Advances in Psychology 心理学进展, 2015, 5, 365-375
Published Online June 2015 in Hans. http://www.hanspub.org/journal/ap
http://dx.doi.org/10.12677/ap.2015.56048

Hans汉斯

Multidimensional Item Response Theory: Psychometric Models, Parameter Estimation and Application

Peng Wang*, Xinli Zhu, Fang Wang

School of Psychology, Shandong Normal University, Jinan Shandong
Email: *pengsdnu@163.com

Received: Jun. 3rd, 2015; accepted: Jun. 23rd, 2015; published: Jun. 26th, 2015

Abstract

Multidimensional Item Response Theory (MIRT) is the new development of modern psychometric theories. The psychometric models, parameter estimation and application of MIRT are overviewed in this paper. It is concluded that the development of MIRT models should be combined with cognitive construct, the method of MCMC should be used to enhance the parameter estimation of MIRT, the research of the mixed MIRT should be strengthened, and the method of maximum information should be used to get the total score of a test.

Keywords

Multidimensional Item Response Theory, Psychometric Model, Parameter Estimation

多维项目反应理论的计量模型、参数估计及应用

王　鹏*，朱新立，王　芳

山东师范大学心理学院，山东　济南
Email: *pengsdnu@163.com

收稿日期：2015年6月3日；录用日期：2015年6月23日；发布日期：2015年6月26日

*通讯作者。

摘　要

多维项目反应理论是现代心理测量理论的新发展，文章对多维项目反应理论的计量模型、参数估计及应用进行了综述，认为多维项目反应理论模型开发应与认知结构相结合，马尔科夫链蒙特卡洛方法能较好地实现多维项目反应理论的参数估计，应加强多种题型、多种多维项目反应模型结合的参数估计研究，建议使用基于最大信息量法的多维项目反应理论模型计算测验的总分。

关键词

多维项目反应理论，计量模型，参数估计

1. 引言

相对经典测验理论(Classic Test Theory, CTT)而言，项目反应理论(Item Response Theory, IRT)在项目和测验质量分析、题库建设、计算机自适应测验编制等方面的作用，越来越受到研究者的青睐(戴海琦，2010)。近年来，随着认知科学、数学和计算机科学的发展，IRT 模型的假设、理论和实际应用也出现一些新的进展，其中之一就是由以往注重单维模型(Unidimensional IRT, UIRT)向单维和多维模型(Multidimensional IRT, MIRT)并重转变。MIRT 的提出是为了更好地对被试在完成一项测验任务时需要的多种能力、项目特征与答对概率之间的关系进行模型化。MIRT 被认为是近 20 年来测验理论发展的主要新进展之一(康春花，辛涛，2010)。被试在对某一特定测验题目作答时，可能使用了不只一种能力；同样地，考试中的问题很可能需要许多技能和能力才能答对。特别是测量复杂的知识领域如自然科学时更是如此。尽管 UIRT 在一定条件下是有用的，但还是需要更复杂的 IRT 模型以准确反映被试和题目之间相互作用的复杂性。MIRT 模型的发展正符合这一需要(康春花，辛涛，2010)。

2. 多维项目反应理论的计量模型

Bock，Horst，McDonald，Samejima 等学者的研究工作，推动了 MIRT 的发展。MIRT 的数学表达式包括多个描述测验中体现的被试知识和能力特征的参数，以及项目难度和区分度等题目特性的参数(Reckase, 2009)。MIRT 建立的是被试在多维空间中的位置与其项目作答概率之间的关系。根据能力各维度$\theta_1,\theta_2,\cdots,\theta_m$之间的关系，目前有两种主要的 MIRT 模型。

如果模型基于θ各维度的线性组合，这种线性组合可用在正态肩形函数或逻辑斯蒂克函数中，并通过这些函数来表示对某反应的概率。这意味着同样的总和可由很多种θ维度的组合得到。如果某一个θ维度值较低，其它值相当高的维度θ仍然可以使总和维持不变。具有这种性质的模型一般被称为“补偿模型”(compensatory models)。

另一种模型的特点是，每一个题目涉及的认知任务可以分割成几个部分构成，每一个部分的作答概率可以使用单维 IRT 模型表示。而正确回答题目的概率则是每一部分概率的乘积。事实上，正确作答项目的概率肯定不会高于作答每一部分的最大概率，这降低了高维度值θ对低维度值θ的补偿。这类模型被称为“非补偿模型”(noncompensatory models)。但一个高维度值的θ确实能产生相对较高的作答概率，有一定程度的补偿作用，所以 Reckase (2009)也把这种模型称为“部分补偿模型”(partially compensatory models)。

除了补偿和非补偿这种分类，MIRT 也可以根据题目记分的形式分为二分模型和多级模型。下面结合这两种分类方法，对 MIRT 模型进行介绍。

2.1. 二分法记分的 MIRT 模型

由于探讨二分数据因素结构的需要，二分数据的 MIRT 模型通过扩展二分 UIRT 的形式开始出现(Yao, Schwarz, 2006)。

2.1.1. 补偿模型

2 参数逻辑斯蒂克模型(M-2PL)的多维扩展形式可表示为：

$$P\left(U_{ij}=1\middle|\boldsymbol{\theta}_j,\boldsymbol{a}_i,d_i\right)=\frac{e^{\boldsymbol{a}_i\odot\boldsymbol{\theta}_j^{\mathrm{T}}+d_i}}{1+e^{\boldsymbol{a}_i\odot\boldsymbol{\theta}_j^{\mathrm{T}}+d_i}}$$

其中，e 的指数部分向量 $\boldsymbol{a}_i$ 和 $\boldsymbol{\theta}_j$ 的运算方式是：

$$\boldsymbol{a}_i\odot\boldsymbol{\theta}_j^{\mathrm{T}}+d_i=a_{i1}\theta_{j1}+a_{i2}\theta_{j2}+\cdots+a_{im}\theta_{jm}+d_i=\sum_{k=1}^{m}a_{ik}\theta_{jk}+d_i$$

其中 m 为维度个数。该指数部分是向量 $\boldsymbol{\theta}$ 元素的线性组合，参数 d 是截距项，向量 $\boldsymbol{a}_i$ 的元素作为斜率参数。

显然，MIRT的被试参数和项目参数要比同样形式的UIRT多很多，但由于补偿模型的各 $\boldsymbol{\theta}$ 元素之间有互补关系，所以可将各 $\boldsymbol{\theta}$ 元素视为一个整体，与其对应的项目参数可以多维项目参数的形式表示。Reckase (1985)提出多维项目区分度(Multidimensional Item Discrimination, MDISC)和多维项目难度(Multidimensional Item Difficulty, MID)的表达式：

$$MDISC_i=\left(\sum_{k=1}^{m}a_{ik}^2\right)^{1/2}$$

$$MID_i=\frac{-d_i}{\left(\sum_{k=1}^{m}a_{ik}^2\right)^{1/2}}$$

$MDISC_i$ 和 MID_i 的值和 UIRT 模型中的 a 参数和 b 参数有相同的解释。在 M-2PL 中，$MDISC_i$ 是 $\boldsymbol{\theta}$ 空间中，从原点出发到项目反应概率曲面上最大斜率点的方向斜率，是各维度综合区分能力的体现。MID_i 是 $\boldsymbol{\theta}$ 空间从原点沿最大斜率点的方向到作答概率 0.5 的等高线的距离。

在多维3参数逻辑斯蒂克模型(M-3PL)中，伪猜测参数 c_i 的意义变为 $\boldsymbol{\theta}$ 向量各维度值都极低时的答对概率(Reckase, McKinley, 1991)。M-3PL 的形式为：

$$P\left(U_{ij}=1\middle|\boldsymbol{\theta}_j,\boldsymbol{a}_i,c_i,d_i\right)=c_i+\left(1-c_i\right)\frac{e^{\boldsymbol{a}_i\odot\boldsymbol{\theta}_j^{\mathrm{T}}+d_i}}{1+e^{\boldsymbol{a}_i\odot\boldsymbol{\theta}_j^{\mathrm{T}}+d_i}}$$

此外，在二分法记分的多维扩展模型中，还有多维Rasch模型(Adams, Wilson, Wang, 1997)、多维正态肩形模型(Bock, Schilling, 2003)。这些多维模型都是补偿模型，也可用于探索性或验证性非线性因素分析。

2.1.2. 非补偿模型

补偿模型的特点是，即使另外 $m-1$ 个维度值都很低，被试如果在m维度上值很高，那么仍然可以在项目上有很高的正确作答概率。但有些题目不符合上述情况，如做一道数学题需要数学计算和阅读两种

能力，数学计算能力很高的人如读不懂题目，仍然无法答对题目。根据这种情况，Sympson (1978)提出了下面的多维模型。

$$P\left(U_{ij}=1\middle|\boldsymbol{\theta}_j,\boldsymbol{a}_i,\boldsymbol{b}_i,c_i\right)=c_i+\left(1-c_i\right)\left(\prod_{k=1}^{m}\frac{e^{1.7a_{ik}\left(\theta_{jk}-b_{ik}\right)}}{1+e^{1.7a_{ik}\left(\theta_{jk}-b_{ik}\right)}}\right)$$

上式中，连乘部分采用的是2参数逻辑斯蒂克模型，连乘的每一部分各代表成功完成项目的一个维度任务的概率，项目的这些维度任务被认为是相互独立的，所以项目的答对概率采取了乘积的形式。但Sympson认为每个题目只有一个非零的下渐近线 c_i，对每个题目不同的维度任务来说，没有相应的“伪猜测参数”。

实际上，上式是在3PL模型的基础上扩展成多维的，后来也有研究者使用了更简化的模型，Whitely (1980)称之为多成分潜在特质模型(multicomponent latent trait model, MLTM)，Maris (1995)称之为“连接式Rasch模型”(conjunctive Rasch model)，用以表示多个相互独立的认知成分对一个题目作答正确概率的影响，其表达式为：

$$P\left(U_{ij}=1\middle|\boldsymbol{\theta}_j,\boldsymbol{b}_i\right)=\prod_{k=1}^{m}\frac{e^{\left(\theta_{jk}-b_{ik}\right)}}{1+e^{\left(\theta_{jk}-b_{ik}\right)}}$$

补偿模型和非补偿模型在人-题交互作用的表达方式上是不同的。非补偿模型使用的题目由不同的维度任务构成，这些不同的维度任务又对应着各自需要的技能或知识，题目的完成依赖于每个维度任务的完成。补偿模型则更侧重整体的作用，所有技能和知识一起对题目的所有方面发生影响。这方面的比较研究不多，如Bolt和Lall (2003)发现在英语用法测验中，补偿模型对数据的拟合比非补偿模型好。归根到底，使用哪种模型取决于人们在实际题目上的反应机制。

2.2. 多级记分的 MIRT 模型

二分法记分的MIRT模型已有多年的发展，但多级记分的MIRT模型的研究相对较晚。到目前为止，有关研究也仅限于补偿模型的探讨。Muraki和Carlson (1993)扩展得到了多维等级反应模型，近年还有研究在一套混合记分的测验中综合应用了多维3参数逻辑斯蒂克模型和多维拓广分部评分模型(Yao, Schwarz, 2006)。

2.2.1. 多维分部评分模型

多维分部评分模型(MPCM)是将Rasch模型扩展到多维多级的形式(Kelderman, Rijkes, 1994)，可表示为：

$$P\left(u_{ij}=k\middle|\boldsymbol{\theta}_j\right)=\frac{e^{\left[\sum_{u=1}^{m}\left(\theta_{ju}-b_{iuk}\right)W_{iuk}\right]}}{\sum_{r=0}^{K_i}e^{\left[\sum_{u=1}^{m}\left(\theta_{ju}-b_{iur}\right)Wiur\right]}}$$

其中，

b_{iuk} 是项目i的维度为u、类别为k的难度参数，W_{iuk} 是预先确定的项目i在维度为u、类别为k的权重。

2.2.2. 多维拓广分部评分模型

分部评分项目 i，能力向量为 $\boldsymbol{\theta}_j$ 的被试，在 MGPC 中作答类别为 k-1 的概率为(Yao, Schwarz, 2006)：

$$P_{ijk}=P\left(X_{ij}=k-1\middle|\boldsymbol{\theta}_j,\boldsymbol{\beta}_i\right)=\frac{e^{\left((k-1)\boldsymbol{\beta}_i\odot\boldsymbol{\theta}_j^T-\sum_{t=1}^{k}\beta_{\delta_t i}\right)}}{\sum_{m=1}^{K_i}e^{\left((m-1)\boldsymbol{\beta}_i\odot\boldsymbol{\theta}_j^T-\sum_{t=1}^{k}\beta_{\delta_t i}\right)}}$$

其中

$X_{ij}=0,\cdots,K_i-1$指被试 j 对项目 i 的作答类别。

$\boldsymbol{\beta}_i=(\boldsymbol{\beta}_{i1},\cdots,\boldsymbol{\beta}_{iD})$指维度数为 D 的项目区分度参数向量。

$\boldsymbol{\beta}_{\delta_{ki}}(k=1,2,\cdots,K_i)$指阈限参数(或称$\alpha$参数)，令 $\boldsymbol{\beta}_{\delta_{1i}}=0$。

Yao 和 Schwarz (2006)对此类模型考察，其参数估计方法已由体现 Rasch 模型特色的条件极大似然估计(CML)变成了马尔可夫链-蒙特卡洛方法(MCMC)。Yao 和 Schwarz 分别使用了多维 3 参数逻辑斯蒂克模型(M-3PL)和多维拓广分部评分模型对多重选择题和论文式题目进行了参数估计。Yao 和 Schwarz 对多维模型的项目参数、项目信息量、测验信息量、拟合评价等问题进行了较全面的计算。

2.2.3. 多维等级反应模型

Muraki 和 Carlson (1993)将单维等级反应模型扩展为多维模型，项目反应函数使用的是正态肩形模型的形式。和单维模型一样，多维模型也假定题目作答由一些步骤构成，能做到第 k 步，必然已经成功完成前 k-1 步。这类模型也可用于后面类别包括前面所有类别的评定量表。比如，要评定一下在某项目上用的时间，如果某一个评定类别是“用了一个小时”，则意味着所有低于一个小时的类别均包含在内。多维等级反应模型计算被试得 k 分的概率与单维模型相似，其表达式为：

$$P\left(u_{ij}=k\middle|\boldsymbol{\theta}_j\right)=P^*\left(u_{ij}=k\middle|\boldsymbol{\theta}_j\right)-P^*\left(u_{ij}=k+1\middle|\boldsymbol{\theta}_j\right)$$

其中 $P^*\left(u_{ij}=0\middle|\boldsymbol{\theta}_j\right)=1$， $P^*\left(u_{ij}=m_i+1\middle|\boldsymbol{\theta}_j\right)=0$

等级反应模型的正态肩形模型形式是，

$$P\left(u_{ij}=k\middle|\boldsymbol{\theta}_j\right)=\frac{1}{\sqrt{2\pi}}\int_{\boldsymbol{a}_i^{\mathrm{T}}\odot\boldsymbol{\theta}_j+d_{i,k+1}}^{\boldsymbol{a}_i^{\mathrm{T}}\odot\boldsymbol{\theta}_j+d_{ik}}e^{\frac{-t^2}{2}}\mathrm{d}t$$

将式 1-4-24 代入 1-4-23，可得，

$$P\left(u_{ij}=k\middle|\boldsymbol{\theta}_j\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\boldsymbol{a}_i^{\mathrm{T}}\odot\boldsymbol{\theta}_j+d_{ik}}e^{\frac{-t^2}{2}}\mathrm{d}t-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\boldsymbol{a}_i^{\mathrm{T}}\odot\boldsymbol{\theta}_j+d_{i,k+1}}e^{\frac{-t^2}{2}}\mathrm{d}t$$

上述各种模型都只适用于某一种情况，如题目二分或题目多级，包括或不包括伪机遇参数等。但实际考试中，为提高测验的有效性，往往将不同类型的题目结合在一起使用。目前单参数逻辑斯蒂克模型和分部评分模型、2 参数逻辑斯蒂克模型和等级反应模型等的结合，已用于混合项目类型(mixed item types)的测验(Baker, Kim, 2004)。但目前 MIRT 的混合项目类型的研究较为少见。

杨向东(2010)认为，现有心理测量学模型，缺少实质理论的支持。他认为，揭示一般规律的认知研究与揭示个别差异的测量学研究相整合，构建认知测量模型(cognitive psychometric model)，是将来模型发展的方向。事实上，前文提到的补偿模型和非补偿模型，在一定程度上也体现了被试问题解决过程的认知模型与心理测量学模型结合的特点。但还需要进一步结合具体领域的认知加工模型，并在测验开发之初就根据认知模型编制测验，在此基础上依据经验数据进行验证。另外，认知测量模型还需逐渐与具体情境结合，以解决复杂认知领域的建模、分析和解释等问题。这里并不是主张抛弃这些心理测量学模型，而是要加强实质性的认知加工模型与心理测量学模型之间的联结。

3. 多维项目反应理论模型的参数估计方法

3.1. 常见的参数估计方法

在项目反应理论发展过程中，出现的参数估计方法主要有(Baker，Kim，2004；漆书青，2003)：条

件极大似然估计(conditional maximum likelihood estimation, CMLE)、联合极大似然估计(joint maximum likelihood estimation, JMLE)、边际极大似然估计/EM 算法(marginal maximum likelihood estimation and an EM algorithm, MMLE/EM)、边际贝叶斯估计(marginalized Bayesian estimation, MBE)等。

上述方法，基于 UIRT 的参数估计提出，有的后来逐渐用于 MIRT 的参数估计。特别是结合几个 MIRT 分析的统计软件，如 TESTFACT (Bock, Schilling, 2003)、NOHARM (Fraser, McDonald, 1988)、ConQuest (Wu, Adams, Wilson, 1997)、POLYFACT (Muraki, 1999)等，降低了掌握统计学知识的要求，这些方法的使用者也越来越多。TESTFACT 采用的估计方法是边际极大似然估计/EM 算法，可用项目间四分相关(interitem tetrachoric correlations)对二分法记分项目进行因素分析。该软件使用的 MIRT 模型是多维 2 参数正态肩形模型，可得到项目和被试参数，使用者如输入下渐近线的信息，也可运行多维 3 参数正态肩形模型，但程序不能估计伪猜测参数，同时 TESTFACT 的维度数限制为 15。NOHARM (Normal-Ogive Harmonic Analysis Robust Method)也是 MIRT 软件的代表之一。NOHARM 的参数估计与 TESTFACT 不同，它通过被试-项目矩阵来拟合多维肩形补偿性 MIRT 模型，拟合时采用的是多项式近似、未加权最小二乘法，可进行探索性因素分析，其维度数可达到 50。但 NOHARM 不提供 $\boldsymbol{\theta}$ 的估计值，所以，可通过 NOHARM 得到项目参数估计，然后输入 TESTFACT 得到能力参数估计。ConQuest 主要用于估计 Rasch 模型一族的参数，这种项目和被试参数的估计有可观察的充分统计量。ConQuest 能估计 Rasch 模型的 MIRT 扩展形式的参数，处理二分法记分的数据，采用的估计方法是极大似然估计。POLYFACT 能处理多级数据，使用的模型是多维等级反应模型或拓广分部评分模型；它使用的是边际极大似然估计，执行的是验证性的参数结构，将题目与某一个特定维度相关联。

上述软件是应用较广的 MIRT 分析工具，可以发现，这些软件包括的 MIRT 模型还相对较少；有的软件不能同时输出项目参数估计和能力参数估计；还有的软件对模型的维度数量有一定的限制；主要处理二分数据的情形；还不能处理测验中包括混合记分类型题目的情况。这些软件的局限性，在一定程度上是由参数估计方法决定的，当 IRT 模型中的参数的个数或维度过多时，上述参数估计方法可能难于或无法实现模型的参数估计。在马尔科夫链蒙特卡洛方法(MCMC)引入 IRT 参数估计研究之后，许多复杂、高维模型的参数估计成为可能，MCMC 方法是一种全新的参数估计方法(Baker，Kim，2004；涂冬波，漆书青，蔡艳，戴海琦，丁树良，2008)。下面介绍用于 IRT 参数估计的 MCMC 方法及其软件的研究情况。

3.2. 基于 MCMC 方法的参数估计

MCMC 方法是一种动态的计算机模拟技术，它是根据任一多元理论分布，特别是以贝叶斯推断为中心的多元后验分布来模拟随机样本的一种方法。Monte Carlo 方法的一个基本步骤是产生随机数，使之服从一个概率分布 $\pi(X)$。目前常用的 MCMC 方法，包括 Metrophs-Hasting (M-H)方法和 Gibbs 方法。M-H 方法就是在以 $\pi(X)$ 为平稳分布的马尔科夫链上产生相互依赖的样本。MCMC 方法本质上是一个 Monte Carlo 综合程序，它的随机样本的产生与一条马尔科夫链有关。基于条件分布的迭代取样是另外一种重要的 MCMC 方法，其中最著名的特殊情况就是 Gibbs 抽样，它的特征是其潜在的马尔科夫链是通过分解一系列条件分布建立起来(赵琪，2007)。MCMC 方法估计 IRT 参数的基本特点是：采用能力参数和项目参数交替迭代的方法生成马尔科夫链；然后采取截然不同于极大似然法的思路，发挥计算机模拟技术的优势，采集足够大的状态样本，用初等的方法来估计模型参数，避开了 EM 算法中的复杂计算，提高了估计的成功率(王权，2006)。

多参数、高维度 IRT 模型的参数估计，能较充分发挥 MCMC 方法的优势。国内外研究者已开始将 MCMC 方法用于 MIRT 的参数估计。Jiang (2005)采用 MCMC 方法的 Gibbs 抽样估计多维 3 参数逻辑斯

蒂克模型的项目参数和被试参数。通过模拟研究考察了 1 个维度、2 个维度和 5 个维度的情况。均方根误差(Root Mean Square Error, RMSE)结果显示，MCMC 方法估计的精度较高。Jiang 将自编 MCMC 程序与 TESTFACT 的估计结果进行了比较，发现前者的估计精度优于后者。Zhang (2007)也试图找到一种估计补偿性多维模型项目参数的合理方法。MCMC 方法的估计结果中，项目区分度和难度的精度很高，伪猜测参数的精度相对较好。但作者对 MCMC 方法的计算负荷表示了担忧，如一般 1 条链的迭代要 10,000 次以上。初值的设置更接近项目参数，可能会减少计算负荷。Bolt 和 Lall (2003)使用 MCMC 方法比较了多维补偿模型(M-2PL)和多维非补偿模型(MLTM)的参数估计。采用模拟研究考察两种模型参数的返真性，设置样本(1000, 3000)，项目数量(25, 50)，能力维度之间的相关(0.0, 0.3, 0.6)。结果显示，采用 MCMC 的 M-H 算法能够保证两类模型较高的返真性，但 MLTM 随着能力维度相关的变高返真性变低。付志慧(2010)探讨了多维 3 参数逻辑斯蒂克模型的 Gibbs 抽样法，模拟实验表明，由 Gibbs 抽样所得的 Bayes 后验估计及其标准差与 BILOG-MG 软件所得结果是具有可比性的；还研究了多维等级反应模型下 MCMC 方法对含缺失数据的参数估计，发现在项目参数估计过程忽略缺失数据会带来相当大的偏差，而且待测潜在特性和缺失数据的相关性越强，偏差越大；反之，如不忽略缺失，并对缺失过程建模，偏差会减少。

目前，多数研究者用 MCMC 方法估计 MIRT 项目参数和被试参数，一般是以自编程序，或通过 WinBUGS 执行 MCMC 过程，一段程序一般考察一个多维模型，未能实现对混合记分项目类型的题目参数和被试参数联合估计，亦即不能解决两个或两个以上 MIRT 模型同时估计的问题。Yao 等人开发的 BMIRT (Bayesian Multivariate Item Response Theory)软件，为问题的解决提供了一种选择(Reckase, 2009; Yao, Schwarz, 2006)。BMIRT 用 MCMC 方法中的 M-H 算法，估计多种 MIRT 模型的项目和被试参数，可处理二分和多级数据，并能实现混合记分项目类型中多个 MIRT 模型的同时估计，现对其进行简要介绍。

BMIRT 使用的是 MCMC 方法中的 M-H 算法，其目标是通过一个稳定的分布模拟观察数据。所以，马尔科夫链的稳定分布是给定作答矩阵的模型参数的后验分布。

所有项目参数矩阵可表示为：

$$\beta = \left(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_i, \cdots, \boldsymbol{\beta}_M\right)^{\mathrm{T}}$$

则似然方程是，

$$P\left(X \mid \theta, \beta\right) = \prod_{j=1}^{N} P\left(\boldsymbol{X}_j \middle| \boldsymbol{\theta}_j, \beta\right) = \prod_{j=1}^{N}\prod_{i=1}^{M} P\left(\boldsymbol{X}_{ij} \middle| \boldsymbol{\theta}_j, \beta_i\right)$$

如测验中有混合记分类型题目，则 1-4-27 式中的 $P\left(X_{ij} \middle| \boldsymbol{\theta}_j, \beta_i\right)$ 可根据不同的 MIRT 模型计算。

记 $P\left(\theta \middle| \lambda\right)$ 为某被试群体能力分布，分布参数为 λ。譬如 θ 符合正态分布，则 $\lambda = \left(\mu, \sigma\right)$，$\mu$ 为平均数，σ 为方差和协方差矩阵。应用贝叶斯原理，其联合后验分布可写为，

$$P\left(\theta, \beta, \lambda \middle| X\right) \propto P\left(X \middle| \theta, \beta, \lambda\right) P\left(\theta \middle| \beta, \lambda\right) P\left(\beta\right) P\left(\lambda\right) = P\left(X \middle| \theta, \beta\right) P\left(\theta \middle| \lambda\right) P\left(\beta\right) P\left(\lambda\right)$$

其中，$\propto$ 代表“正比于”。

$$P\left(X \middle| \theta, \beta\right) = \prod_{j=1}^{N}\prod_{i=1}^{M} P\left(X_{ij} \middle| \boldsymbol{\theta}_i, \boldsymbol{\beta}_j\right)$$

$$P\left(\theta \middle| \lambda\right) = \prod_{j=1}^{N} P\left(\boldsymbol{\theta}_j \middle| \lambda\right)$$

MCMC方法中的转移核t (transition kernel)是在某个给定当前参数值的状态的条件密度：

$$t\left[\left(\boldsymbol{\theta}^k, \boldsymbol{\beta}^k\right), \left(\boldsymbol{\theta}^{k+1}, \boldsymbol{\beta}^{k+1}\right)\right] = P\left[\left(\boldsymbol{\theta}^{k+1}, \boldsymbol{\beta}^{k+1}\right) \middle| \left(\boldsymbol{\theta}^k, \boldsymbol{\beta}^k\right)\right]$$

其中上标 k 和 $k+1$ 表示前一状态和后一状态。在 BMIRT 中，将转移核设置为两个条件概率的乘积：

$$t\left[\left(\boldsymbol{\theta}^{k},\boldsymbol{\beta}^{k}\right),\left(\boldsymbol{\theta}^{k+1},\boldsymbol{\beta}^{k+1}\right)\right]=P\left(\boldsymbol{\theta}^{k+1}\middle|\boldsymbol{X},\boldsymbol{\beta}^{k}\right)P\left(\boldsymbol{\beta}^{k+1}\middle|\boldsymbol{X},\boldsymbol{\theta}^{k+1}\right)$$

因为转移核分成了两个成分，对观察值的抽样也可以分两步。第一步从条件分布中抽取 $\boldsymbol{\theta}$；然后这些抽取的值可以设定项目参数的条件分布。

如果转移核可以用于马尔科夫链的模拟步骤，那么每一步的观察数据作为一个抽样样本保存，一般需要很长的马尔科夫链才能达到稳定分布。有时初始值的设置也会影响结果。在迭代初期的观察数据一般应予丢弃，这一阶段叫做“烧制”(burn-in)阶段。“burn-in”阶段之后的观察数据，用来估计稳定分布的项目参数和被试参数。分布的标准差即为参数估计的标准误。

基于 MCMC 方法进行参数估计，BMIRT 还可以使用对含有缺失值的二分、多级数据，以及 M-2PL、M-3PL、M-PCM、M-GRM 等进行同时估计。在同样条件下与 TESTFACT、NOHARM 等软件进行比较，有良好的参数返真性。BMIRT 在处理参数较多的多级、高维、混合模型时优势比较明显。

4. 多维项目反应理论在测验分析中的应用

随着 MIRT 的参数估计问题得以解决，它在测验分析中的应用价值也通过相关研究体现出来。MIRT 有助于理解项目和测验究竟测什么这个问题，能够分析不同能力维度的区分能力，并可以报告被试在各个能力维度上的表现(Zhang, 2007)。MIRT 几乎涵盖了 UIRT 发挥作用的所有方面，如测验结构效度证据、测验记分、测验等值(谢晶，张厚粲，2009)、计算机化自适应测验开发(computerized adaptive test, CAT)、项目功能差异检测(differential item functioning, DIF)，等等。根据本文的目的，这里主要综述 MIRT 在测验结构效度证据、测验记分两方面的研究。

4.1. 多维项目反应理论的结构效度证据研究

检验测验的维度，为测验提供结构效度证据，几乎是目前文献中 MIRT 应用的最多的情况。

Akerman (1992)认为根据项目参数(如区分度参数)所表示的每个项目向量的角度代表了该项目测量哪个能力维度。一个测验最有效的项目，可以从项目向量图中发现。也就是说，测量相似能力维度的有效项目应在一个范围明确的扇区中，Akerman 称之为“效度扇区”。

基于 IRT 的项目因素分析，亦称“全息项目因素分析”(Full-information item factor analysis)，能充分利用每名被试的作答向量，可深入考察潜在因素的数量(Bock, Gibbons, Muraki, 1988)，越来越受到研究者的重视。目前，为进一步考察特定潜在特质的多维性，研究者们借 MIRT 的一种简化形式——双因子模型(Bifactor Model)，逐渐积累着相关的证据(Gibbons, Bock, Hedeker, et al., 2007)。双因子分析模型要求：每个题目在普通因素(the general factor)(下面矩阵的第一列)上都有非零负荷；但每个题目仅在一个群因素(the group factors)上有非零负荷；普通因素之间及与群因素之间正交。双因子模型的一个优势就是能够简化似然方程(Akerman, 1992)。

$$\alpha=\begin{bmatrix}\alpha_{11} & \alpha_{12} & 0\\ \alpha_{21} & \alpha_{22} & 0\\ \alpha_{31} & \alpha_{32} & 0\\ \alpha_{41} & \alpha_{42} & 0\\ \alpha_{51} & 0 & \alpha_{53}\\ \alpha_{61} & 0 & \alpha_{63}\\ \alpha_{71} & 0 & \alpha_{73}\end{bmatrix}$$

这种结构在教育、心理测量中并不少见，如阅读理解测验往往考核一个主要的目标技能和不同段落内容体现的多种知识领域技能(Gibbons, Hedeker, 1992)。可以说，涉及到题组或分量表的教育与心理测验，往往都可以通过 MIRT 的双因子模型考察其结构。如，Fukuhara (2009)通过双因子结构的 MIRT 分析了带有题组的数据，模拟研究表明对题目效应以及 DIF 的考察可获得较高的精度。以往研究已考察了双因子模型与多维 Samejima 等级评分模型的结合，显然，双因子模型也可以与其他 MIRT 模型相结合，如前文提到的多维分部评分模型等。

在分析测验维度时，如何判断不同维度对数据的拟合情况，是很有争议的问题。Berger 和 Knol (1990)用模拟数据比较了 MIRT 的几种拟合指标，结果显示 χ^2 检验相对不可信，χ^2 差异、AIC (Akaike's information criterion)指标表现较好，但仍需更多的研究确认。另外，区分度参数矩阵的结构也是非常重要的参考信息。

4.2. CTT、UIRT、MIRT 在测验记分中的应用的比较

由于测验理论的发展，有的测验使用者主要基于 CTT 的方法记分，有的则开始使用 IRT 方法指导测验记分。下面结合 CTT、UIRT、MIRT 框架，介绍不同记分方法的比较研究。

在 CTT 框架中，将被试在每个题目上的得分累加得到总分是一种常见的方式，这种分数实际上将每一个题目的得分视为同等精度、同等重要的分数，显然与客观情况不符。Rotou 等人(2001)假定真分数已知，通过模拟标准化测验的二分数据，比较 CTT 方法未加权得分、CTT 方法加权得分与 MIRT 方法得分对真分数的“返真性”。其中 CTT 的加权是被试项目得分乘以项目与总分的点二列(point biserial)相关累加起来，MIRT 的得分是根据项目反应函数计算的正确作答概率累加之和。结果显示，基于 MIRT 方法的加权分的返真性优于 CTT 未加权方法和加权方法。

DeMars (2005)用 Bifactor 方法、MIRT、UIRT 对包括两个分测验(subscale)的考试进行了分析，发现所有方法计算出的得分相关几乎都很高，但 UIRT 计算出的得分与其他结果的相关较低。由于实测数据无法比较三种方法的优劣，Demars 又使用模拟数据评价三种估计方法的项目偏差(Bias)和均方根误差(RMSE)，发现 UIRT 的 Bias 和 RMSE 最大，而 Bifactor 和 MIRT 估计的精度较高、比较接近。

一般考试都由几个部分或分测验(subscale)构成，学科专家编制的考试结构与因素分析的结构往往不一致(Reckase, 2009)，但人们记分的时候还是习惯用考试的各部分或分测验报告分数，并在分测验得分的基础上得到总分。上述研究证明，直接用原始分累加的方法可能误差较大。Yao (2010)通过对几种 IRT 模型的比较，建议使用 MIRT 模型报告分测验的分数和并用基于 MIRT 的最大信息量法报告测验总分。Yao 使用模拟和真实数据对四种方法进行比较：a) UIRT 模型；b) 高阶 IRT 模型(HO-IRT)，即维度 θ_{jl} 值是表示被试总能力的 θ_j 值的线性函数，$\theta_{jl}=\lambda_l\theta_j+\eta_{jl}$，同时在每个分测验内部使用 UIRT 模型；c) MIRT 模型；d) Bifactor 模型。其中 MIRT 模型与 Bifactor 模型的区别是，MIRT 模型不设置一个普通因素，而只是将各个分测验作为相应的维度，如 Yao 使用的测试包括 5 个分测验，则 MIRT 的维度为 5，每个维度对应一个分测验包括的项目；Bifactor 则包括 6 个维度，其中一个是普通因素。模拟研究显示，所有条件下参数估计、分测验得分的返真性 MIRT 比 HO-IRT 略好，但差异不大，二者比 UIRT、Bifactor 精度高；总分的返真性比较发现，Bifactor 和 UIRT 的精度接近，基于最大信息量法的 MIRT 要显著优于 Bifactor、UIRT、HO-IRT。

5. 未来研究展望

项目反应理论考察被试特质与项目特征之间交互作用的方式，描述二者影响项目作答概率的数学模型。根据项目记分的类别，项目反应理论可分为二分法记分和多级记分两种模型；根据被试特质的维度

数量，可分为单维项目反应理论和多维项目反应理论。随着模型的复杂化，模型参数的估计方法也经历了由条件极大似然估计、联合极大似然估计、边际似然估计、贝叶斯估计等的变化，近年来 MCMC 方法在高维、多级的多参数 IRT 模型中的作用，越来越受到研究者的重视(涂冬波，蔡艳，戴海琦，丁树良，2011)。

多维项目反应理论模型的发展需与心理学实质理论特别是认知加工理论相结合，这样不仅能发挥 MIRT 模型在选拔性考试、题库建设等方面的作用，而且能具体分析 MIRT 模型所考察的能力结构及其内部加工过程，从而为测验编制、认知诊断、补救教学等方面提供更明确的参考信息。这种途径开发的认知测量模型，将逐渐摆脱以往测验开发中“数据驱动”模型的影响。

现实测验的形式不断演变，有些复杂题目的记分类型往往不是单一的，被试体现出的实践技能一般也是多维的，需要加强多种题型、多种 MIRT 模型结合的参数估计研究及其他心理测量学分析。

项目反应理论在测验的误差分析方面，通过项目信息函数选取合适的项目，通过测验信息函数对特定能力点被试的估计误差进行分析，对整个测验的估计准确性进行考察，这是经典测量理论无法实现的方面；项目反应理论还可以通过 MIRT 对测验的维度进行分析，以提供测验结构效度证据；使用 MIRT 计算分测验的得分，在此基础上使用基于最大信息量法的 MIRT 计算测验的总分，相比其它模型的计算方法能提供较高的估计精度。

基金项目

全国教育科学“十二五”规划教育部重点 2012 年度课题(DIA120273)。

参考文献 (References)

戴海琦(2010). **心理测量学**. 北京: 高等教育出版社.

付志慧(2010). **多维项目反应模型的参数估计**. 硕士论文, 吉林大学, 吉林.

康春花, 辛涛(2010). 测验理论的新发展: 多维项目反应理论. **心理科学进展**, 3 期, 530-536.

漆书青(2003). **现代测量理论在考试中的应用**. 武汉: 华中师范大学出版社.

涂冬波, 蔡艳, 戴海琦, 丁树良(2011). 多维项目反应理论: 参数估计及其在心理测验中的应用. **心理学报**, 11 期, 1329-1340.

涂冬波, 漆书青, 蔡艳, 戴海琦, 丁树良(2008). IRT 模型参数估计的新方法——MCMC 算法. **心理科学**, 1 期, 177-180.

王权(2006). “马尔科夫链蒙特卡洛”(MCMC)方法在估计 IRT 模型参数中的应用. **考试研究**, 4 期, 45-63.

谢晶, 张厚粲(2009). 测验等值: 从 IRT 到 MIRT. **心理学探新**, 5 期, 67-71.

杨向东(2010). 测验项目反应机制与心理测量模型假设的对应性分析. **心理科学进展**, 8 期, 1349-1358.

赵琪(2007). **MCMC 方法研究**. 硕士论文, 山东大学, 济南.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 4*, 255-278.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-24.

Akerman, T. A. (1992). Assessing construct validity using multidimensional item response theory. *Proceedings of the Annual Meeting of the American Educational Research Association*, San Francisco, 20-24 April 1992.

Baker, F. B., & Kim, S. (2004). *Item response theory parameter estimation techniques* (2nd ed.). New York: Marcel Dekker, Inc.

Berger, M. P. F., & Knol, D. L. (1990). *On the assessment of dimensionality in multidimensional item response theory models*. Research Report from the Division of Educational Measure and Data Analysis, Enschede: University of Twente.

Bock, R. D., & Schilling, S. G. (2003). IRT based item factor analysis. In M. du Toit (Ed.), *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measure-*

ment, 12, 261-280.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27,* 395-414.

DeMars, C. E. (2005). Scoring subscales using multidimensional item response theory models. *Proceedings of the Annual Meeting of the American Psychological Association (ED496242),* Washington DC, 18-21 August 2005.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23,* 267-269.

Fukuhara, H. (2009). *A differential item functioning model for testlet-based items using a bi-factor multidimensional item response theory model: A Bayesian approach*. Ph.D. Thesis, Tallahassee, FL: Department of Educational Psychology and Learning Systems, Florida State University.

Gibbons, R. D., & Hedeker, D. (1992). Full information item bi-factor analysis. *Psychometrika, 57,* 423-436.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K. et al. (2007). Full-information item bi-factor analysis of graded response data. *Applied Psychological Measurement, 31,* 4-19.

Jiang, Y. L. (2005). *Estimating parameters for multidimensional item response theory models by MCMC methods*. Ph.D. Thesis, East Lansing, MI: Michigan State University.

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59,* 149-176.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 45,* 479-494.

Muraki, E. (1999). *POLYFACT version 2 (Computer program)*. Princeton, NJ: Educational Testing Service.

Muraki, E., & Carlson, J. E. (1993). Full-information factor analysis for polytomous item responses. *Proceedings of the Annual Meeting of the American Educational Research Association,* Atlanta, 12-16 April 1993.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 30,* 469-492.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer Science + Business Media, LLC.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15,* 361-373.

Rotou, O., Elmore, P. B., & Headrick, T. C. (2001). Number correct scoring: Comparison between classical true score theory and multidimensional item response theory. *Proceedings of the Annual Meeting of the American Education Research Association,* Seattle, 10-14 April 2001.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis, MN: University of Minnesota.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479-494.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ConQuest: Generalized item response modeling software*. Victoria: ACER.

Yao, L. (2010). *Reporting valid and reliable overall scores and domain scores*. Monterey, CA: CTB/McGraw-Hill.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 30,* 469-492.

Zhang, L. T. (2007). *The estimation of multidimensional item response theory models*. Ph.D. Thesis, Columbia, SC: University of South Carolina.