D]. IL: University of Illinois at Urbana-Champaign,2002.

[7] dela Torre J. The generalized DINA model framework [J]. Psychometrika,2011,76(2): 179-199.

[8] Feng Yuling,Habing B T,Huebne A. Paramter estimation of the reduced RUM using the EM algorithm [J]. Applied Psychological Measurement,2014,38(2): 137-150.

[9] Henson R,Douglas J. Test construction for cognitive diag-

nosis [J]. Applied Psychological Measurement,2005,29(4):262-277.

[10] McGlohen M,Chang,Huahua. Combining computer adaptive testing technology with cognitively diagnostic assessment [J]. Behavior Research Methods,2008,40(3):808-821.

[11] Wang Changjiang,Chang Huahua,Huebner A. Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing [J]. Journal of Educational Measurement,2011,48(3):255-273.

[12] dela Torre J,Douglas J A. Higher order latent trait models for cognitive diagnosis [J]. Psychometrika,2004,69(3):333-353.

[13] Kuhn D. Why development does (and does not) occur: evidence from the domain of inductive reasoning [A] // McClelland J L,Siegler R. Mechanisms of cognitive development: Behavioral and neural perspectives [C]. Hillsdale,NJ: Erlbaum,2001:221-249.

[14] Vosniadou S,Brewer W F. Mental models of the earth: a study of conceptual change in childhood [J]. Cognitive Psychology,1992,24(4):535-585.

[15] DiBello L V,Stout W. Arpeggio documentation and analyst manual (Ver. 3. 1. 001) [Computer software]. St. Paul,MN: Assessment Systems Corporation.

[16] Gelman A,Carlin J B,Stern H S,et al. Bayesian data analysis [M]. London: Chapman & Hall Ltd,1995.

[17] Cui Ying,Leighton J P,Gierl M J,et al. A person-fit statistic for the attribute hierarchy method: the hierarchy consistency index [EB/OL]. [2015-09-12]. http://www.ualberta.ca/~mgierl/files/conferences/NCME06_YC.pdf.

[18] Rupp A A,Templin J L. The effects of *Q*-matrix misspecification on parameter estimates and classification accuracy in the DINA model [J]. Educational and Psychological Measurement,2008,68(1):78-96.

The Revision of the Reduced RUM Based on the Attribute Hierarchy

CAI Yan,TU Dongbo

(School of Psychology of Jiangxi Normal University,Lab of Psychology and Cognition Science of Jiangxi,Nanchang Jiangxi 330022,China)

Abstract: The Monte Carlo simulation method and empirical study were both used here to investigate the accuracy and the power of output explanation of diagnosis. Three simulation studies were conducted: Study 1 was based on the fix sample size,test length and the number of attributes. While Study 2 and Study 3 were based on the variable sample size and test length and the number of attributes respectively. Different attribute hierarchies were contained in all three simulation studies. At last the real data about syllogistic reasoning was chosen as an example to illustrate the application and comparison of the rRUM and the rRUM-AH. The findings showed that: If the relationship between attributes satisfied some hierarchy structure,then under any experiment design condition,the corrected ratios under rRUM-AH model were greater than those under rRUM model. When the relationship between attributes satisfied some hierarchy structure,the average of pattern corrected match ratios under rRUM model was less than 80 percent,otherwise,it under rRUM-AH model was greater than 90 percent. To satisfy the need in practice,the rRUM-AH model is a good choice for practitioner. Based on the analysis on the syllogistic reasoning data,the outputs indicated that the rRUM-AH model could fit the data better than rRUM model,and the power of result explanation under it was more reasonable.

Key words: cognitive diagnosis models; attribute hierarchy structure; the reduced RUM; classification accuracy

(责任编辑: 冉小晓)