Advances in Psychology 心理学进展, 2021, 11(6), 1428-1437
Published Online June 2021 in Hans. http://www.hanspub.org/journal/ap
https://doi.org/10.12677/ap.2021.116160

Hans汉斯

多水平混合IRT模型及其发展与应用

齐媛媛，陈德枝*

浙江师范大学杭州幼儿师范学院，浙江 杭州
Email: *cdezhi@zjnu.cn

收稿日期：2021年5月5日；录用日期：2021年6月16日；发布日期：2021年6月23日

摘 要

多水平混合IRT模型(MMIRTM)将IRT与潜在类别分析和阶层线性模型相结合，能够同时对嵌套在多水平下的被试分类并量化其潜在特质。它是近年来教育心理测量学的研究热点与重点之一。在梳理多水平混合IRT模型的发展由来、基本概念和原理的基础上，对多水平混合IRT模型在项目功能差异检测等领域的应用做了相关阐述，并对当前多水平混合IRT模型的应用与拓展进行了述评与展望。

关键词

多水平混合IRT模型，IRT，潜在类别分析，多水平分析

Multilevel Mixture IRT Model and Its Development and Applications

Yuanyuan Qi, Dezhi Chen

Hangzhou College for Kindergarten Teachers, Zhejiang Normal University, Hangzhou Zhejiang
Email: *cdezhi@zjnu.cn

Received: May 5th, 2021; accepted: Jun. 16th, 2021; published: Jun. 23rd, 2021

Abstract

The Multilevel Mixture IRT Model (MMIRTM) combines IRT with latent category analysis and hierarchical linear model, which can simultaneously classify and quantify the potential characteristics of subjects nested in multiple levels. It is one of the research hotspots and key points of educational psychometrics in recent years. On the basis of sorting out the development origin, basic

*通讯作者。

文章引用: 齐媛媛, 陈德枝(2021). 多水平混合 IRT 模型及其发展与应用. *心理学进展, 11(6),* 1428-1437.
DOI: 10.12677/ap.2021.116160

concepts and principles of the multilevel mixture IRT model, this paper elaborates the application of the multilevel mixture IRT model in the field of Differential Item Functioning (DIF) detection and other fields, and makes an review and outlook on the application and expansion of the current multilevel mixture IRT model.

Keywords

Multilevel Mixture IRT Model, IRT, Latent Category Analysis, Multi-Level Analysis

1. 引言

项目反应理论(IRT)模型是依据被试在各个项目上的实际作答反应结果，经数学模型的运算，统一估计出被试的能力水平或潜在的心理特质水平以及项目的计量学参数(Lord, 1980；罗照盛，2012)。潜在类别分析(LCA) (Lazarsfeld & Henry, 1968；邱皓政，2008)是把类别数据与潜在变量的观念加以结合，用潜在的类别变量来解释外显的类别变量之间的关联，使外显变量之间的关系通过潜在类别变量 X 来估计(张洁婷，焦璨，张敏强，2010)。阶层线性模型(HLM) (Anthony & Stephen, 1992)则是用来解决多层次数据和多水平数据的统计分析问题，例如教育测量学中常见的学生嵌套于班级，班级嵌套于学校所形成的三层数据结构等(温福星，2009)。为了更好地拟合那些作答被试内部存在不同潜在特征类别且具有多层嵌套结构的数据，只在 IRT 自身的框架内发展新模型已不能满足实际需要，因此 IRT 逐步与其它心理统计方法，如与 LCA、HLM 相结合构建新的 IRT 模型，这已逐渐成为 IRT 的发展趋势。

如将 IRT 模型的理论优势与 LCA 的分类能力相结合，所形成的混合 IRT 模型(Mixture IRT Model, MixIRTM) (Rost, 1990)。MixIRTM 用于识别具有相同项目反应模式的潜在被试类别，每个潜在类别的被试能力结构不同，反应策略也不同。MixIRTM 基于这一理论基础来考查潜在类别间的差异，但 MixIRTM 的一个重要局限是，它本质上违反了局部独立性假设，忽略了在许多教育测试数据中存在除学生水平之外的其他多层嵌套结构，不能解释在一个多水平(层次)结构中存在的依赖关系(Lee, Cho, & Sterba, 2018)。因此，将 IRT 模型嵌套在 HLM 中的多水平 IRT 模型(Multilevel IRT Model)应需而生(Kamata, 2001)。多水平 IRT 模型虽然解决了数据嵌套问题，但是从混合 IRT 建模的角度来看，多水平 IRT 模型的一个局限性是，除了模型中所包含的显性估计参数给出的信息外，不能提供关于群体成员的信息(Cho & Cohen, 2010)。

为了能够克服混合 IRT 和多水平 IRT 的局限性，并同时拥有两者的优势，研究者们开始尝试将多水平模型和混合 IRT 模型相结合，即多水平混合 IRT 模型(Multilevel Mixture Item Response Theory Model, MMIRTM) (Vermunt, 2007)。该方法一提出即被用于教育教学测量，并取得了更精确的分析结果，为教育教学提供了更加科学的评价信息(Vermunt, 2007)。有研究结果进一步表明(Cho & Cohen, 2010)，MMIRTM 不仅可以提高测量结果的精确性，同时可以获得不同潜在类别群体的测量特征，MMIRTM 不仅是 IRT 模型、LCA 和 HLM 三种方法的结合，也是近年来心理、教育测量应实际需求发展计量模型的新方向。梳理与讨论 MMIRTM 的研究现状与进展，是探索 MMIRTM 的首要工作，对 MMIRTM 的发展与应用均具有重要意义。

2. 多水平混合 IRT 模型简介

多水平混合 IRT 模型是建立在混合 IRT 模型和多水平 IRT 模型两个重要概念之上的现代心理、教育计量学模型。认识多水平混合 IRT 模型首先要了解混合 IRT 模型和多水平 IRT 模型。以下对这两种模型进行简要说明。

2.1. 混合 IRT 模型

Rost (Rost, 1990)在成人物理知识测验中，为提高模型对数据的拟合性能，首先将 Rasch 模型与 LCA 相结合，提出了混合 Rasch 模型(Mixture Rasch Model, MRM) (黄明明，王立君，2015)。即在原始的 Rasch 模型中引入“被试所属的潜在类别 c”，用来诊断说明不同潜在类别被试之间质的差异，同时量化同一类别内被试在相同项目下的能力结构差异。自此之后，其他 IRT 模型也都相继与 LCA 相结合，并最终形成了混合 IRT 模型体系。这种混合模型体系使得 IRT 模型具有了潜在类别的特性。因此，混合 IRT 的形成丰富了 IRT 的理论体系，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力的分析工具(王霞，谭国华，王旭，张敏强，骆聪，2014)。混合 IRT 模型中，以最为典型的 MRM 为例作为说明，1990 年 Rost (Rost, 1990)首次提出的 MRM 表达式如下：

$$P\left(X_{ij}=1\middle|\theta_{jc},\beta_{ic},C\right)=\sum_{c}^{C}\pi_{c}\left[\frac{\exp\left(\theta_{jc}-\beta_{ic}\right)}{1+\exp\left(\theta_{jc}-\beta_{ic}\right)}\right] \tag{1}$$

θ_j 为被试 j 的能力，β_i 是项目 i 的难度，c 表示被试所属的潜在类别 $\left(c=1,2,\cdots,C\right)$，$\pi_c$ 表示第 c 类别的大小，也称为混合比例大小，$0<\pi_c<1$。$P\left(X_{ij}=1\middle|\theta_{jc},\beta_{ic},C\right)$ 是被试 j 在所有类别 c 中的项目 i 上正确作答(0、1 计分时得 1 分)的概率之和，反映出被试的整体能力水平。

Mixture IRT 已经被用来描述和分析多种类型的数据特征，并应用在跨文化研究(Smit, Kelderman, & Flier, 2003)、教育测评(Cho & Cohen, 2010)、人格测验(Austin, Deary, & Egan, 2006)和网络成瘾潜在结构探索(马文超，边玉芳，骆方，2012)等研究中。

2.2. 多水平 IRT 模型

多水平 IRT 模型是将 IRT 嵌套在 HLM 中，根据所结合的 IRT 模型和 HLM 的不同，多水平 IRT 发展出多种表现形式(刘慧，简小珠，张敏强，熊悦欣，2012)。其中，1991 年，Zwinderman (Zwinderman, 1991)将 Rasch 模型和多元回归模型相结合，实现了 IRT 模型和 HLM 的初步结合；Adams 等人(Adams, Wilson, & Wu, 1997)提出两水平 IRT 模型，Kamata (Kamata, 2001)在广义线性模型的框架下对 Rasch 模型进行重组，提出了三水平 Rasch 模型，三水平分别为项目水平、个体水平和群体水平(刘慧等，2012)。Fox 和 Glas (Fox & Glas, 2003)将预测变量也定义为 IRT 模型中的潜在特质参数，从而提出了 HLM 与多个 IRT 模型结合的复杂多水平 IRT 模型。紧接着 Fox (Fox, 2005)提出将等级反应模型与 HLM 相结合的多水平 IRT 模型，该模型主要用于多级计分项目。刘红云等(刘红云，骆方，2008)进一步探讨了多水平 IRT 模型与一般 IRT 模型参数之间的关系，并初步探索了多水平 IRT 模型的推广模型。

目前来说，对于多水平 IRT 的应用已不在少数，模型大多被用在探讨测验等值(Chu & Kamata, 2000; Chu & Kamata, 2005)、项目功能差异(differential item functioning, DIF) (Cheong, 2006)和学校效能评估(Fox, 2004)中。在单维多水平 IRT 模型不断开发的同时，研究者们陆续提出了一些涉及多个潜在特质维度的多维多水平 IRT 模型(刘慧等，2012)。Cheong 和 Raudenbush (Cheong & Raudenbush, 2000)使用儿童行为清单将行为问题外化，运用多维多水平 Rasch 模型测量和建模了来自 79 个城市的 2177 名 9~15 岁儿童和青

少年问题行为的分析策略。Raudenbush 等(Raudenbush, Johnson, & Sampson, 2003)在研究社会行为、态度和信念的相关性时，也提出了“多水平 Rasch 模型”这一具有随机效应的多变量模型，并将其应用在犯罪行为的自我报告中。

2.3. 多水平混合 IRT 模型

2007 年，Vermunt 在国际统计学会第五十六届会议公报上首次将多水平 IRT 模型的第二水平潜变量看做类别型数据，即初步形成多水平的混合 IRT 模型(Vermunt, 2007)。Vermunt (Vermunt, 2008)将初期的多水平混合 IRT 模型应用于数学测验。De Jong 和 Steenkamp (Jong & 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(王霞等，2014)。Cho 和 Cohen 于 2010 年(Cho & Cohen, 2010)将多水平结构引入混合 IRT 模型，并将该模型扩展为多水平混合 IRT 模型，明确提出了 MMIRTM 的概念，提出的 MMIRTM 混合了两个水平(学生水平和学校水平)的潜在类别，并谈及该模型在 DIF 中的应用。总之，根据所结合的 IRT 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和多级计分的度量模型等(王霞等，2014)。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。并最终形成了混合 IRT 模型体系，MMIRTM 也是混合 IRT 模型体系中的一种，单参数的 MMIRTM 就是多水平混合 Rasch 模型(Multilevel Mixture Rasch Model, MMRM)，标准 Rasch 模型如公式(2)所示，MMRM 是将 Rasch 模型和潜在类别分析以及多水平模型相结合，不仅可以提高模型参数估计的精确性，同时可以获得不同潜在类别群体的测量特征(Cho & Cohen, 2010；李美娟，刘玥，刘红云，2020)，在 Cho 和 Cohen (Cho & Cohen, 2010)研究提出的单维 MMRM 中，被试在 0~1 计分的项目上正确作答反应的概率如公式(3)所示：

$$p_{ij}=\frac{1}{1+\exp\left[-\left(\theta_j-\beta_i\right)\right]} \tag{2}$$

$$P=\left(y_{ijt}=1\middle|g,k,\theta_{jtgk}\right)=\frac{1}{1+\exp\left[-\left(\theta_{jtgk}-\beta_{igk}\right)\right]} \tag{3}$$

式(2)中 θ_j 为被试 j 的能力值，β_i 为项目 i 的难度参数，p_{ij} 为被试 j 在项目 i 上正确作答反应的概率；式(3)中，$g=1,\cdots,G$ 是第一水平潜在类别指标，$k=1,\cdots,K$ 是第二水平潜在类别指标，$j=1,\cdots,J$ 表示被试，$t=1,\cdots,T$ 表示学校，$i=1,\cdots,I$ 表示项目，θ_{jtgk} 表示被试 j 在学校 t 和在潜在类别 g、k 的能力，β_{igk} 表示潜在类别 g 和 k 的项目 i 难度。MMRM 不同水平和潜在类别的混合结构如表 1 所示：

Table 1. Mixed proportional structure of MMRM
表 1. MMRM 的混合比例结构

	K = 1	K = 2	.....	K = K
G = 1	Π1|1	Π1|2	.....	Π1|K
G = 2	Π2|1	Π2|2	.....	Π2|K
.....	.....	.....	.....	.....
G = G	ΠG|1	ΠG|2	.....	ΠG|K
总和	$\sum_{g=1}^{G}\pi_{g\|1}=1$	$\sum_{g=1}^{G}\pi_{g\|2}=1$	.....	$\sum_{g=1}^{G}\pi_{g\|k}=1$

表 1 中，纵列表示第二水平潜在类别，横列表示第一水平的潜在类别，在第一水平的潜在类别确定后，在此基础上进行第二水平潜在类别的确定，进而形成多水平的模型。由此，可更加清晰得到多水平混合 Rasch 模型与 Rasch 模型的关系，即多水平混合 Rasch 模型是在 Rasch 模型的基础上加入了多个水平(水平数 ≥ 2)，每个水平上又分别加入了其潜在类别参数，因此被试最终作答的正确率同时受到两水平嵌套以及其内部潜在类别的影响。对于两水平的数据，MMIRTM 可以在第一水平和第二水平进行非连续潜在变量(潜在类别)和连续潜在变量(能力)的分析，第一水平的潜类别分析主要基于被试作答反应之间的关系，第二水平的潜类别分析主要基于组内被试作答反应之间的关系(李美娟，刘玥，刘红云，2020；Vermunt, 2003)。

目前来说，MMIRTM 的参数估计方法有极大似然估计(Maximum likelihood Estimation, MLE) (Finch & Finch, 2013)、贝叶斯估计(Bayesian Estimation, BE)，还可以使用 WinBUGS 1.4 编写的马尔可夫链蒙特卡罗(MCMC)算法进行估计(Spiegelhalter, Thomas, & Best, 2003)。大多数估计方法都通过 MPLUS 实现，也可以通过在 R 中自主编程来实现(李美娟，刘玥，刘红云，2020)，最近有研究者(Chung & Houts, 2020)开发出适用于多维 IRT 模型的 flexMIRT 软件包，它可以处理一维和多维 IRT 模型的二分、多级和混合项目，并支持多水平和多题组分析，就项目模型而言，两参数和三参数的 logistic 模型都可适用。

3. 多水平混合 IRT 模型的发展

3.1. 多水平混合 IRT 模型的多维性发展

在一些大型的统一教育考试中，虽然某一场考试可能只是一门学科，但考试中被试的反应却同时受到被试多个维度技能发展水平的影响，如英语水平测试中，被试的英语成绩不仅受到词汇认知能力和语法结构掌握程度的影响，还受到被试其他许多内在技能的影响，比如：阅读理解能力、语境感知能力和简易计算能力。而这些能力受到社会经济地位、性别等个体协变量影响，个体又嵌套在更高水平的群体(比如：班级、学校)中(Lu, Zhang, & Tao, 2018)，在这种情况下，通常 IRT 所假设的潜在特质的单维性将可能是无效的，因此将导致在项目参数和个体能力参数评估上的潜在困难(Ackerman, 1994; Reckase, 1985)。此外，当数据是多维的时，一般常用的 DIF 检验程序可能也不是有效的(例如：Mantel-Haenszel [MH]检验、logistic 回归检验、似然比检验(Steven & Howard, 2013))，因为这些方法的基本假设是被试的潜在能力是单维的，把被试者的能力限定在一个单一的潜在特征上。在这种情况下，推荐使用多维 IRT 模型，因为它既能提供准确的参数估计，也能提供关于被试在不同维度上的能力信息(Reckase, 2007)。考虑到多维 IRT 模型提供的潜在优势，以及上文提到的 MMRM 在多层数据的 DIF 评估方面的优势，Finch 等人(Finch & Finch, 2013)描述了既能适用多维数据也能适用多水平数据的多维多水平混合 Rasch 模型(Multidimensional and Multilevel Mixture Rasch Model, MMMixRM)，该模型与 MMRM 的使用方式基本相同，但适用于同时评估多个不同维度潜在特征的情况。MMMixRM 可以表示为

$$P=\left(y_{ijgt}=1\middle|g,k,\theta_{jtgkm}\right)=\frac{1}{1+\exp\sum\left[-\left(\theta_{jtgkm}-\beta_{igk}\right)\right]} \tag{4}$$

公式(4)中各参数的定义如公式(3)，其中 θ_{itgkm} 表示来自潜在类别 g、k 的被试 j 在学校 t 的 m 维度潜在特征上的能力参数，β_{igk} 表示潜在类别 g 和 k 的项目 i 难度；与 MMRM 一样，MMMixRM 提供了第一水平和第二水平在不同潜在类别上正确作答的概率估计，以及特定类别的项目难度估计，在 MMMixRM 这种情况下，项目的难度估计是基于正确维度进行的，并非一定是单维。此外，在多维多水平混合 Rasch 模型中，不同水平潜在类别的划分关系是由一组测量多个维度潜在特征的项目决定的，而不是由只评估一个维度潜在特征的项目决定。

按照项目所测特质的维度数可分为单维多水平混合 IRT 模型和多维多水平混合 IRT 模型。在 MMIRTM 中，当所有项目都测量同一个潜在特质维度时则为单维 MMIRTM；项目测量不同潜在特质维度时则为多维 MMIRTM。根据当前的研究进展来看(Jang, Kim, & Cohen, 2018)，MMIRTM 大多用来测量多维潜在特质，多维 MMIRTM 又分为项目内多维和项目间多维，前者是指测验内部的单个项目测量多种潜在特质；后者是指测验分割成几个子测验，使得每个子测验内部的项目测量同一种潜在特质(付志慧，2010)。Jang 等人(Jang, Kim, & Cohen, 2018)证明维度对 MMIRTM 中潜在类别的提取有着重要影响，不同的多维结构、每个维度的项目数和维度之间相关程度都会影响模型中潜在类别的提取数量，当所有项目都测量单个维度(即项目间多维结构)时，混合 IRT 模型提取的潜在类别的数量小于某些项目测量两个以上维度(即项目内多维结构)时的数量。此外，无论多维结构是什么样的类型，由混合 IRT 模型提取的潜在类别的数量都会随着维度之间的相关性的增加而减少。也就是说当数据的多维结构复杂或维度之间的相关性较弱或独立时，混合 IRT 模型往往会提取更多的潜在类别。MMMixRM 除了应用在 DIF 中外(Finch & Finch, 2013)，在美国，社会科学研究人员还将该模型应用在大规模的跨文化研究中，在其他文化背景下检验他们的理论的普遍性和限制条件，而不是只局限在它们最初开发的地方。

3.2. 多水平混合 IRT 模型的改进拓展

为了满足研究者的客观需求，对原始的 MMIRTM 做出相应的改进以应用在实际问题解决中，刘红云等(Liu, Liu, & Li, 2018)为了分析 PISA 2012 中一个复杂问题解决任务的过程性数据，对 MMIRTM 进行改进，将 IRT 的“项目”概念改为个人反应中的每个动作，动作再依据某些规则进行评分，通过一系列动作来探索学生在解题过程中能力和反应策略的变化情况，结果表明改进后的 MMIRTM 可以更好地探索学生解决问题的具体策略，以及这些策略的优缺点，这些发现也可以进一步用于设计有针对性的教学干预。2020 年的最新研究中，李美娟等(李美娟等，2020)又使用拓展的 MMIRTM 来分析 PISA 2012 问题解决测验中的一道交通问题，模型中包含两个水平：过程水平和个体水平。在过程水平，定义潜类别来描述不同步骤的异质性，从而对不同策略进行分类；在个体水平，定义连续潜变量来估计个体的能力。

实际上，传统的 MMIRTM 是拓展 MMIRTM 的一种特例，拓展 MMIRTM 在过程水平和个体水平都具有其独特性，过程水平中每一当前步骤的潜在类别状态是前面各步骤的累积，而并非是只针对当前单独步骤的类别划分；个体水平潜变量的定义所采用的测量指标与传统的 MMIRTM 也不同(李美娟等，2020)。传统模型中，个体水平的潜变量是由第一水平(项目水平)的“向量”(观测变量)估算得到(Lee, Cho, & Sterba, 2018)，而拓展模型中可以定义更加自由的设计“矩阵”(过程水平)来决定个体层面能力估计值。还有研究者在传统 IRT 模型中加入协变量参数，进而对预测变量中的测量误差进行建模(Fox & Glas, 2003)。MMIRTM 的拓展不仅丰富了整个项目反应理论体系，而且极大地延伸了 IRT 模型的应用领域和适用范围、更加精确拟合具有复杂结构的被试数据以减小误差。

4. 多水平混合 IRT 模型的应用

项目功能差异(Differential Item Functioning, DIF)是指来自不同群体但能力水平相同的受测者，在同一题目上具有不同正确作答概率的现象(Steven & Howard, 2013；曾秀芹，孟庆茂，1999)。DIF 的研究一般针对两个团体，如两种性别、种族以及其它特征的被试进行分析，指目标组和参照组在测验所测的特质相同的前提下，两组在某题上的答对率有差异，这种差异不是由于特质差异引起的，而是由与测验无关的因素引起的(骆方，张厚粲，2006)。DIF 的检测方法有很多，一般来说常用到的有验证性因素分析(CFA)和 IRT，研究主要讨论 IRT 在 DIF 中的应用。早在 2005 年，已经有研究者将混合 IRT 模型应用在 DIF 中，结果表明混合 IRT 模型在 DIF 中的应用比一般 IRT 误差更小(Cohen & Bolt, 2005)；刘红云等人(刘红

云，骆方，2008)又用多水平 IRT 模型对 DIF 进行分析，结果表明，多水平 IRT 模型对变量的类型没有任何限制，可以同时分析多个变量的 DIF，而且可以分析造成 DIF 的潜在原因；MMIRTM 由于能够正确地解释多水平数据结构中数据之间的依赖关系而引起了研究人员的兴趣(Bacci & Gnaldi, 2015; Bennink, Croon, Keuning, & Vermunt, 2014; Cho & Cohen, 2010; Cho, Cohen, & Bottge, 2013; Finch & Finch, 2013; Jilke, Meuleman, & Walle, 2015; Liu et al., 2018; Tay, Diener, Drasgow, & Vermunt, 2011; Varriale & Vermunt, 2012)。2010 年，Cho 和 Cohen (Cho & Cohen, 2010)正式将 MMIRTM 应用在实际的 DIF 实验条件中，对 MMIRTM 的性能进行了模拟研究和实证研究，证明其参数估计效果极佳，并详细说明了如何使用 MMIRTM 来识别和描述潜在群体的特征。也有研究者(Bennink et al., 2014; Cho & Cohen, 2010; Finch & Finch, 2013)描述了不同类型的 MMIRTM 在检测项目功能差异中的应用，其中 Finch 等人(Finch & Finch, 2013)更是进一步扩展了用于 DIF 检测的 MMIRTM，将包含多个测量维度的多维 MMIRTM 应用于完成数学和语言测试的三年级学生的全国样本中，分析结果表明，多维模型比单独的单维模型提供了更完整的关于 DIF 性质的信息；除应用在 DIF 中外，还有不少研究者将 MMIRTM 应用在其他领域，比如有研究(Bacci & Gnaldi, 2015; Vermunt, 2008)利用 MMIRTM 来分析教育测量数据，根据学生的满意度水平，利用 MMIRTM 将大学课程划分为同质类；Tay 等人(Tay et al., 2011)则使用 MMIRTM 分析来自 116 个国家的 121740 份个人自我报告的情绪数据，结果发现了 4 个个体类别和 5 个国家类别，并讨论了 MMIRTM 对跨文化、多水平和测量等值研究的理论以及方法启示；Jilke 等人(Jilke et al., 2015)将 MMIRTM 应用在公共行政研究中，以市民对公共服务的满意度和对公共机构的信任两种测量结构为例，分别采用多组验证因子分析和 MMIRTM 来检验和纠正测量的不等值性。

5. 总结与展望

MMIRTM 不仅保留了混合 IRT 模型的特性，而且兼具多水平 IRT 模型的优良特性，可以处理具有多水平嵌套结构，且各水平内部具有不同潜在类别的数据，是 IRT 模型的又一次突破性拓展，使得模型对数据的拟合性能更加完善；也可以看作是 LCM 的一种拓展模型，现已成功应用在 DIF、跨文化和测量等值等研究中，具有广阔的应用前景。它还是一个可用于教育测验，尤其是教育和心理评估数据分析的优良测量模型。随着模型发展与应用的推广，相继出现了多维多水平混合 IRT 模型(MMMixIRTM)等，并成功应用于大规模的跨文化研究和多维度潜在特征数据的分析中。在近期的研究中，MMIRTM 不断进行拓展改进，在模型开发和应用方面发展势头强盛，而且不断与其他测量模型结合，形成更加复杂的测量模型。应用领域也不再局限于教育和心理测量方面，正在向个体发展、公共服务等方面进行拓展。随着 MMIRTM 本身的逐渐完善及其应用领域的不断拓展，在以下几个方面，MMIRTM 有待做进一步的探索。

首先是 MMIRTM 的模型参数估计及其影响因素。发展与应用 IRT 的首要工作是模型参数估计问题，虽然当前参数估计的相关研究提出采用 MLE，BE 和 MCMC 方法，但是这些方法的适用条件，估计精度的比较及其影响因素目前还少有探索。

其次，MMIRTM 的拓展。近年来随着人们对 MMIRTM 的逐步熟知，该模型应实际测量发展需要，对被试潜在能力维度进行了拓展。实际应用情况复杂，如何应实际需求对 MMIRTM 做相应的拓展，这将是 MMIRTM 未来发展与应用重点探讨的内容之一。

此外，已有研究表明 MMIRTM 能更深入准确地进行项目功能差异探讨。但 MMIRTM 在计算机自适应测验和测验等值等相关技术的应用方面目前还有大量未知有待探究。

最后，MMIRTM 是 IRT、LCA 和 HLM 三种方法的结合，认识与理解 MMIRTM 首先需具备 IRT、LCA 和 HLM 的基本知识，这样一来不仅加深了人们对 MMIRTM 的认识难度，也在一定程度上限制了

它的实际应用与推广。总之，如何更好地认识、发展和应用 MMIRTM，这都有待我们做进一步深入的探索。

基金项目

本文为浙江省教育科学规划课题“幼儿执行功能与数学学习的关系及其应用”的研究成果之一，项目批准号：2019SCG313。

参考文献

曾秀芹，孟庆茂(1999). 项目功能差异及其检测方法. *心理学动态, (2),* 41-47+57.

付志慧(2010). *多维项目反应模型的参数估计*. 博士学位论文，长春：吉林大学.

黄明明，王立君(2015). 混合 IRT 模型的研究进展与应用. *考试研究, (4),* 61-68.

李美娟，刘玥，刘红云(2020). 计算机动态测验中问题解决过程策略的分析：多水平混合 IRT 模型的拓展与应用. *心理学报, 52(4),* 528-540.

刘红云，骆方(2008). 多水平项目反应理论模型在测验发展中的应用. *心理学报, 40(1),* 92-100.

刘慧，简小珠，张敏强，等(2012). 多水平 IRT 的发展与应用述评. *心理科学进展, 20(4),* 627-632.

罗照盛(2012). *项目反应理论基础*(p. 4). 北京：北京师范大学出版社.

骆方，张厚粲(2006). 检验项目功能差异的两类方法——CFA 和 IRT 的比较. *心理学探新, 26(1),* 74-78.

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.

邱皓政(2008). *潜在类别模型的原理与技术*(p. 29). 北京：教育科学出版社.

王霞，谭国华，王旭，张敏强，骆聪(2014). 混合 IRT 潜在模型及其应用轨迹. *心理科学进展, 22(3),* 540-548.

温福星(2009). *阶层线性模型的原理与应用*. 北京：中国轻工业出版社.

张洁婷，焦璨，张敏强(2010). 潜在类别分析技术在心理学研究中的应用. *心理科学进展, 18(12),* 1991-1998.

周韵，译(2013). *项目功能差异(第 2 版)* (pp. 93-99). 上海：格致出版社，上海人民出版社. (Steven, J. O., & Howard, T. E., 2013).

Ackerman, T. A. (1994). Using Multidimensional Item Response Theory to Understand What Items and Tests Are Measuring. *Applied Measurement in Education, 7,* 255-278. https://doi.org/10.1207/s15324818ame0704_1

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel Item Response Models: An Approach to Errors in Variables Regression. *Journal of Educational and Behavioral Statistics, 22,* 47-76. https://doi.org/10.3102/10769986022001047

Anthony, S. B., & Stephen, W. R. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods* (p. 32). London: Sage Publication.

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual Differences in Response Scale Use: Mixed Rasch Modelling of Responses to NEO-FFI Items. *Personality and Individual Differences, 40,* 1235-1245. https://doi.org/10.1016/j.paid.2005.10.018

Bacci, S., & Gnaldi, M. (2015). A Classification of University Courses Based on Students' Satisfaction: An Application of a Two-Level Mixture Item Response Model. *Quality & Quantity: International Journal of Methodology, 49,* 927-940. https://doi.org/10.1007/s11135-014-0101-0

Bennink, M., Croon, M. A., Keuning, J., & Vermunt, J. K. (2014). Measuring Student Ability, Classifying Schools, and Detecting Item Bias at School Level, Based on Student-Level Dichotomous Items. *Journal of Educational and Behavioral Statistics, 39,* 180-202. https://doi.org/10.3102/1076998614529158

Cheong, Y. F. (2006). Analysis of School Context Effects on Differential Item Functioning Using Hierarchical Generalized Linear Models. *International Journal of Testing, 6,* 57-79. https://doi.org/10.1207/s15327574ijt0601_4

Cheong, Y. F., & Raudenbush, S. W. (2000). Measurement and Structural Models for Children's Problem Behaviors. *Psychological Methods, 5,* 477-495. https://doi.org/10.1037/1082-989X.5.4.477

Cho, S.-J., & Cohen, A. S. (2010). A Multilevel Mixture IRT Model with an Application to DIF. *Journal of Educational and Behavioral Statistics, 35,* 336-370. https://doi.org/10.3102/1076998609353111

Cho, S.-J., Cohen, A. S., & Bottge, B. (2013). Detecting Intervention Effects Using a Multilevel Latent Transition Analysis with a Mixture IRT Model. *Psychometrika, 78,* 576-600. https://doi.org/10.1007/s11336-012-9314-0

Chu, K. L., & Kamata, A. (2000). Nonequivalent Group Equating via 1-P HGLLM. *The Annual Meeting of the American Educational Research Association*, New Orleans, 24-28 April 2000.

Chu, K. L., & Kamata, A. (2005). Test Equating in the Presence of DIF Items. *Journal of Applied Measurement, 6*, 342-354.

Chung, S., & Houts, C. (2020). flexMIRT: A Flexible Modeling Package for Multidimensional Item Response Models. *Measurement: Interdisciplinary Research and Perspectives, 18*, 40-54. https://doi.org/10.1080/15366367.2019.1693825

Cohen, A. S., & Bolt, D. M. (2005). A Mixture Model Analysis of Differential Item Functioning. *Journal of Educational Measurement, 42*, 133-148. https://doi.org/10.1111/j.1745-3984.2005.00007

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*, 3-32. https://doi.org/10.1007/s11336-009-9134-z

Finch, W. H., & Finch, M. E. H. (2013). Investigation of Specific Learning Disability and Testing Accommodations Based Differential Item Functioning Using a Multilevel Multidimensional Mixture Item Response Theory Model. *Educational and Psychological Measurement, 73*, 973-993. https://doi.org/10.1177/0013164413494776

Fox, J.-P. (2004). Applications of Multilevel IRT Modeling. *School Effectiveness and School Improvement, 15*, 261-280. https://doi.org/10.1080/09243450512331383212

Fox, J.-P. (2005). Multilevel IRT Using Dichotomous and Polytomous Response Data. *The British Journal of Mathematical and Statistical Psychology, 58*, 145-172. https://doi.org/10.1348/000711005X38951

Fox, J.-P., & Glas, C. A. W. (2003). Bayesian Modeling of Measurement Error in Predictor Variables Using Item Response Theory. *Psychometrika, 68*, 169-191. https://doi.org/10.1007/BF02294796

Jang, Y., Kim, S. H., & Cohen, A. S. (2018). The Impact of Multidimensionality on Extraction of Latent Classes in Mixture Rasch Models. *Journal of Educational Measurement, 55*, 403-420. https://doi.org/10.1111/jedm.12185

Jilke, S., Meuleman, B., & Walle, S. V. D. (2015). We Need to Compare, but How? Measurement Equivalence in Comparative Public Administration. *Public Administration Review, 75*, 36-48. https://doi.org/10.1111/puar.12318

Kamata, A. (2001). Item Analysis by the Hierarchical Generalized Linear Model. *Journal of Educational Measurement, 38*, 79-93. https://doi.org/10.1111/j.1745-3984.2001.tb01117.x

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis* (pp. 10-11). Boston, MA: Houghton Mifflin.

Lee, W. Y., Cho, S. J., & Sterba, S. K. (2018). Ignoring a Multilevel Structure in Mixture Item Response Models: Impact on Parameter Recovery and Model Selection. *Applied Psychological Measurement, 42*, 136-154. https://doi.org/10.1177/0146621617711999

Liu, H., Liu, Y., & Li, M. (2018). Analysis of Process Data of PISA 2012 Computer-Based Problem Solving: Application of the Modified Multilevel Mixture IRT Model. *Frontiers in Psychology, 9*, 1372. https://doi.org/10.3389/fpsyg.2018.01372

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems* (pp. 16-20). Mahwah, NJ: Lawrence Erlbaum.

Lu, J., Zhang, J., & Tao, J. (2018). Slice-Gibbs Sampling Algorithm for Estimating the Parameters of a Multilevel Item Response Model. *Journal of Mathematical Psychology, 82*, 12-25. https://doi.org/10.1016/j.jmp.2017.10.005

Raudenbush, S. W., Johnson, C., & Sampson, R. J. (2003). A Multivariate, Multilevel Rasch Model with Application to Self-Reported Criminal Behavior. *Sociological Methodology, 33*, 169-211. https://doi.org/10.1111/j.0081-1750.2003.t01-1-00130.x

Reckase, M. D. (1985). The Difficulty of Test Items That Measure More than One Ability. *Applied Psychological Measurement, 9*, 401-412. https://doi.org/10.1177/014662168500900409

Reckase, M. D. (2007). *Multidimensional Item Response Theory* (pp. 60-63). New York: Springer.

Rost, J. (1990). Rasch Models in Latent Classes: An Integration of Two Approaches to Item Analysis. *Applied Psychological Measurement, 14*, 271-282. https://doi.org/10.1177/014662169001400305

Smit, A., Kelderman, H., & Flier, H. V. D. (2003). Latent Trait Latent Class Analysis of an Eysenck Personality Questionnaire. *MPR-Online, 8*, 23-50.

Spiegelhalter, D., Thomas, A., & Best, N. (2003). *WinBUGS (Version 1.4) [Computer Program]* (pp. 31-42). Cambridge: MRC Biostatistics Unit, Institute of Public Health.

Tay, L., Diener, E., Drasgow, F., & Vermunt, J. K. (2011). Multilevel Mixed-Measurement IRT Analysis: An Explication and Application to Self-Reported Emotions across the World. *Organizational Research Methods, 14*, 177-207. https://doi.org/10.1177/1094428110372674

Varriale, R., & Vermunt, J. K. (2012). Multilevel Mixture Factor Models. *Multivariate Behavioral Research, 47*, 247-275. https://doi.org/10.1080/00273171.2012.658337

Vermunt, J. K. (2003). Multilevel Latent Class Models. *Sociological Methodology, 33*, 213-239.

https://doi.org/10.1111/j.0081-1750.2003.t01-1-00131.x

Vermunt, J. K. (2007). Multilevel Mixture Item Response Theory Models: An Application in Education Testing. *The Bulletin of the International Statistical Institute 56th Session*, Lisbon, 25 September 2007.

Vermunt, J. K. (2008). Multilevel Latent Variable Modeling: An Application in Education Testing. *The Australian Journal of Statistics, 37*, 285-299.

Zwinderman, A. H. (1991). A Generalized Rasch Model for Manifest Predictors. *Psychometrika, 56*, 589-600. https://doi.org/10.1007/BF02294492