

高阶项目反应模型的发展与应用

陈飞鹏¹ 詹沛达^{1,4} 王立君¹ 陈春晓² 蔡 毛³

(¹浙江师范大学心理系, 金华 321004) (²杭州第四中学, 杭州 310002)

(³杭州第九中学, 杭州 310020) (⁴北京师范大学认知神经科学与学习国家重点实验室, 北京 100875)

摘 要 在测量具有层阶结构的潜质时, 标准项目反应模型对项目参数估计和能力参数估计都具有较低的效率, 多维项目反应模型虽然在估计第一阶潜质时具有高效性, 但没有考虑到潜质层阶的情况, 所以它不适合用来处理具有层阶结构的潜质; 而高阶项目反应模型在处理这种具有层阶结构的潜质时, 不仅能够高效准确地对项目参数和能力参数进行估计, 而且还能同时获得高阶潜质与低阶潜质。目前存在的高阶项目反应模型有高阶 DINA 模型、高阶双参数正态肩型层阶模型、高阶逻辑斯蒂模型、多级评分的高阶项目反应模型和高阶题组模型。未来对高阶项目反应模型的研究方向应注意多水平高阶项目反应模型、项目内多维情况下的高阶项目反应模型以及高阶认知诊断模型。

关键词 项目反应理论; 高阶项目反应模型; 多维项目反应模型; 低阶潜质; 高阶潜质

分类号 B841

1 引言

经典测量理论(classical test theory, CTT)是在随机抽样理论基础上建立起来的一套心理与教育测量理论体系, 其所有数据分析的基础是被试在项目上作答的观察分数。CTT 认为观察分数等于真分数加上误差分数, 但在实际操作中真分数是无法获得的。CTT 的主要局限性是观察分数等权重性累加的不合理性, 测验对被试的评价依赖于测验的具体项目组合和项目数量, 测验与项目性能指标的估计依赖于具体的被试样本, 被试能力与项目难度两个指标含义的非统一性以及测量误差估计的不精确性和笼统性五个方面(罗照盛, 2012; 漆书青, 戴海崎, 丁树良, 2002)。20 世纪 60 年代以来, 心理测量学在克服 CTT 这些缺点的基础上提出了项目反应理论(item response theory, IRT), 项目反应模型(item response model, IRM)实现了在项目水平上对被试能力评估的模型化, 目前已成为心理测量学的核心内容。标准项目反应模型(standard item response model, SIRM)要求一

个测验只对一种潜质进行测量, 即需要满足“单维性假设”。但该假设与许多心理或教育测验的实际情况并不相符, 会给项目参数估计甚至能力参数估计带来严重的后果(Ansley & Forsyth, 1985)。为了满足一个测验测量多种能力的需要, 有学者(Whitely, 1980, 1981; Christoffersson, 1975; McDonald, 1985, 1997; Reckase & McKinley, 1982, 1991; Reckase, 1997)提出了多维项目反应模型(multidimensional item response model, MIRM), MIRM 能同时估计被试在多个维度上的能力, 并且考虑各能力维度之间的关系, 因此 MIRM 能更有效地估计多维能力。此外在实际测验中还有一些潜质具有层阶结构(Golay & Lecerf, 2011), 如: 韦氏成人智力量表(Wechsler Adult Intelligence Scale, WAIS)中就测量了 3 阶潜质: 第 1 阶中包含了 13 个子测验并分别测量一种潜质, 在第 2 阶中的 13 种潜质被归为 4 种外延更广的潜质(言语能力、知觉推理、工作记忆和信息加工速度), 而在第 3 阶中这 4 种潜质又包含在一般智力(general mental)之中(Ryan & Schnakenberg-Ott, 2003; Huang & Wang, 2013)。针对具有层阶结构的潜质, 传统的处理方法是分步采用 SIRM 或直接采用 MIRM 来处理。分步采用 SIRM 时, 需先将某测验分成若干个子测验,

收稿日期: 2014-01-07

通讯作者: 王立君, E-mail: frankwlj@163.com

然后多次采用 SIRM 来分析不同子测验的数据并得到相应的低阶潜质值,最后再用 SIRM 来分析整个测验数据以求得到高阶潜质值。但由于该方法忽略了各低阶潜质间的相关,所以当测验不够长或子测验数目过多时,就无法获得准确的参数估计值(Yao, 2010; Zhang, 2012; de la Torre & Hong, 2010)。另一种方法是直接采用 MIRM 来处理该情况,由于 MIRM 考虑到了子测验间的相关性,所以相对于分步采用 SIRM 能更有效地得到更为精准的低阶潜质估计值(Haberman & Sinharay, 2010; Yao, 2011; Yao, 2013; de la Torre, Song, & Hong, 2011; Yao & Boughton, 2009; Okan, 2013),进而可对被试做出更准确地诊断。但由于 MIRM 并没有考虑潜质的阶层结构,所以被试的高阶潜质值是无法被直接估计的,而高阶潜质恰是选拔性考试所更为强调的(Huang, Chen, & Wang, 2012)。为了更合理、有效地解决该问题,可直接处理具有层阶结构潜质的高阶项目反应模型(high order item response model, HO-IRM)得到了研究者的关注(de la Torre & Douglas, 2004; Yanyan & Winkle, 2008; de la Torre & Song, 2009; de la Torre & Hong, 2010; Huang & Wang, 2013; Huang, Wang, Chen & Su, 2013)。本文对 HO-IRM 涉及的基本概念和目前常用的模型进行了较为详细地介绍,有利于国内学者全面、清晰地了解 HO-IRM,为国内学者更好地应用 HO-IRM 来指导心理或教育测验的开发、编制及测验分析提供了理论参考。

2 高阶项目反应模型

2.1 HO-IRM 的基本概念

为了能合理、有效地处理具有层阶结构的潜质,HO-IRM 将 SIRM 和 MIRM 的优势相结合,用被试在各个子测验中的得分来分析不同的低阶潜质,且假设低阶潜质之间的相关性可由更高一阶的潜质来解释说明(de la Torre et al., 2009b)。HO-IRM 在分析具有层阶结构的潜质时具有明显的优势,它相对于 SIRM 来说,无论是估计高阶潜质还是估计低阶潜质都可以更高效地得到更精准的估计值(de la Torre et al., 2009b; Sheng et al., 2008)。而相对于 MIRM,HO-IRM 不仅能与 MIRM 一样高效地获得第 1 阶潜质的参数估计值,还能同时获得第 2 阶或更高阶潜质的参数估计值(Huang et al., 2013)。为了使读者更清晰地了解

HO-IRM,下文对目前已开发的 5 个模型进行了较为详细地介绍。

2.2 高阶 DINA 模型

目前,大部分认知诊断模型(cognitive diagnosis models, CDM)只报告了被试的微观认知状态,并没有关注潜在的更高阶的一般能力(general aptitude),而一般能力在解决问题过程中也是极其重要的。在众多的 CDM 中,DINA 模型(deterministic inputs, noisy “and” gate model) (Haertel, 1989, Junker & Sijtsman, 2001)因其简约性受到了研究者的广泛关注,该模型可描述为:

$$P(Y_{ij}=1|\alpha)=(1-s_j)^{\eta_{ij}}g_j^{1-\eta_{ij}} \quad (1)$$

η_{ij} 是描述被试 i 与项目 j 的关系,即被试 i 是否掌握项目 j 所考核的所有属性。 $\eta_{ij}=\prod_{k=1}^K\alpha_{ik}^{\eta_{jk}}$:若 $\eta_{ij}=1$,说明被试 i 掌握了项目 j 所考核的所有属性;若 $\eta_{ij}=0$,则说明被试 i 未掌握项目 j 所考核的所有属性。 $S_j=P(Y_{ij}=0|\eta_{ij}=1)$ 表示被试 i 在项目 j 上失误的概率,即被试 i 掌握了项目 j 所考核的所有属性,但答错的概率。 $g_j=P(Y_{ij}=1|\eta_{ij}=0)$ 表示被试 i 在项目 j 上猜对的概率,即被试未全部掌握项目 j 考核的所有属性,但答对的概率。

为了能够同时报告宏观的一般能力和微观的认知状态,de La Torre 和 Douglas (2004)以 DINA 模型为基础提出了高阶 DINA 模型(high order DINA; HO-DINA),该模型假设属性间相互独立且从属于一更高阶的一般能力,则认知属性与一般能力之间存在如下关系:

$$P(\alpha|\theta)=\prod_{k=1}^K P(\alpha_k|\theta) \quad (2)$$

$$P(\alpha_k|\theta)=\frac{\exp(\lambda_{0k}+\lambda'_k\theta)}{1+\exp(\lambda_{0k}+\lambda'_k\theta)} \quad (3)$$

在上面两个公式中 λ_{0k} 为属性 k 的截距, λ'_k 表示属性 k 在能力维度上的负荷。HO-DINA 模型建立在传统 DINA 模型基础上,并增加了比属性更高阶的能力参数,因此该模型不仅能描述被试的一般能力 θ ,还能描述被试的属性掌握情况以及属性与一般能力间的关系,为使用者提供更为丰富的诊断信息。实际应用该模型时,研究者应对 Q 矩阵及属性阶层关系进行界定,确定诊断所涉及的认知属性及属性间的逻辑关系;并进行项目设计及测验开发;最后需对所测试的实际数据

与 HO-DINA 模型的拟合度进行检验及评估。模型检验可使用 log-odds -ration (LOR, 对数差异比) 指标检验方法, 该指标主要是比较待检验的项目与其它所有项目所构成项目对的观察 LOR 指标与期望 LOR 指标间的差异, 若所有项目对观测的 LOR 与期望的 LOR 的平均绝对差异足够小, 则说明该项目资料模型拟合, 否则不拟合。涂冬波、蔡艳、戴海琦和丁树良(2011)的研究发现: 诊断的属性个数越多, 诊断的模式正确率越低, 而诊断的项目数越多, 诊断的模式正确率越高, 所以在实际应用中应根据实际情况来决定适当的项目数及属性数。

2.3 高阶双参数正态肩型层阶模型

虽然 HO-DINA 模型相对于传统 DINA 模型来说, 能够提供更为丰富的信息, 这也是第一个能同时估计高阶潜质和低阶潜质的项目反应模型。它虽然在第 2 阶上能处理连续潜质但在第 1 阶上只适合处理二分类别潜质(掌握或没掌握)。为了使 HO-IRM 能够处理两阶都为连续潜质的情况, Yanyan 和 Wikle (2008)提出了高阶双参数正态肩型模型(two-parameter normal ogive hierarchical model, 2PNOHM)。该模型对高阶潜质和低阶潜质作了更具体地限制, 即假设两者成线性关系:

$$\theta_{nv} = \beta_v \theta_n^{(2)} + \varepsilon_{nv}^{(1)} \quad (4)$$

上式中, n 指第 n 个被试, v 是指第 1 阶的第 v 个潜变量, 高阶潜质 $\theta_n^{(2)} \sim N(0,1)$, $-1 < \beta_v < 1$ 是指第 v 个低阶潜质对高阶潜质的回归系数, ε_{nv} 是指误差, 且各误差之间相互独立。2PNOHM 共有两阶, 第 1 阶就是洛德在 1952 年提出的两参数正态肩型模型(two-parameter normal ogive model, 2PNOM), 即

$$P(y_{vni}=1) = \phi(\alpha_{vi}\theta_{nv} - \gamma_{vi}) = \int_{-\infty}^{\alpha_{vi}\theta_{nv} - \gamma_{vi}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \quad (5)$$

上式中, y_{vni} 是指第 n 个被试对第 v 个子测验中第 i 个项目的二分反应, α_{vi} 和 γ_{vi} 分别指项目参数中的区分度和猜测度, $\theta_{nv}^{(1)}$ 为能力参数, 这里可指各个子测验所测的低阶潜质。将公式(4)代入上式即可得到 2PNOHM 的第 2 阶模型, 即:

$$P(y_{vni}=1) = \phi(\alpha_{vi}\beta_v\theta_n^{(2)} - \gamma_{vi} + \alpha_{vi}\varepsilon_{nv}^{(1)}) \quad (6)$$

上式中, n 指第 n 个被试, v 是指第 1 阶的第 v 个潜变量, i 指测验中第 i 个项目, α_{vi} 和 γ_{vi} 分别为

区分度和猜测度, $\theta_n^{(2)}$ 为高阶潜质, ε_{nv} 为误差且 $\varepsilon_{nv} \sim N(0,1)$ 。2PNOHM 可以用来分析具有层阶结构的连续潜质, 这从很大程度上拓展了 HO-IRM 的应用, 使 HO-IRM 能更好地应用于实践。如: Yanyan 等人分别使用 2PNOHM 与 SIRM 分析成就测验中的大学基础学科考试(College Basic Academic Subjects Examination, CBASE)。研究结果表明 2PNOHM 相对于 SIRM 在高阶潜质和低阶潜质上都更具有高效性和准确性。该模型在分析 CBASE 时认为该测验由阅读和写作两个子测验组成, 此时英语成绩就可看做是 2PNOHM 中的第 2 阶潜质(高阶潜质), 而两个子测验对应的能力就可以看做是第 1 阶的各低阶潜质。但需要注意的是 2PNOHM 只能用来处理只有两个低阶潜质的情况, 而在实际中, 高阶潜质通常是由多个低阶潜质构成的。为了将 HO-IRM 应用于更广的领域, de la Torre 等人提出了高阶 logistic 模型; Huang 等人提出了高阶广义分部评分模型(High Order Generalized Partial Credit Model, HO-GPCM)及其它的高阶多级评分模型; Huang 等人提出了高阶题组反应模型(higher order testlet model, HTM)。下文将对这 3 种模型及其应用做出详细地介绍。

2.4 高阶 logistic 模型

2PNOHM 除了只能处理两个低阶潜质的不足之外, 还有以下两点不足: 首先, 2PNOHM 是基于 2PNOM 提出的, 而在实际应用中 2PNOM 没有 3 参数逻辑斯蒂模型的适用范围广; 其次, 2PNOHM 没有对高阶潜质与低阶潜质之间的相关系数(β_v)进行限制, 这导致低阶潜质与高阶潜质的边缘分布和各低阶潜质间的边缘分布不在同一量尺(same scale)上。为了使 HO-IRM 能够有更广泛地应用, de la Torre 等人在克服 2PNOHM 不足的基础上提出了高阶 3 参数逻辑斯蒂模型。这个模型不仅在 2PNOHM 的基础上增加了猜测系数(c_{iv}), 而且在假设高阶潜质与低阶潜质的回归系数为 β_v 的同时, 还假设子测验 v 与子测验 v' 之间的相关系数为 $\beta_v \times \beta_{v'}$, 这就可以确保所有低阶潜质与高阶潜质的边缘分布在同一量尺上, 还使该模型能够在高阶潜质具有多个低阶潜质的情况下高效地获取精确的参数估计值。该模型的第 1 阶为:

$$P_{niv} = c_{iv} + (1 - c_{iv}) \frac{\exp[\alpha_{iv}(\theta_{nv}^{(1)} - b_{iv})]}{1 + \exp[\alpha_{iv}(\theta_{nv}^{(1)} - b_{iv})]} \quad (7)$$

将公式(4)代入到上式, 就可以得到高阶 3 参

数逻辑斯蒂模型的第 2 阶项目反应函数：

$$P_{ninv} = c_{iv} + (1 - c_{iv}) \frac{\exp[a_{iv}(\beta_v \theta_n^{(2)} - b_{iv} + \varepsilon_{nv})]}{1 + \exp[a_{iv}(\beta_v \theta_n^{(2)} - b_{iv} + \varepsilon_{nv})]} \quad (8)$$

公式(7)和公式(8)中的下标 n , i 和 v 的意义同上, a_{iv} 为区分度, b_{iv} 为难度系数, c_{iv} 为猜测度, ε_{nv} 是指误差, 且 $\varepsilon_{nv} \sim N(0, 1 - \beta_v^2)$, $\theta_n^{(2)}$ 指高阶潜质, $\theta_{nv}^{(1)}$ 为低阶潜质, 且 $\theta_{nv}^{(1)} \sim N(0, 1)$ 。当 $c_{iv} = 0$ 时, 上式就变为高阶 2 参数逻辑斯蒂模型的第 2 阶; 当 $c_{iv} = 0, c_{iv} = 0, a_{iv} = 1$ 时, 上式就为高阶 1 参数逻辑斯蒂模型的第 2 阶。

当项目参数已知时, 能力参数和回归系数的分布不变, 但此时需要注意的是当测验只有两个子测验时, 两个子测验的回归系数 β_1 和 β_2 并不是唯一的, 这导致模型不确定性的同时还决定了这两个回归系数不能被分别估计。所以, 当 HO-IRM 只有两个低阶潜质时, 需要对回归系数进行额外的限制。当 HO-IRM 有 3 个低阶潜质时, 低阶潜质之间就有 3 个相关系数, 这正符合目前 HO-IRM 的公式适合估计 3 个回归系数的要求, 即 HO-IRM 最适合此类数据。当低阶潜质超过 3 个时, 线性模型就不再适合分析这种复杂数据。当项目参数和回归系数都已知时, 可以使用传统的参数估计方法(如最大似然估计)来获取准确的能力参数。为了进一步说明这种构建模型方式的有效性, de la Torre 等人使用高阶 logistic 模型对美国加州考试局(California Testing Bureau, CTB)所提供的 2255 名九年级学生的数学成绩进行了分析, 并与对应的 logistic 模型分析结果进行了比较。结果显示, 高阶 logistic 模型在估计低阶能力时相对于对应的单层模型具有更好的精确性和高效性, 且当子测验与其他子测验之间相关高时, 这种优势更明显。

2.5 高阶 IRT 模型的新进展——多级评分的 HO-IRM

高阶双参数正态肩型模型和高阶 logistic 模型将 HO-IRM 拓展到连续变量数据上, 但它们仅能处理二级评分数据, 而教育测量中经常出现多级评分的情况(如简答题和作文题等)。为了使 HO-IRM 能够处理多级评分项目测量层阶潜质的情况, Huang 等人(2013)提出了以多级评分模型为第 1 阶的高阶模型, 如广义分部评分模型(generalized partial credit model, GPCM) (Muraki, 1992) :

$$\log\left(\frac{P_{nijv}}{1 - P_{ni(j-1)v}}\right) = \alpha_{iv}\theta_{nv}^{(1)} - \delta_{ijv} \\ \equiv \alpha_{iv}\theta_{nv}^{(1)} - (\delta_{ijv} + \tau_{ijv}) \quad (9)$$

上式中, P_{nijv} 和 $P_{ni(j-1)v}$ 分别指被试 n 在测验 v 中的第 i 个项目获得 j 分和 $j-1$ 分的概率, δ_{ijv} 是指测验 v 中的第 i 个项目的第 j 步的项目参数, δ_{iv} 是指测验 v 中的第 i 个项目的整体难度参数。将公式(4)代入到上式可得到高阶广义分部评分模型第 2 阶的项目反应函数：

$$\log\left(\frac{P_{nijv}}{1 - P_{ni(j-1)v}}\right) = \alpha_{iv}(\beta_v \theta_{nv}^{(2)} + \varepsilon_{nv}^{(1)}) - (\delta_{iv} + \tau_{ijv}) \quad (10)$$

如果高阶模型是以分部评分模型(partial credit model, PCM) (Masters, 1982)为第 1 阶而建立的, 即 $\alpha_{iv} = 1$ 时, 就可把公式(10)转化为高阶分部评分模型第 2 阶的项目反应函数; 如果高阶模型是以评定量表模型(rating scale model, RSM) (Andrich, 1978)为第 1 阶而建立的, 即 $\alpha_{iv} = 1, \tau_{ijv} = \tau_{iv}$ 时, 公式(10)就转化为高阶评定量表模型第 2 阶的项目反应函数。

如果高阶模型是以等级反应模型(graded response model, GRM) (Samejima, 1969)为第 1 阶而建立的, 则所得到的高阶模型为高阶等级反应模型。该模型的第 1 阶即为 GRM：

$$\log\left(\frac{P_{nijv}^*}{1 - P_{nijv}^*}\right) = \alpha_{iv}\theta_{nv}^{(1)} - \eta_{ijv} \quad (11)$$

P_{nijv}^* 是被试 n 在测验 v 中的第 i 个项目获得 j 分的概率, η_{ijv} 是测验 v 中的第 i 个项目的第 j 个类别临界参数(category threshold parameter)。将公式(4)代入上式就可得到高阶等级反应模型第 2 阶的项目反应函数。

Huang 等人(2013)在 WinBUGS 软件中(Spiegelhalter, Thomas, & Best, 2003)使用 MCMC 算法不仅对多级评分的 HO-IRM 进行了准确的参数估计, 而且还对高阶潜质与低阶潜质之间是非线性关系的 IRM 以及各子测验符合不同模型的 IRM 进行了准确的参数估计。Huang 等人在分析台湾中学生升高中的基本能力测验(Basic Competence Tests, BCT)的成绩和病理性上网数据时, 详细地介绍了如何选择适合的 HO-IRM 以及如何选择低阶潜质与高阶潜质之间的关系。这可以使 HO-IRM

适用于更复杂的真实数据,如 CET-6 由听力、阅读、翻译与写作组成,其中听力与阅读属于二级评分,翻译与写作为多级评分,可以将听力、阅读、翻译和写作看做是第一阶的 4 个子潜质,整体测验则是处于第二阶的潜质。在使用 HO-IRM 分析此测验时可先分别假设子潜质与整体潜质之间呈线性关系和非线性关系(如一元二次方程关系),然后根据贝叶斯模型选择方法确定最适合分析此类数据的模型。

2.6 高阶题组模型

以上模型都是在 SIRM 的基础上建立的,也就是说需要满足标准项目反应模型的局部独立性假设。但在实际的心理或教育测验中经常不能满足这个假设,如出现题组(testlet)的情况(英语中的篇章阅读理解题型、数学中的分步证明题),而且题组可以比独立的选择题提供更为复杂的、逻辑关系更强的阅读材料,更容易实现对层阶潜质的测量。Huang 和 Wang (2013)将潜质的层阶概念和题组概念相结合,提出了高阶题组反应模型(higher-order testlet model, HTM)。因为 HTM 的复杂性,本文将先介绍题组模型中的逻辑斯蒂克贝叶斯题组模型(Logistic Bayesian Testlet Model) (LBTM; Bradlow et al., 1999; Wainer & Wang, 2000),让读者能够对题组模型有一定的了解,以便更好地理解 HTM。

2.6.1 逻辑斯蒂克贝叶斯题组模型

Bradlow 等(1999)提出的双参数逻辑斯蒂克贝叶斯题组模型(two-parameter logistic Bayesian testlet model, 2-PLBTM)实现了对题组项目间的相依性与被试能力进行分离的目的,Wainer 和 Wang (2000)将 2-PLBTM 拓展为 3 参数逻辑斯蒂克贝叶斯题组模型(three parameter Logistic Bayesian testlet model, 3-PLBTM)。3-PLBTM 除了在 2-PLBTM 的基础上引入下渐近线参数外,其更大的进步在于承认了同一被试在不同题组内受到的题组效应存在差异(詹沛达,王文中,王立君,2013)。此时

$$P(y_{ni}=1) = c_{iv} + (1-c_{iv}) \frac{\exp[a_{iv}(\theta_{nv} - b_{iv} + \gamma_{nd(i)v})]}{1 + \exp[a_{iv}(\theta_{nv} - b_{iv} + \gamma_{nd(i)v})]} \quad (12)$$

上式中 a_{iv} 为题组 v 中项目 i 的区分度,但其实质上与 SIRM 中的区分度概念并不相同,关于两种不同的详细介绍可参见詹沛达等(2013)一文。 b_{iv} 为项目 i 的难度, c_{iv} 为项目 i 下渐近线参数,

$\gamma_{nd(i)v}$ 为被试 n 在题组 $d(i)$ 上的题组效应参数(同一被试在题组内不同项目上共享的成分), $\gamma_{id(i)v} \sim N(0, \sigma_\gamma^2)$, 其中 σ_γ^2 反应了题组效应的大小,即 σ_γ^2 值越大,题组效应越大。当 $\sigma_\gamma^2 = 0$ 时,即 $\gamma_{id(i)v} = 0$ 时,3-PLBTM 等价于 3PLM; 当同一被试在所有题组内受到相同的题组效应,即 σ_γ^2 相同时,3-PLBTM 就变为 2-PLBTM。

对比上式与高阶 3 参数逻辑斯蒂模型第 2 阶的项目反应函数(公式 8),当公式(8)中的 $\beta_v = 1$ 时,公式(8)就等价于上式,且在进一步的代数运算上两者是相等的(Frank, 2009)。但这两个模型之间有着不同的意义,题组模型把特定子测验上的变异看做是被试与题组的相互作用造成的随机效应,而 HO-IRM 认为被试在各子测验成绩上的变异是由各子集中的固定成分造成的(de la Torre et al., 2009b)。

2.6.2 高阶题组模型

Huang 和 Wang (2013)以 3-PLBTM 为第 1 阶,将公式(4)代入到公式(12)得到三参数逻辑斯蒂克贝叶斯高阶题组模型的第 2 阶,即

$$P_{nijk} = c_{iv} + (1-c_{iv}) \times \frac{\exp[\alpha_{iv}(\beta_v \theta_n^{(2)} - b_{iv} + \varepsilon_{nv}^{(1)})]}{1 + \exp[\alpha_{iv}(\beta_v \theta_n^{(2)} - b_{iv} + \varepsilon_{nv}^{(1)})]} \quad (13)$$

上式中以 3PLM 为基础模型发展出来的 HTM,当基础模型为多级评分模型(如 GPCM)时,则:

$$\log \left(\frac{P_{nijk}}{P_{ni(j-1)v}} \right) = \alpha_{iv}(\beta_v \theta_n^{(2)} - \delta_{iv} - \tau_{jiv} + \varepsilon_{nv}^{(1)}) \quad (14)$$

方程(13)和方程(14)中各参数的意义与上文一致,上式即是以 GPCM 为基础模型建立的 HTM 的第 2 阶的项目反应函数,即广义分部评分高阶题组模型(generalized partial credit high order testlet model,),如果 HTM 是以 PCM 为基础模型而建立的,即 $\alpha_{iv} = 1$ 时,就可把上式转化为分部评分高阶题组模型(partial credit high order testlet, PC-HTM)第 2 阶的项目反应函数;如果 HTM 是以 RSM 为基础模型而建立的,即 $\alpha_{iv} = 1, \tau_{jiv} = \tau_{jv}$ 时,上式就转化为评定量表高阶题组模型(rating scale high order testlet model, RS-HTM)第 2 阶的项目反应函数。

Huang 等人(2013)在 Winbugs 中获得了 HTM 参数估计值,这可以让 HO-IRM 适用于具有更复杂结构的测验。他们分别使用 3 参数逻辑斯蒂模型所对应的 HTM 和高阶 3 参数逻辑斯蒂模型分析

了5000名台湾中学生升入高中时的BCT成绩,可能因为该测验的题组效应不强,这两个模型得到了相似的结果。但使用这两个模型分析题组效应强的病理性上网数据时,结果显示,当忽略题组效应时,会明显低估低阶潜质的因素负荷且会高估低阶潜质的信度。大陆的考试也会存在题组现象如高考的理科综合卷是由物理、化学和生物三部分构成,而每个部分又都是由独立项目和题组项目构成的。HO-IRM可以将三门学科的项目分别看做测量第1阶的子潜质,而三个子潜质构成第2阶的潜质,即理科综合能力,HTM的应用就可以避免直接使用HO-IRM而忽略题组效应而带来的误差,从而更准确地分析此类数据。

3 总结与展望

3.1 总结

虽然国内外对HO-IRM的研究是近些年才兴起,但其在理论和应用上都已取得了较大的发展。首先,HO-IRM将传统的层阶因素分析方法应用到IRT框架中,实现了在IRT框架中分析被试的高阶潜质值和低阶潜质值,这既是对IRT的进一步拓展,也改善了传统方法中以分数代替能力的不足。其次,在测量具有层阶结构的潜质时,HO-IRM能够同时准确地估计高阶潜质和低阶潜质且具有高效性,这弥补了SIRT估计的低效性和MIRT不能直接估计高阶潜质的不足。本文主要介绍了5种HO-IRM,其中HO-DINA模型不仅是高阶认知诊断模型的代表,而且首先提出被试能否正确作答除了受具体的低阶潜质影响还受更高阶的一般能力影响,这为HO-IRM的发展提供了基本思路。2PNOHM将高阶潜质与低阶潜质之间的关系进行了具体化,即假设两者之间是线性关系,而后3种HO-IRM都是在这个假设前提发展出来的,它们将HO-IRM进一步拓展到更多领域。

虽然HO-IRM对IRT的发展具有重要的影响,但它也具有一定的不足:首先,它假设高阶潜质与低阶潜质是呈线性关系,但并不是所有的高阶潜质与低阶潜质都是线性关系,两者之间还可能非线性关系(Huang et al., 2013),在使用HO-IRM分析数据之前应先确定两者之间的关系,从而选取最适合的模型;其次,HO-IRM假设测验是由多个单维子测验构成的,即每个项目只能属于一个子测验,但处理真实数据时,有些项目经

常属于两个甚至更多子测验,因此,当所测潜质没有明确的层阶结构时,要谨慎使用HO-IRM。

3.2 展望

3.2.1 发展能够处理嵌套数据的HO-IRT

查阅现有的国内外文献,发现国内仅有涂冬波等人(2011)以及涂冬波,蔡艳和戴海琦(2013)两篇文献对HO-DINA模型有进一步的研究,而没有相关文献对其他HO-IRM做介绍,因此加强HO-IRM在国内的研究与应用,对于推动国内心理统计和测量的发展具有重要的意义。而国外对HO-IRM的研究也主要集中于它在分析具有层阶结构的潜质时相对于SIRM和MIRM的优势。这也就是说现有HO-IRM只是对所测潜质有了进一步的考虑,而在被试方面仍然假设在同一整体中随机选取被试。但在教育和心理测量中,选取的被试(学生)往往嵌套于班级之中,而班级又嵌套于学校之中。在处理嵌套数据时,通常会因为分组取样、同时受到外来的援助或干扰、群体间不同的学习机会或不同的问题解决策略等因素导致群聚效应(person clustering effect, PCE)。忽略群聚效应会带来不准确的参数估计,而且群聚效应会减小有效的样本大小(Cyr & Davies 2006)。为了让IRM能够处理这类数据,有研究者(Fox, 2010; Raudenbush & Bryk, 2002; Wang & Qiu, 2013)提出了多水平项目反应模型(multilevel item response model, MUIRM)来处理具有嵌套结构的数据。多水平IRM将项目反应理论和阶层线性模型相结合,是对项目反应理论的进一步拓展,也是对阶层线性模型的有效改进(刘慧, 简小珠, 张敏强, 熊悦欣, 2012)。但多水平IRM只是考虑了被试的嵌套问题并没有考虑所测潜质的层阶结构,所以进一步发展出既能处理所测潜质具有层阶结构,又考虑到被试嵌套情况的项目反应模型是HO-IRM接下来的一个研究方向。

3.2.2 高阶认知诊断模型以及项目间多维的高阶项目反应模型的发展

SIRM假设测验的所有项目只测量一种潜质,但通常需要评估的能力是由不同低阶潜质组成的,MIRM可以对这种情况下的能力值进行估计。有些测验是由多个子测验构成,每个子测验只测量一种潜质,这被称之为项目间多维(multidimensional between-item)。而有时会出现一个项目受多种潜质影响的情况,这被称之为项目内多维(multidimen-

sional within-item)。现有的 HO-IRM 仅能处理项目间多维的情况，而没有研究是针对项目内多维而提出的。所以进一步将 HO-IRM 拓展到项目内多维的情况可成为下一步的一个研究方向。

Huang 等人指出任何项目反应函数都可以发展成对应的高阶模型，即将公式(4)代入第 1 阶的项目反应函数即可得到对应的 HO-IRM 的第 2 阶。但这只是就项目间多维而提出的。而对于认知诊断模型而言，目前仅有 HO-DINA 模型，还没有开发出其他认知诊断模型对应的高阶模型。所以进一步开发高阶认知诊断模型可以成为将来的一个研究方向。

现有的研究在对低阶潜质进行估计时，充分考虑到各个子测验之间的相关性，从而提高测量的效率和精确度；但此时假设任意两子测验之间的相关都是相等的，而没有对各子测验之间相关不等的情况做进一步研究。已有研究使用贝叶斯模型选择指标来选取最适合的 HO-IRM，但对具体条件下使用何种指标并没有做出详细的研究，所以对 HO-IRM 的模型选择值得国内研究者关注。

参考文献

- 罗照盛. (2012). *项目反应理论基础*. 北京: 北京师范大学出版社.
- 漆书青, 戴海崎, 丁树良. (2002). *现代教育与心理测量学原理*. 北京: 高等教育出版社.
- 刘慧, 简小珠, 张敏强, 熊悦欣. (2012). 多水平 IRT 的发展与应用述评. *心理科学进展*, 20(4), 627-632.
- 涂冬波, 蔡艳, 戴海琦, 丁树良. (2011). HO-DINA 模型的 MCMC 参数估计及模型性能研究. *心理科学*, 34(6), 1476-1481.
- 涂冬波, 蔡艳, 戴海琦. (2013). 基于 HO-DINA 模型的多级评分认知诊断模型的开发. *心理科学*, 36(4), 984-988.
- 詹沛达, 王文中, 王立君. (2013). 项目反应理论新进展之题组反应理论. *心理科学进展*, 21(12), 2265-2280.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-573.
- Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9(1), 37-48.
- Bradlow, E. T., Wainer, H., & Wang, X. H. (1999). A bayesian random effects model for testlets. *Psychometrika*, 64(2), 153-168.
- Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40(1), 5-32.
- Cyr, A., & Davies, A. (2006). *Item Response Theory and Latent variable modeling for surveys with complex sampling design the case of the National Longitudinal Survey of Children and Youth in Canada*. Paper presented at the conference of the Federal Committee on Statistical Methodology, Office of Management and Budget, Arlington, VA.
- de La Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3), 333-353.
- de la Torre, J. (2009a). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement*, 33, 465-485.
- de la Torre, J., & Song, H. (2009b). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, 33(8), 620-639.
- de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size a Higher-Order IRT model approach. *Applied Psychological Measurement*, 34(4), 267-285.
- de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscore. *Applied Psychological Measurement*, 35(4), 296-316.
- Frank, R. (2009). *Three multidimensional models for Testlet-based tests: Formal relations and an empirical comparison*. Princeton, New Jersey: ETS.
- Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.
- Golay, P., & Lecerf, T. (2011). Orthogonal higher order structure and confirmatory factor analysis of the French Wechsler Adult Intelligence Scale (WAIS-III). *Psychological Assessment*, 23, 143-152.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 301-323.
- Haberman, S. J., & Sinharay, S. (2010). Reporting subscores using multidimensional item response theory. *Psychometrika*, 75(2), 209-227.
- Huang, H. Y., Chen, P. H., & Wang, W. C. (2012). Computerized adaptive testing using a class of high-order item response theory models. *Applied Psychological Measurement*, 36(8), 689-706.
- Huang, H. Y., & Wang, W. C. (2013). Higher order testlet response models for hierarchical latent traits and testlet-based items. *Educational and Psychological Measurement*, 73, 491-511.
- Huang, H. Y., Wang, W. C., Chen, P. H., & Su, C. M. (2013). Higher-order item response models for hierarchical latent traits. *Applied Psychological Measurement*, 37, 619-637.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258-272.
- Masters, G. N. (1982). A Rasch model for partial credit

- scoring. *Psychometrika*, 47, 149–174.
- McDonald, R. P. (1985). Factor analysis and related methods. *Applied Psychological Measurement*, 9(4), 435–437.
- McDonald, R. P. (1997). Normal-ogive multidimensional model. *Handbook of Modern Item Response Theory*, 257–269.
- Eiji Muraki (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Okan, B. (2013). *Between-person and within-person subscore reliability: comparison of unidimensional and multidimensional IRT models* (Unpublished doctoral dissertation). Conservancy. University of Minnesota, Education.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Reckase, M. D., & McKinley, R. L. (1982). *Some latent trait theory in a multidimensional latent space*. Iowa City, IA: American College Service.
- Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, 15(4), 361–373.
- Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25–36.
- Ryan, J. J., & Schnakenberg-Ott, S. D. (2003). Scoring reliability on the Wechsler Adult Intelligence Scale-Third Edition (WAIS- III). *Assessment*, 10(2), 151–159.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 35, 139.
- Spiegelhalter, D. J., Thomas, A., & Best, N. (2003). WinBUGS version 1.4 [Computer Program.]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.
- Wainer, H., & Wang, X. H. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, 37(3), 203–220.
- Wang, W.-C., & Qiu, X.-L. (2013). A multidimensional and multilevel extension of a random-effect approach to subjective judgment in rating scales. *Multivariate Behavioral Research*, 48, 398–427.
- Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45(4), 479–494.
- Whitely, S. E. (1981). Measuring aptitude processes with multicomponent latent trait models. *Journal of Educational Measurement*, 18, 67–84.
- Yao, L. H. & Boughton, K. A. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement*, 46, 177–197.
- Yao, L. H. (2010) Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement*, 47(3), 339–360.
- Yao, L. H. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement*, 35, 48–66.
- Yao, L. H. (2013). Multidimensional item response theory for score reporting. In Y. Chang & H. H. Cheng (Eds.), *Advances in Modern, International Testing: Transition from Summative to Formative Assessment*. Charlotte, NC: Information Age Publishing.
- Yanyan, S., & Wickle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, 68(3), 413–430.
- Zhang, J. M. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement*, 36(5), 375–398.

The Development and Application of Higher-Order Item Response Models

CHEN Feipeng¹; ZHAN Peida^{1,4}; WANG Lijun¹; CHEN Chunxiao²; CAO Mao³

(¹ Department of Psychology, Zhejiang Normal University, Jinhua 321004, China)

(² Hangzhou NO.4 High School, Hangzhou 310002, China) (³ Hangzhou NO.9 High School, Hangzhou 310020, China)

(⁴ National Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing 100875, China)

Abstract: When measuring hierarchical latent trait, the standard IRT models have low efficiency in estimating item parameters and ability parameters. Without considering hierarchical structure, MIRT models have a high efficiency in estimating the first-order latent traits. It is not suitable for dealing with hierarchical latent trait. The HO-IRM not only can produce more accurate item parameter and ability parameters estimates, but also can obtain the first-order and high-order latent traits. The currently-existing HO-IRM includes high-order DINA model, high-order two parameter normal ogive hierarchical model, high-order logistic model, high-order polytomous items model and higher-order testlet model. In the future HO-IRM study, attention must be paid to the multi-level HO-IRM, HO-IRM within an item in the multidimensional situation and high-order cognitive diagnostic models.

Key words: IRT; HO-IRM; MIRT; first-order latent traits; high-order latent traits