

混合 IRT 潜在模型及其应用轨迹*

王 霞¹ 谭国华² 王 旭³ 张敏强¹ 骆 聪¹

(¹华南师范大学心理学院/心理应用研究中心, 广州 510631)

(²广州市教育局教学研究室, 广州 510030) (³广州市教育信息中心, 广州 510030)

摘 要 项目反应理论是测量被试潜在特质的现代测量理论, 潜在类别分析是基于模型的潜在特质分类技术。混合项目反应理论将项目反应理论与潜在类别分析相结合, 能够同时对被试分类并量化其潜在特质。在阐述混合项目反应理论概念、原理的基础上, 介绍了 MRM、mNRM 和 mPCM 等几种常见混合模型及其参数估计方法, 并从心理与行为特征分类、项目功能差异检测、测验效度评价等方面评述了其在心理测验中的应用发展轨迹。

关键词 项目反应理论(IRT); 潜在类别分析(LCA); 混合 IRT; 潜在结构

分类号 B841

1 引言

项目反应理论(item response theory, IRT)是一种现代心理测量理论。较之经典测量理论, IRT 具备众多优势, 如项目与特质参数的不变性; 被试能力参数和项目难度参数的配套性; 对测量误差的精确估计等。近年来, IRT 越来越广泛地被应用于心理与教育测量, 它在指导测验编制、题库建设、计算机自适应测验和估计测验偏差等方面的贡献尤为突出。

IRT 的基本原理是通过建立数学模型来揭示被试作答反应与所测特质之间的关系, 从而实现对被试潜在特质的估计(漆书青, 戴海琦, 1992)。这种数学模型被称为项目反应模型。为了拟合实测数据资料, 多种项目反应模型相继被提出。随着理论发展和实践应用的需要, IRT 已不满足于在自身的框架内发展新的模型, 而转向与其它心理统计方法相结合。其中, IRT 与潜在类别分析

(latent class analysis, LCA)相结合形成的混合项目反应理论(mixture item response theory, 混合 IRT)就是心理测量与心理统计融合发展的一个新趋势。

LCA 是一种基于模型的聚类方法, 旨在将被试个体划分为组内尽可能同质而组间存在差异的几类或组(邱皓政, 2008)。混合 IRT 将 IRT 与 LCA 相结合, 意味着各潜在类别服从同一 IRT 模型, 但不同潜在类别的模型参数存在差异。混合 IRT 能够同时对被试分类并量化其能力特质。因此, 混合 IRT 是对 IRT 的进一步拓展, 也是对 LCA 的有效改进, 它的提出不仅弥补了 IRT 仅能处理连续潜变量和 LCA 仅能处理分类潜变量的不足, 也为心理、社会、医学等领域的研究者面对复杂数据时提供了一种新的思路和有利的分析工具。

混合 IRT 常被用于评估测量结构的多维性和检测被试作答风格的差异性。例如, 当一份心理测验的项目可使用不同策略进行解答时, 由于解题策略不同, 不同群组的项目参数将会呈现出差异(Glück, Machat, Jirasko, & Rollett, 2001; Bonnefon, Eid, Vautier, & Jmel, 2008)。混合 IRT 已被用于检验 IRT 模型的拟合度(Rost & von Davier, 1995; Spiel, Gittler, Sirsch, & Glück, 1997), 以及与测验标度相关的经典问题中(Rost, Carstensen, & von Davier, 1997; Hong & Min, 2007; Schultz-Larsen, Kreiner, & Lomholt, 2007; Meiser &

收稿日期: 2013-05-25

*全国教育科学“十二五”规划教育部重点课题(GFA111009)、广州卓越教育项目: 学生学业水平认知诊断评价、2012 年度教育部人文社会科学研究青年基金项目(12YJC190016)、广州市基础教育学业质量监测系统项目(GZIT2012-ZB0292)。

通讯作者: 张敏强, E-mail: zhangmq1117@qq.com

Machunsky, 2008)。此外, Bonnefon 和 Vautier (2010) 还探讨了该模型在心理实验中的应用。混合 IRT 模型同样可用于鉴别人格问卷中被试的作答风格和作伪现象(Zickar, Gibby, & Robie, 2004; Eid & Zickar, 2007)。

Rost (1990)首次尝试将 Rasch 模型与潜在类别模型相结合, 提出了混合 Rasch 模型(mixed Rasch model, MRM), 以期结合两种模型的理论强度来提高模型对实测数据的拟合度。Kelderman 和 Macready (1990)采用了一个相同构思但不一样的呈现方式, 即对数线性潜在类别模型, 分析并解释项目功能差异。Smit, Kelderman 和 van der Flier (1999)将人口统计学变量作为协变量纳入 MRM 中, 发现加入与潜在分类有关的协变量有利于潜在类别的划分和降低参数估计的标准误。Smit, Kelderman 和 van der Flier (2003)探讨了混合称名反应模型和混合广义分部评分模型在艾森克人格问卷中的应用。Vermunt (2008)将多水平混合 IRT 模型应用于数学测验。De Jong 和 Steenkamp (2010)将多水平多维混合 IRT 模型应用于跨文化研究, 分析潜变量的结构。总之, 根据所结合的项目反应模型的不同, 混合 IRT 模型发展出多种表现形式, 从混合 Rasch 模型拓展到多维混合 IRT、多水平混合 IRT、多维多水平混合 IRT, 从外显变量为二分变量拓展到称名变量(mNRM, a mixture version of the nominal response model)、多级计分(mPCM, a mixture version of the partial credit model)的度量模型等等。此外, 还形成了用于研究某些外显分组变量与潜变量之间关系的带协变量的混合模型。

2 混合 IRT 潜在模型

2.1 常见混合 IRT 潜在模型介绍

2.1.1 二级计分项目的度量模型

由 IRT 中的 Rasch 模型和 LCA 相结合衍生而来的混合 Rasch 模型(MRM)是研究最深入、应用最广泛的单维混合 IRT 模型之一。公式(1)为 MRM 的数学表达式:

$$P(X_{ij}=1|\theta_{jm}, \delta_{im}, m) = \frac{\exp(\theta_{jm} - \delta_{im})}{1 + \exp(\theta_{jm} - \delta_{im})} \quad (1)$$

其中, m 表示被试所属的潜在类别($m=1, 2, \dots$,

M), π_m 是第 m 个潜在类别的大小, 也称混合比例, 满足 $\sum_{m=1}^M \pi_m = 1$ 。 θ_{jm} 是第 m 个潜在类别中被试 j 的特质水平, δ_{im} 是第 m 个潜在类别中项目 i 的难度, $P(X_{ij}=1|\theta_{jm}, \delta_{im}, m)$ 是类别 m 中被试 j 在项目 i 上得 1 分的概率。

MRM 与带协变量的 MRM 的差别在于混合比例的不同, 带协变量的 MRM 中的 π_m 受到协变量的影响。假定被试样本中存在两种潜在类别(π_{j1} 指被试 j 属于潜在类别 1 的概率, π_{j2} 指被试 j 属于潜在类别 2 的概率), 协变量 y 为二分变量(0 或 1), 那么, 可以建立 π_{jm} 对 y_j 的逻辑斯蒂回归模型:

$$\logit(\pi_{jm}) = \beta_{0m} + \beta_{1m}y_j \quad (2)$$

$$\pi_{jm} = \frac{\exp(\beta_{0m} + \beta_{1m}y_j)}{\sum_{m=1}^M \exp(\beta_{0m} + \beta_{1m}y_j)} \quad (3)$$

令 $\beta_{01}=0$, $\beta_{11}=0$, 则可以得到公式(4)、(5)、(6)、(7):

$$\pi_{j1} = P(m=1|y_j=0) = \frac{1}{1 + \exp(\beta_{02})} \quad (4)$$

$$\pi_{j2} = P(m=2|y_j=0) = \frac{\exp(\beta_{02})}{1 + \exp(\beta_{02})} \quad (5)$$

$$\pi_{j1} = P(m=1|y_j=1) = \frac{1}{1 + \exp(\beta_{02} + \beta_{12})} \quad (6)$$

$$\pi_{j2} = P(m=2|y_j=1) = \frac{\exp(\beta_{02} + \beta_{12})}{1 + \exp(\beta_{02} + \beta_{12})} \quad (7)$$

于是, 可以计算发生比率:

$$\text{Odds ratio} = \frac{P(m=1|y_j=0)/P(m=2|y_j=0)}{P(m=1|y_j=1)/P(m=2|y_j=1)} \quad (8)$$

2.1.2 多级计分项目的度量模型

对于多级计分项目而言, 最常见的混合模型为混合称名反应模型(mNRM)。令 X_{ij} 为被试 $j \in \{1, \dots, J\}$ 对项目 $i \in \{1, \dots, I\}$ 的作答反应, $x_{ij} = k \in \{0, \dots, K_i\}$, Y_j 是一个随机变量, 表示被试 j 的潜在类别属性, $y_j = m \in \{1, \dots, M\}$, 则 mNRM 可表示为:

$$P(X_{ij}=k|Y_j=m, \theta_{jm}) = \frac{\exp(\alpha_{ikm}\theta_{jm} + \delta_{ikm})}{\sum_{h=0}^{K_i} \exp(\alpha_{ihm}\theta_{jm} + \delta_{ihm})} \quad (9)$$

其中, δ_{ikm} 和 α_{ikm} 分别表示潜在类别 m 中项目 i 上第 k 等的难度参数和区分度参数, 且满足 $\sum_k \delta_{ikm} = 0$, $\sum_k \alpha_{ikm} = 0$ 。混合分部评分模型 (mPCM) 是 mNRM 的一种特殊形式, 它将区分度参数设定为与项目等级相等。基于 PCM 的通式, mPCM 描述了类别 m 中被试 j 在项目 i 上选择第 k 等的概率:

$$P(X_{ij} = k | Y_j = m, \theta_{jm}) = \frac{\exp\left[\sum_{r=1}^k (\theta_{jm} - \tau_{irm})\right]}{1 + \sum_{h=1}^{K_i} \exp\left[\sum_{r=1}^h (\theta_{jm} - \tau_{irm})\right]} \quad (10)$$

τ_{irm} 为 阈 限 参 数, $\tau_{i0m} = 0$, $\tau_{irm} = \delta_{i,k-1,m} - \delta_{ikm}$, $\delta_{i0m} = 0$ 。于是, mPCM 的数学表达式就转换成了公式(11)所示:

$$P(X_{ij} = k | Y_j = m, \theta_{jm}) = \frac{\exp(k\theta_{jm} + \delta_{ikm})}{\sum_{h=0}^{K_i} \exp(h\theta_{jm} + \delta_{ihm})} \quad (11)$$

其它模型(如广义分部评分模型的混合形式等)可以通过对 mNRM 或 mPCM 中的项目参数进行设限而获得。

外显分组变量的纳入将提高测量模型的稳定性, 降低参数估计的标准误, 并且有利于潜在类别的划分和潜在类别特征的描述。假定存在外显分组变量 A 和 B, 对于多级计分模型而言, 可以建立一个饱和的对数线性模型来描述潜在类别变量与外显分组变量之间的联系, 公式如下:

$$\log P(Y_j = m | A_j = a, B_j = b) = u_{ab} + \lambda_m + \lambda_{ma} + \lambda_{mb} + \lambda_{mab} \quad (12)$$

u_{ab} 是对比例常数取对数的结果, λ_m 表示潜在类别的主效应, 而潜在类别与协变量 A、潜在类别与协变量 B、潜在类别与协变量 A 和协变量

B 的交互作用分别为: λ_{ma} 、 λ_{mb} 、 λ_{mab} 。

同时, 可以建立一个潜在特质变量在外显变量上的线性回归模型, 公式如下:

$$\theta_{jm} = \beta_0 + \beta_a + \beta_b + \varepsilon \quad (13)$$

其中, $\varepsilon \sim N(0,1)$, β_0 为截距, β_a 和 β_b 分别表示协变量 A 和协变量 B 对潜在特质的影响效应。

2.2 模型选择与参数估计

如图 1 所示, 混合 IRT 将 IRT 与 LCA 相结合, 意味着各潜在类别服从同一 IRT 模型(被试的作答反应受相同潜在特质变量的影响), 但不同潜在类别的模型参数存在差异(被试的作答反应还可能受到潜在分组变量的影响, 如解题策略、作答风格等; 此外, 被试的作答反应也可能受到外显分组变量的影响, 如性别、种族等)。

概括地讲, 混合 IRT 分析方法的基本过程是: 首先, 通过模型适配检验确定最佳模型, 即确定潜在类别的数量; 然后, 依据被试的作答模式将被试划分到相应的潜在类别并估计其潜在特质得分。因此, 混合 IRT 模型包括两个方面的重要内容: 模型选择、参数估计。

通常, 模型的选择是一个模型比较的过程。首先, 通过比较不同潜在类别数量的潜在类别模型与数据的拟合程度, 选择最佳的潜在类别模型; 然后, 通过比较不同潜在类别数量的混合 IRT 模型与数据的拟合程度, 选择最佳的混合 IRT 模型; 最后, 比较 IRT 模型、最佳潜在类别模型以及最佳混合 IRT 模型对数据的拟合情况, 从而选择最恰当的模型(马文超, 边玉芳, 骆方, 2012)。

不同的方法可用于评价模型的拟合度和选择恰当数量的潜在类别。首先, 先验理论。模型的构建源于一个坚实的理论基础, 理论不仅可以界

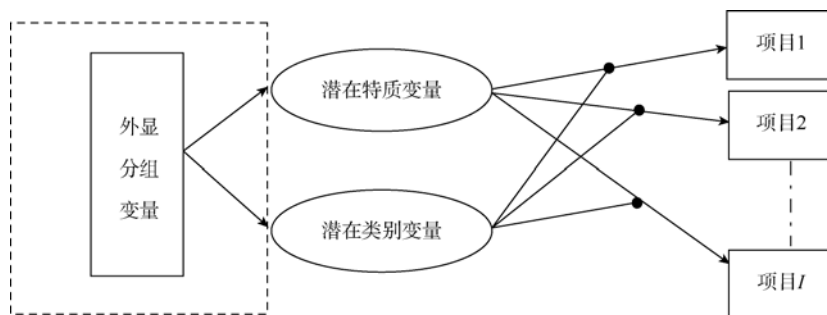


图 1 混合 IRT 示意图

定模型的形式，也可以详细说明模型中的因果关系。因此，在混合 IRT 模型中首先需要依据理论建立基础模型(IRT 模型)，然后根据数据特征用公式表现出来形成假设，选择适当的模型参数形式。其次，统计手段。在混合 IRT 中可以通过模型适配检验确定最佳模型，常见的模型适配指标包括信息评价指标 (information evaluation criteria)、似然比卡方检验、误分类率等。其中，AIC、BIC、CAIC 指标是较常用的信息评价指标，它们的值越小说明模型对数据资料的拟合度越好。Li, Cohen, Kim 和 Cho (2009)对混合 IRT 模型中常用信息评价指标的有效性进行了比较，结果表明 BIC 在大多数情况下都表现良好，而 AIC 在某些情况下倾向于选择复杂模型。这一结论与前人研究一致，BIC 优于其它指标，因为它考虑了样本大小，避免了过度参数化(McLachlan & Peel, 2000)。总而言之，评价模型时，研究者首先应该考虑的是先验理论而非统计数值，不能牺牲理论而屈就统计数值，也不能过分依赖某一指标，最好的方法是对多种指标做综合性的判断。

混合 IRT 模型的参数估计方法随着模型的建立也得到了相应的发展。其中，比较常见的参数估计方法包括：采用 EM 算法的边际极大似然估计方法(MMLE)，可在 mdltm (Multidimensional Discrete Latent Trait Model)软件(von Davier, 2005)以及 Mplus 软件(Muthén & Muthén, 2007)中实现；条件极大似然估计方法，可在 WINMIRA2001 软件中实现(von Davier, 2001)；马尔可夫链蒙特卡罗(Markov Chain Monte Carlo, MCMC)方法，可在

WINBUGS 软件中实现(Bolt, Cohen, & Wollack, 2002; Cohen & Bolt, 2005; Samuelsen, 2005)。长期以来，极大似然估计方法被认为是估计统计模型的一种有效方法。Rost (1990, 1991)提出混合 Rasch 模型时，通过研究实例展示了如何使用 EM 算法获得模型参数的极大似然估计值，但并没有用于评估该算法质量的详细模拟研究结果。Preinerstorfer 和 Formann (2012)通过模拟研究检测了混合 Rasch 模型中条件极大似然估计方法(cMLE)的质量，结果表明增加被试样本容量和测验长度会提高项目参数的估计精度。但是，当模型中参数个数或维度过多时，传统的 EM 算法一般难以或无法实现模型的参数估计。Patz 和 Junker (1999)研究表明 MCMC 方法能够简单明了地估计复杂的 IRT 模型。因此，许多研究者尝试将该方法推广到更为复杂的混合 IRT 模型中。Li 等人 (2009)采用 MCMC 方法对混合 IRT 模型进行参数估计，模拟研究结果表明 MCMC 方法的估计精度良好。其他研究者 (如 Cohen et al., 2005; Samuelsen, 2005; Dai, 2009)也先后将 MCMC 方法应用于多个混合 IRT 模型的参数估计中。表 1 呈现了混合 IRT 模型中 EM 算法与 MCMC 算法的优势与不足，总的来说，MCMC 算法是一种全新的参数估计方法，可以克服 EM 算法运算复杂、难以实现等局限性。但是，参数估计程序运行时间长是限制 MCMC 方法推广的一个重要因素。因此，进一步优化 MCMC 算法、缩短估计时长是今后研究的一个重要方向。

随着模型的发展及其相应的参数估计方法的

表 1 混合 IRT 模型中 EM 算法与 MCMC 算法的比较

参数估计方法		优点	缺点
EM 算法	应用较早较稳定的一种参数估计方法；	可估计复杂、高维模型； 避开了 EM 算法复杂运算，仅用初等数学方法来估计模型参数，从而提高参数估计的成功率； 包含待估参数的先验信息，使得参数估计更加稳定，其值更合理、更符合实际； 对初始值选择不敏感，提高了算法的适应性； 能够处理特殊的反应模式(全答对，全答错)。	当模型中参数多、维度多时，难以或无法实现模型的参数估计； 对初始化要求较高，且容易陷入局部最优； 对于特殊的反应模式(全答对，全答错)无法正确估计其项目参数。
	耗时较短。		
MCMC 算法			耗时太长。

应用,越来越多的研究者开始探讨参数估计准确性的问题。参数估计的质量可以通过分析推导或模拟研究来确定,然而分析推导很难甚至是不可能实现的。Preinerstorfer 等人(2012)模拟研究结果表明,参数估计的准确性依赖于被试样本容量、测验长度和潜在类别分布比例。随着被试样本的增加,参数估计的准确性提高;较多的项目同样会得到更精确的参数估计值;当不同潜在类别的大小一致(即被试样本容量一致)时,参数估计的准确性较高。此外,不同潜在类别被试的能力分布也是一个值得探讨的影响因素,考虑不同潜在类别被试能力分布的差异性有利于模型收敛、加速算法、估计模型参数(Rost, 1990; Dai, 2009)。

3 混合IRT潜在模型的应用

混合IRT作为心理统计和测量理论的结合,主要被应用于人格测量、能力测量以及临床诊断等领域,用以探讨项目功能差异(Differential Item Functioning, DIF)检测、心理与行为特征分类以及测验效度评价等问题。

3.1 项目功能差异检测

检验DIF的方法有多种,如多群组CFA方法、Mantel-Haenszel程序、logistic回归方法、SIBTEST,以及IRT的DIF分析方法等等。这些方法有一个相同的前提假设:目标群体中的群组差异事先得知。它们的基本做法是:首先确定一个先验分组信息,根据这个外显标准划分被试样本(如分成男、女两组),然后比较不同群组的模型参数是否有变化,若有变化,则存在DIF。

但是,有研究者发现与DIF有关的外显特征和实际的优势组与不利组的联系较弱,外显的分组信息也并不如通常所假定的那样总是保持一致,被研究的群组也可能并不是受DIF影响的群组(De Ayala, Kim, Stapleton, & Dayton, 2002; Samuelsen, 2005)。而且,事先假定被试作答反应的差异是由某个或某些外显分组变量所造成的,研究者往往只能局限于少数外显分组变量(如性别、年龄、种族等)的多群组比较中探讨样本异质性。但是,实际情形中还可能存在着其它显著的变量或者变量间的交互作用,它们能够更有效地区分不同质的子群体。

混合IRT相对于传统的DIF分析方法,最大的优势是不需要事先假定被试的哪些外显特征与

潜在类别的划分有关,换言之,在没有明确指定外显分组信息的前提下发掘被试作答行为之间的质性差异。此外,混合IRT将LCA和IRT相结合,能够同时处理连续型和类别型的潜变量,不仅可以检测出被试样本中的质性差异,而且能探测出各群组内部的量性差异,因而能挖掘出更多关于潜变量的信息。越来越多的研究者对潜在分类信息所导致的DIF问题感兴趣。例如,在跨文化研究中,Smit等人(2003)将种族背景当成协变量纳入混合IRT模型中探究DIF,发现文化差异与被试在项目上的得分差异有关。同样,混合IRT在教育测评中也具有很大的研究价值,它可以探测出学生在特殊教育问题上掌握能力的质性差异,并通过一个量化的方法来检测问题的根源,提出相应的解决措施(如 van Nijlen & Janssen, 2008; Cho & Cohen, 2010)。

3.2 心理与行为特征分类

心理测验中,样本异质性是不可避免的,如果忽略这种样本异质性,统计分析结果可能会出现严重的偏差。因此,不能随意采用传统的多组分析方法分析具有异质性的样本,除非可以明确指定每个被试属于哪一个组。

混合IRT作为IRT与LCA相结合的产物,同样也具备LCA的基本功能,它可以通过对被试外显作答反应的分析将被试划分为不同的类别,并描述不同类别所具有的心理行为特征。例如,Austin, Deary 和 Egan (2006)将混合Rasch模型应用于人格测验(NEO-FFI),深入分析了被试的作答反应。结果发现,神经质、宜人性、尽责性和外向性量表的模型建构最终都得到了两个潜在类别,其中一个类别的被试倾向于选择反应量尺上的极端选项,而另一个类别的被试倾向于选择中间选项;对于开放性量表而言,含三个潜在类别的模型与数据资料的拟合度最优,不过这种模型难以解释,因此将其归因于开放性量表的多维性特点上。

但与传统的LCA不同的是,混合IRT模型除了可以检测出被试样本中存在的多种潜在类别之外,还能揭示各类别内被试潜在特质上的量性差异。例如,马文超等人(2012)关于网络成瘾潜在结构的研究结果发现:(1)被试样本中包含两个存在着质性差异的群组。这种质性差异体现为两个群组中的被试对题目所描述的行为有不同的认识。

比如正常组中的中等成瘾程度的个体即会认同第6题所描述的内容(由于上网影响了自己的学习状态或朋友关系),而在成瘾倾向组中只有成瘾程度很高的个体才会对其表示认同。(2)两个群组还呈现出量的差异。一方面,群组间显示出明显不同的网络成瘾程度;另一方面,无论成瘾倾向组还是正常组,组内个体之间都表现出成瘾程度上的差异,Rasch模型度量了这种个体异质性。该研究说明,基于混合Rasch模型的分析有助于研究者认识网络成瘾的本质,深入了解成瘾倾向组和正常组中的个体差异,从而为临床治疗提供更为有意义的参考,对不同成瘾程度的个体采取更有针对性的干预和治疗手段。

3.3 测验效度评价

因子分析是检验测验效度的一种有效方法。不过,传统的因子分析依赖于施测样本,因此,基于因子分析结果得出的结论在测量工具的推广性方面受到了限制。理想的解决措施就是寻找支持效度推广的依据,即对大量不同的被试进行施测,这种方法被称为多群组CFA,可以用来检验不同观测群体的结构联系是否相同。但是,多群组CFA方法要求样本异质性的来源应事先得知。所以,有研究者认为,通过混合IRT检验测验的结构效度其推广性比多群组CFA方法优越(Sawatzky, Ratner, Kopec, & Zumbo, 2012)。

混合IRT模型作为项目反应模型和潜在类别模型的结合,同时考虑了被试样本的量性差异和质性差异,因此,相比单一类别模型,混合IRT模型能更详细地描述被试的作答模式。于是,研究者们假定与单一类别模型估计得到的潜在特质得分相比,混合IRT模型估计得到的潜在特质得分能更精确地预测外显行为。为了验证这一假设,Maij-de Meij, Kelderman 和 van der Flier (2008)将混合IRT应用于荷兰人格测验ABV (Amsterdamse Biografische Vragenlijst)的外向性量表和神经质量表,区分了具有不同作答风格的三类被试。同时,邀请一名心理学家在5点量表上对被试的外向性和情绪稳定性水平进行评定,评定结果作为效标变量。结果发现,与单一类别IRT模型相比(即假设所有受测样本均同质),含三个潜在类别的混合模型能够提高神经质量表的预测效度。但是,Maij-de Meij 等人的研究结果也显示了:混合模型虽然能更详细地描述被试在外向性量表上的作答

反应,但其估计得到的潜在特质得分并没有显著提高外向性量表的预测效度。

由此可见,与单一类别模型相比,混合IRT模型确实能够更详细地描述被试在测验上的作答行为,但是,它所估计得到的潜在特质得分究竟能否显著提高测验的预测效度还有待检验,可能不同的测验会得到不同的结果。

4 小结与展望

作为心理测量与心理统计相结合的产物,混合IRT在IRT与LCA的基础上继承和发展了新的优势:第一,混合IRT考虑到施测过程中的样本异质性问题,深入考察被试的作答反应,因而可能提高测验对被试心理、行为特征的预测力;第二,相对于传统的IRT或LCA方法,混合IRT分析方法具有一定的优越性,它可以使很多复杂的变量综合明朗化,克服了变量描述的单一性。更重要的一点是,混合IRT模型不仅可以通过构造分类潜变量发掘潜在的类别,同时也可以对不同潜在类别之间的连续潜变量进行对比研究;第三,混合IRT是基于概率模型对数据进行处理,因此存在多种统计指标,通过比较这些统计指标,能够确定最优模型,这种分析方法更为客观、符合实际。第四,与传统的DIF分析方法相比,混合IRT由于结合了LCA的思想,可以提炼出被试的潜在分类信息,而不用事先假定被试的分组信息。因此,混合IRT更符合实际情况,能提供更多的信息解释DIF。

当然,混合IRT作为IRT与LCA的结合,在继承两者优势的同时,也不可避免地吸收了一些不足。第一,我们在寻找可能存在的潜在类别群体时,得到的结果可能并不是最佳的人群分类,模型只能作为一种可能的方式解释观测数据。此外,我们所采用的方法限定了一些假设条件,这样也可能导致没有可识别的潜在类别存在。第二,由于不同潜在类别的模型参数不一样,因此,类别之间被试的潜在特质得分不能直接进行比较。第三,潜在分类信息利用不均匀。在划分类别时,各被试归属于某一潜在类别的条件概率不一致,可能会降低测验的预测精度。第四,混合IRT将LCA与IRT相结合,模型中待估参数增加了,同时,对研究者的理论素养与技术水平的要求更高,需要具备较好的数学功底和计算机编程

技能。混合IRT模型存在的这些不足从理论上和技术上都对研究者提出了挑战,但同时也指明了发展方向,研究者对这一方法进行应用和推广时需予以注意并发展出相应的解决措施。

综观上述,混合IRT在理论研究和方法技术上都存在发展的空间。总的来说,混合IRT还可以从以下几个方面作进一步探讨:

第一,加强对混合IRT模型参数估计方法的研究与应用。目前混合IRT模型的参数估计程序仍相对复杂,因此,进一步简化参数估计过程,开发参数估计方法对混合IRT的发展与应用具有极为重要的意义,是混合IRT的一个重要研究方向。

第二,拓展混合IRT的应用领域。混合IRT已被应用于二级计分项目和多级计分项目,但实际应用中往往会发现,有的测验材料既有0—1计分试题,也有多级评分题,并且含题组的测验也不乏其例。伴随混合IRT理论体系的完善与研究技术的发展,可以尝试将混合IRT应用于这种混合题型或含题组的测验项目分析。

第三,将外显分组变量与潜在分类信息相结合。在混合IRT模型中加入不同的外显分组变量作为协变量,探测协变量对潜在类别划分以及参数估计的影响,尤其是检测DIF时,应考虑到不同协变量与潜变量的交互作用。以往研究采用的外显分组变量多为类别型变量,如种族、性别等,而极少考虑连续型外显分组变量,今后的研究可探讨连续型外显分组变量(如社会赞许性倾向)对模型的影响效应。

第四,深化对混合IRT方法的理论介绍与研究。虽然国外对混合IRT基本理论体系的介绍与研究已趋于完善,国内关于LCA和IRT的研究也已积累了不少成果,但国内关于混合IRT的研究仍较为薄弱。因此,加强混合IRT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。

参考文献

- 马文超,边玉芳,骆方.(2012).网络成瘾的潜在结构:连续的还是分类的?《心理发展与教育》,28(5),554-560.
- 漆书青,戴海崎.(1992).《项目反应理论及其应用研究》.南昌:江西高校出版社.
- 邱皓政.(2008).《潜在类别模型的原理与技术》.北京:教育科学出版社.
- Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modeling of responses to NEO-FFI items. *Personality and Individual Differences*, 40(6), 1235-1245.
- Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, 39(4), 331-348.
- Bonnefon, J.-F., Eid, M., Vautier, S., & Jmel, S. (2008). A mixed Rasch model of dual-process conditional reasoning. *The Quarterly Journal of Experimental Psychology*, 61(5), 809-824.
- Bonnefon, J.-F., & Vautier, S. (2010). Modern psychometrics for the experimental psychology of reasoning. *Acta Psychologica Sinica*, 42, 99-110.
- Cho, S.-J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, 35(3), 336-370.
- Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42(2), 133-148.
- Dai, Y. Y. (2009). *A Mixture Rasch Model with a Covariate: A Simulation Study via Bayesian Markov Chain Monte Carlo Estimation*. Unpublished doctoral dissertation, University of Maryland.
- De Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential Item Functioning: A mixture distribution conceptualization. *International Journal of Testing*, 2, 243-276.
- De Jong, M. G., & Steenkamp, J.-B. E. M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. *Psychometrika*, 75(1), 3-32.
- Eid, M., & Zickar, M. J. (2007). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 255-270). New York: Springer.
- Glück, J., Machat, R., Jirasko, M., & Rollett, B. (2001). Training-related changes in solution strategy in a spatial test: An application of item response models. *Learning and Individual Differences*, 13(1), 1-22.
- Hong, S., & Min, S.-Y. (2007). Mixed Rasch modeling of the self-rating depression scale: Incorporating latent class and Rasch rating scale models. *Educational and Psychological Measurement*, 67(2), 280-299.
- Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups.

- Journal of Educational Measurement*, 27(4), 307–327.
- Li, F. M., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, 33(5), 353–373.
- Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*, 32(8), 611–631.
- McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: John Wiley & Sons Inc.
- Meiser, T., & Machunsky, M. (2008). The personal structure of personal need for structure. *European Journal of Psychological Assessment*, 24(1), 27–34.
- Muthén, L. K., & Muthén, B. O. (2007). *Mplus User's Guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.
- Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24(2), 146–178.
- Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, 65(2), 251–262.
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271–282.
- Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, 44(1), 75–92.
- Rost, J., Carstensen, C. H., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324–332). Münster, Germany: Waxmann.
- Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In *Rasch models* (pp. 257–268). New York: Springer.
- Samuelsen, K. M. (2005). *Examining differential item functioning from a latent class perspective*. Unpublished doctoral dissertation, University of Maryland, College Park.
- Sawatzky, R., Ratner, P. A., Kopec, J. A., & Zumbo, B. D. (2012). Latent variable mixture models: A promising approach for the validation of patient reported outcomes. *Quality of Life Research*, 21(4), 637–650.
- Schultz-Larsen, K., Kreiner, S., & Lomholt, R. K. (2007). Mini-Mental Status Examination: mixed Rasch model item analysis derived two different cognitive dimensions of the MMSE. *Journal of Clinical Epidemiology*, 60(3), 268–279.
- Smit, A., Kelderman, H., & van der Flier, H. (1999). Collateral information and mixed Rasch models. *Methods of Psychological Research Online*, 4(3), 19–32.
- Smit, A., Kelderman, H., & van der Flier, H. (2003). Latent trait latent class analysis of an Eysenck Personality Questionnaire. *Methods of Psychological Research Online*, 8(3), 23–50.
- Spiel, C., Gittler, G., Sirsch, U., & Glück, G. (1997). Application of the Rasch model for testing Piaget's theory of cognitive development. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 111–117). Münster, Germany: Waxmann.
- van Nijlen, D., & Janssen, R. (2008, March). *Mixture IRT-models as a means of DIF-detection: Modeling spelling in different grades of primary school*. Paper presented at annual meeting of the National Council on Measurement in Education location, New York.
- Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research*, 17(1), 33–51.
- von Davier, M. (2001). *WINMIRA user manual*. Kiel: IPN.
- von Davier, M. (2005). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models* [Computer software]. Princeton, NJ: Educational Testing Service.
- Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods*, 7(2), 168–190.

The Mixture Item Response Theory Models and Its Application Traces

WANG Xia¹; TAN Guohua²; WANG Xu³; ZHANG Minqiang¹; LUO Cong¹

(¹ Center for Studies of Psychological Application;

School of Psychology, South China Normal University, Guangzhou 510631, China)

(² Teaching Research Office of Guangzhou Education Bureau, Guangzhou 510030, China)

(³ Guangzhou Municipal Education Information Center, Guangzhou 510030, China)

Abstract: Item response theory (IRT) is a modern measurement theory for accurately estimating individuals' ability, and latent class analysis (LCA) is a statistical technique used to identify subtypes of individuals, which is based on statistic models. Mixture item response theory (Mixture IRT)[0] combining LCA and IRT is able to classify the subjects as well as quantifying their traits. The concept and principle of Mixture IRT is elaborated in this paper. Besides, several common mixed models are introduced here, such as the mixed Rasch model (MRM), a mixture version of the nominal response model (mNRM) and a mixture version of the partial credit model (mPCM). Furthermore, different parameter estimation methods are described and the application traces in psychological test of Mixture IRT is evaluated from the perspectives of classifying psychological or behavioral traits, detecting differential item functioning and estimating test validity.

Key words: item response theory (IRT); latent class analysis (LCA); Mixture IRT; latent structure