

多维测验项目参数的估计： 基于 SEM 与 MIRT 方法的比较*

刘红云 骆方 王玥 张玉

(北京师范大学心理学院, 应用实验心理北京市重点实验室, 北京 100875)

摘要 作者简要回顾了 SEM 框架下分类数据因素分析(CCFA)模型和 MIRT 框架下测验项目和潜在能力的关系模型, 对两种框架下的主要参数估计方法进行了总结。通过模拟研究, 比较了 SEM 框架下 WLS_c 和 WLSMV 估计方法与 MIRT 框架下 MLR 和 MCMC 估计方法的差异。研究结果表明: (1) WLS_c 得到参数估计的偏差最大, 且存在参数收敛的问题; (2)随着样本量增大, 各种项目参数估计的精度均提高, WLSMV 方法与 MLR 方法得到的参数估计精度差异很小, 大多数情况下不比 MCMC 方法差; (3)除 WLS_c 方法外, 随着每个维度测验题目的增多参数估计的精度逐渐增高; (4)测验维度对区分度参数和难度参数的影响较大, 而测验维度对项目因素载荷和阈值的影响相对较小; (5)项目参数的估计精度受项目测量维度数的影响, 只测量一个维度的项目参数估计精度较高。另外文章还对两种方法在实际应用中应该注意的问题提供了一些建议。

关键词 多维项目反应理论; 验证性因素分析; 参数估计; 分类数据

分类号 B841

1 引言

在心理学和教育学的研究中, 测验所采用的问卷或量表往往具有多维度的结构, 且测验项目反应的数据本质上也常常不具备连续性和等距性的特点(如心理测验中的同意与反对、是与否, 教育测验中的正确与错误)。项目反应理论(Item Response Theory, IRT)和因素分析(Factor Analysis, FA)是常用的分析项目特征的方法, 然而经典的项目反应理论往往假设测验具有单维结构, 当测验含有多个维度时, 则需要用到较为复杂的估计方法。当测验项目不连续或不等距时, 测验题目测验结果的数据类别往往不满足一些统计模型对数据测量水平和分布特征的要求, 如传统的因素分析的方法要求数据满足: (1)观测变量之间的协方差可以表示为少数潜在变量的函数; (2)特殊因子服从正态分布; (3)结果变量服从条件正态分布; (4)观测变量与潜变量之间

存在线性关系。

因素分析方法是心理学中常用来分析测验工具结构和质量的方法, 然而大多数的研究却往往忽略了数据本身的特点(Gierl & Mulvenon, 1995)。如果忽略分析的前提条件, 在结构方程模型(Structural Equation Model, SEM)的框架下采用传统的极大似然(Maximum Likelihood, ML)估计方法对数据进行分析, 这样得到的参数估计及其对应的标准误就可能出现偏差(Bollen, 1989), 而不能正确解释观测变量与潜变量之间的关系(Distefano, 2002)。大量的研究表明, 如果测量指标变量的分类数目为五类, 且服从正态分布, 可以直接作为连续变量来处理, 得到模型参数估计的结果是比较稳定的(Muthén & Kaplan, 1992)。然而当测量指标变量分类数目较少时, 得到的参数估计和拟合指数等则会存在较大偏差(Green, Akey, Fleming, Hersherberger, & Marquis, 1997)。另外, 分类(二分)数据也很难保证正态性,

收稿日期: 2010-11-24

* 国家自然科学基金项目(31100759)资助。

通讯作者: 骆方, E-mail: luof@bnu.edu.cn

如果正态假设不能被满足,那么在 SEM 框架下基于 ML 估计的 χ^2 就不再服从中心 χ^2 分布。当数据是分类的,且非正态分布时,拟合指数会受到很大影响,基于 ML 估计的 χ^2 、RMR 会夸大(Green et al., 1997), NNFI, GFI 和 CFI 也会被低估(Hutchinson & Olmos, 1998)。这意味着实际上可能正确的模型却不能很好的拟合数据,这就可能会导致研究者放弃正确的模型。随着样本量的降低、分类数目的减少以及非正态程度的增加,模型分析结果的偏差会更为明显(DiStefano, 2002)。

随着统计技术的发展,因素分析技术从最初提出概念发展到现在期间发生了质的飞跃,不仅有探索性的因素分析,还有验证性的因素分析。而且对于不同的数据类型,其方法也得到了广泛的拓展。主要体现在估计方法的改进,目前常用的分类数据的估计方法有(1)渐进分布自由估计法(ADF),这可看作是一种没有数据分布要求的 ML 估计法,是由 Browne (1984)提出的;(2)S-B 修正,根据观测数据的多元峰度水平,修正基于 ML 得到的卡方值和标准误,包括 MLM (均值调整过的 ML)和 MLMV (均值和方差调整过的 ML),同样也可用来修正模型的拟合指数;(3)Robust WLS (RWLS)估计方法,包括 WLSM (均值调整过的 WLS)和 WLSMV (均值和标准误调整过的 WLS);(4)对每个参数的实际分布进行逐步估计,以便计算更准确的卡方值和标准误。前三种估计方法可以看作是在基于正态假设的估计方法上的修正,而最后的逐步估计法则是对参数估计过程的改进。

项目反应理论基于被试作答,将被试的潜在能力描述为关于项目的非线性函数,它对题目类型也没有连续性和正态性的要求,实质上是描述项目和潜在能力间的关系。另外,项目反应理论可以提供诸如测验信息函数、项目特征曲线等指标,用来描述测验和测验项目的质量。但是经典的项目反应理论由于受局部独立性和单维性等条件的限制,在描述测验工具质量和测验结构时较少用到。随着多维项目反应理论(Multidimensional Item Response Theory, MIRT)的发展和 MCMC (Markov chain Monte Carlo)算法的推广应用,用项目反应理论分析测验项目的特征和质量逐渐受到重视(Reckase, 2009)。随着 MIRT 分析软件的开发(TESTFACT, NOHARM, ConQuest, BMIRT 等),这类方法更加受到重视,一些大型测评项目就采用了 MIRT 的模型(如 PISA, TIMSS 等)。

由于 IRT 分析得到的项目参数和 FA 得到的项目参数具有一定的联系, Takane 和 de Leeuw (1987)阐述了分类数据因素分析和项目反应参数之间的关系。已有一些研究对基于 IRT 的参数估计方法和 FA 的参数估计方法之间的差异进行了比较。Tate (2003)比较了三参数模型中,在探索性分析的框架下斜交旋转后 TESTFACT 和 NOHARM 得到的项目参数与在验证性框架下采用 MPLUS 中 RWLS 估计方法得到的项目参数; Gosz 和 Walker (2002), Miller (1991)比较了不同情境下 TESTFACT 和 NOHARM 之间的差异; Knol 和 Berger (1991)比较了不同情境下 NOHARM、TESTFACT 和基于四分相关的因素分析方法与多维 MAXLOG (采用联合极大似然估计)方法的差异,研究发现,NOHARM、TESTFACT 和基于四分相关的因素分析方法能得到较准确的参数估计结果; Finch (2010)比较了不同情境下基于验证性的分析框架,NOHARM 的 ULS (Unweighted Least Squares)估计方法、MPLUS 的 RWLS 估计方法间的差异,但其假设的测验结构相对简单,每个题目只测一个维度,并且在其研究中采用了分析单维测验的方法处理测验多维度的问题。

本文目的是在 SEM 和 MIRT 的模型框架下,首先对多维度分类指标变量的分析模型、参数估计以及参数之间的关系进行简要说明。然后通过模拟研究,对不同条件下的参数估计精度及其影响因素进行分析和比较,旨在给实际应用者提供如何选取合适的分析模型和参数估计方法以及应该注意的相关问题等方面的建议。

2 分类指标变量的因素(潜变量)分析模型

基于分类数据的因素分析模型和项目反应模型为描述分类反应项目与潜变量之间的关系提供了合适的方法(Mislevy, 1986)。在 IRT 的理论框架下,有一系列描述项目与潜在特质之间关系的模型,包括等级反应模型(Samejima, 1969)、非等级反应模型(Bock, 1972)、分步计分模型(Masters, 1982; Muraki, 1992)、带有猜测参数的模型(Birnbaum, 1968)等。本文所考虑的模型主要是在多维条件下的分类数据验证性因素分析模型(Categorical Confirmatory Factor Analysis, CCFA)和两参数的 Logistic (2PL)二分等级反应模型。

2.1 分类数据的因素分析模型

2.1.1 模型定义 CCFA 假设分类的项目反应是潜

在的连续特征反应的表现, 即个体 i 在项目 j 上的反应是一连续的潜在变量 x_{ij}^* (x_{ij}^* 是连续且服从正态分布的潜变量), 观测的项目反应 x_{ij} 是潜在连续反应的分类型表现。因此对于某一项目, 项目分类反应变量的分布是该项目潜变量 x_{ij}^* 分布的表现。可以通过阈限参数 τ 定义 x_{ij}^* 和 x_{ij} 之间的关系。如果 $\tau_{jc} < x_{ij}^* < \tau_{jc+1}$, 则 $x_{ij}=c$ 。这里, τ_{jc} 用来定义 C 等级的反应, $\tau_{j0}=-\infty$, $\tau_{jC}=\infty$, $c=0,1,2,\dots,C-1$ 。对于二分类变量来说, $C=2$ 。 τ 参数表示连续潜在反应与观测的离散性反应的分界点。假设某一潜变量的分布为正态分布, 如果对于某一二分类观测变量(同意和不同意)做出“不同意”反应的概率是 0.1587, 阈值 τ 就等于 -1 (即概率 0.1587 对应的 Z 分数)。此时, τ 表示平均而言, 潜变量分数低于均值一个标准差的个体在这个项目上将会选择“不同意”, 而潜变量标准分数大于 -1 的个体, 将会做出“同意”的选择。对于两个二分反应的项目, 可以用表 1 的列联表表示观测分类变量与潜变量之间的关系。

表 1 两个二分反应项目, 阈值、潜变量和观测变量之间的关系

		题目 1	
		$x_1^* > \tau_1$	$x_1^* \leq \tau_1$
题目 2	$x_2^* > \tau_2$	(1, 1)	(0, 1)
	$x_2^* \leq \tau_2$	(1, 0)	(0, 0)

CCFA 模型将潜在反应变量之间的相关结构用模型表述为一系列参数的线性组合。无条件 CCFA 模型可以定义为: $\Sigma_{xx}^* = \Lambda_x^* \Phi_{\xi\xi}^* \Lambda_x^{*'} + \Theta_{\delta\delta}^*$ 其中, Σ_{xx}^* 表示 J 维潜在反应变量(即理论上的潜在连续反应变量)的总体协方差矩阵。 Λ_x^* 表示因素载荷矩阵, $\Phi_{\xi\xi}^*$ 表示正态分布的潜在因素之间的协方差矩阵。 $\Theta_{\delta\delta}^*$ 表示特殊因子间的协方差矩阵, 一般为对角矩阵。模型从形式上来看与传统的因素分析模型没有区别, 然而 CCFA 模型中由于方程两边均是潜变量, 因此在参数估计上比传统因素分析要复杂。

2.1.2 参数估计方法 有多种方法可以用来估计 CCFA 模型中的参数, 本文主要考虑两种常用的估计方法: 分类数据的加权最小二乘估计(Weighted Least Squares for Categorical data, WLSc)和修正的分类数据加权最小二乘估计(Modified Weighted Least Squares for Categorical data, MWLSc), 这两种方法被认为是目前用来估计 CCFA 模型参数较好的两种方法。

(1) 分类数据的加权最小二乘估计(WLSc)

WLSc 是一种最常用的估计 CCFA 参数的方法。这一方法直接由连续变量的加权二乘法推广而来。连续变量加权二乘法的拟合函数定义为:

$$F_{WLS} = (s - \sigma)' W^{-1} (s - \sigma)$$

这里, s 是由样本协方差矩阵中的 $p \times p$ 个元素组成的向量, σ 是由模型定义的协方差矩阵中的 $p \times p$ 个元素组成的向量, W^{-1} 是一权重矩阵, 是 $u \times u$ 阶正定矩阵的逆矩阵(其中 $u=p(p+1)/2$), 是渐进协方差矩阵的估计值(Browne, 1984)。这一估计方法即使是对连续的指标变量, 对有限的样本容量, 在估计上也存在一系列的问题。其最大的困难就在于很难得到权重矩阵的精确估计。这主要是由于权重矩阵有 $u(u+1)/2$ 个不同的元素, 随着指标变量个数的增加, 权重矩阵中需要估计的元素快速增加。只有当样本量超过 u 时, 才能使得权重矩阵可逆。一般为了保证得到权重矩阵的精确估计, 样本量 n 应该远远大于 u (Browne, 1974, 1984)。然而在实际中, 即使在样本量较大, 权重矩阵可逆的情况下, 权重矩阵估计的精度也经常受到质疑。

对于分类的数据, 上面的拟合函数可以重写为: $F_{WLSc} = (r - \rho)' W^{-1} (r - \rho)$ 。这里 ρ 是由模型定义的相关矩阵中 $p \times p$ 个元素组成的向量(Jöreskog, 1994), r 是由样本相关矩阵(二列相关或多系列相关)中 $p \times p$ 个元素组成的向量。也有研究者(Muthén, du Toit, & Spisic, 1997)将观测的和模型估计的阈限值分别包含在向量 r 和 ρ 中。不论阈限的值是否包含在拟合函数中, 估计正定的加权矩阵的逆矩阵仍是必须的。和连续变量的 WLS 估计方法一样, 分类变量的 WLSc 需要很大的样本量来得到权重矩阵的精确估计(Muthén et al., 1997)。在心理学的研究中, 当测验中题目比较多时, WLSc 估计在应用过程中不只有样本容量和估计精确度的问题, 而且对于上述的分类数据模型来说, 在理论上也无法得到近似协方差矩阵的唯一解。近年来, 有多种方法可用来估计分类数据或等级数据的权重矩阵(Jöreskog, 1990, 1994; Lee, Poon, & Bentler, 1990, 1995; Muthén, 1984; Muthén et al., 1997)。虽然不同的方法在理论上都是渐近收敛的, 但对有限的样本来说, 得到的结果可能不同(Oranje, 2003)。

(2) 修正的分类数据加权最小二乘估计(MWLSc)

为了克服估计权重矩阵中所有元素所带来的计算量大的局限性, 研究者提出了只估计主对角线上的元素, 而将其它元素固定为零的方法, 这里主

要介绍这种基于 WLS_c 的修正方法。MWLS_c 将拟合函数定义为: $F_{MWLS_c} = (r - \rho)'W_D^{-1}(r - \rho)$ 。这里 ρ 和 r 的定义与 WLS_c 相同, 不同的是加权矩阵 W_D^{-1} 中只包含对角线上的元素, 这一修正大大减少了权重矩阵中非零元素的个数, 因而提高了计算速度和降低了估计方法对样本量的要求。对角化 WLS_c (Jöreskog & Sörbom, 2001; Christofferson, 1975) 和稳健 WLS_c 估计 (Muthén et al., 1997) 就是采用这一策略的两个代表性的方法, 又称为均值校正的 WLS (WLSM) 和均值方差校正的 WLS 估计 (WLSMV)。然而这一简化方法带来的问题是采用 MWLS_c 得到的估计不再是一个充分的统计量, 参数估计的标准误和检验也是有偏的。为此, 研究者提出了一系列校正的方法 (Satorra & Bentler, 1994; Yuan & Bentler, 1998; Oranje, 2003), 然而这些方法虽然修正了卡方统计量和标准误, 却没有对模型的自由度做出修正。Muthén 等 (1997) 提出的均值方差校正不仅修正了卡方统计量、标准误, 而且对模型的自由度也作了校正。Flora 和 Curran (2004) 的研究发现 Muthén 等 (1997) 的方法可以提供很好的参数估计结果。本文主要考察 WLS_c 和 WLSMV 方法对参数估计的精度 (WLSM 和 WLSMV 得到的结果完全相同)。

2.2 多维项目反应理论模型

2.2.1 模型定义 项目反应理论框架下的项目因素分析模型主要是针对分类反应变量发展起来的模型。对于二分反应 (0-1 计分) 数据, 一般的两参数多维项目反应模型可以表示为 (Ackerman, 1994):

$$P_i(U_{ij} = 1 | \theta_j) = \frac{1}{1 + \exp(-\sum a_{im}(\theta_{jm} - b_i))}$$

这里 θ_{jm} 是第 j 个被试在第 m 个维度上的潜在能力; a_{im} 是第 i 个项目在第 m 个维度上的区分度参数, 描述了第 i 个项目与第 m 个潜在能力间关系的强弱; b_i 是第 i 个项目的难度参数。

2.2.2 参数估计方法 IRT 框架下的项目因素分析模型的参数估计依赖于原始数据, 这一估计方法考虑所有原始数据的信息, 因此又称为全息估计 (Full-information estimators)。最早的估计 IRT 参数的方法是联合极大似然估计 (Joint Maximum Likelihood, JML), 在联合极大似然估计中, 项目参数和能力参数同时估计, 这时有可能导致得到的估计量不能满足一致性的特点。在 IRT 框架下, 最常用的 IRT 参数估计方法是采用 EM (Expectation Maximization) 算法的边际极大似然估计 (Maximum

Marginal Likelihood, MML) (MML/EM)。MML/EM 方法不仅可以用于单维 IRT 模型的参数估计, 还可以推广到多维 IRT 的情境。Bock, Gibbons 和 Muraki (1988) 最早将这一估计方法应用于探索性的 MIRT 模型, 并称其为全息项目因素分析。Gibbons 和 Hedeker (1992) 将其进一步推广到具有一定限定条件的验证性 MIRT 模型, 即双因素 (Bifactor) 模型。然而, 标准的 MML/EM 方法仍是基于 MML 的原理采用 Gauss-Hermite 积分, 且其对因素个数有较大的限制, 在此基础上采用数值积分的修正 MML 估计是 TESTFACT 和 MPLUS 采用的多维 IRT 模型参数估计方法, 在 MPLUS 中, 称为 MLR (Maximum Likelihood estimation with Robust standard errors)。从估计方法的进展来看, 近年来 MCMC 估计法受到许多研究者的青睐。本文主要考察的 MIRT 参数估计方法有 MLR 和 MCMC。

从估计方法上来看, 基于 SEM 的估计方法和基于 IRT 的估计方法最主要的区别就在于 IRT 估计法是在潜在维度 (因素) 上求积分, 而 SEM 估计方法是基于潜在项目 x_{ij}^* 而求积分。

2.3 CCFA 和 IRT 模型参数之间的关系

关于项目与潜变量之间的关系, 虽然在 SEM 和 IRT 框架下有两种不同的模型描述, 但是二者在理论上却也有非常密切的关系。CCFA 模型的参数与两参数 IRT 等级反应模型的参数之间满足一定的数学关系 (Lord & Novick, 1968; Takane & de Leeuw, 1987)。在两参数 0-1 计分的 MIRT 模型中的区分度参数向量 a_i 与 CCFA 中的因素载荷向量 λ_i^* 之间的关系可以表示为 $a_i = (\frac{\lambda_i^*}{\sqrt{1 - \lambda_i^{*'} \phi \lambda_i^*}})D$, 这里 λ_i^* 为第

i 个项目所对应的因素载荷向量, ϕ 是潜在因素之间的协方差矩阵, 项目难度 b_i 与 CCFA 参数之间的关系可以表示为: $b_i = \frac{\tau_i}{\sqrt{1 - \lambda_i^{*'} \phi \lambda_i^*}}$ 。对于 Logistic

项目反应模型, $D=1.7$ 。对于正态拱形 IRT 模型, $D=1$ 。

3 研究设计

3.1 研究方法

在多维测验情境下, 为了比较 SEM 框架下 WLS_c 和 WLSMV 与 IRT 框架下 MLR 和 MCMC 估计方法得到的参数估计的精度, 探索影响参数估计精度的因素, 本研究采用 Monte Carlo 模拟研究的

方法, 根据已知的区分度和难度参数, 生成被试反应矩阵。

3.2 模拟设计

本研究考虑测验长度和测验维度对参数估计精度的影响, 共生成四种类型的数据集: (1)两个维度, 每个维度 10 个题目; (2)两个维度, 每个维度 20 个题目; (3)三个维度, 每个维度 10 个题目; (4)三个维度, 每个维度 20 个题目。项目和维度间的关系见表 2, 项目在所测量维度上具有相同的区分度, 区分度真值为 1; 对于 5 个项目, 难度分别为-2, -1, 0, 1, 2; 对于 10 个项目, 难度分别为-2, -1.5, -1, -0.5, 0, 0, 0.5, 1, 1.5, 2。多维的能力服从均值为 0, 协方差矩阵为 Σ 的正态分布, Σ 矩阵的对角线元素为 1, 表示每个维度能力的方差为 1, 非对角线元素为 0.5, 表示不同维度能力间的相关为 0.5。如果第 j 个被试在第 i 个题目上的反应的概率 $p_i(\theta_j) > u_{ij}$, 则 $x_{ij} = 1$, 反之 $x_{ij} = 0$, 其中 u_{ij} 为取自均匀分布[0, 1]的随机数。为了平衡不同估计方法所受样本量的影响, 本研究考虑样本量分别为 250, 500, 1000 和 2000 四种情况。每种条件重复 30 次。

表 2 四种不同条件下的模拟设计

数据集	项目个数	维度	模式	A1	A2	A3
1	15	2	(5x)	1	1	
			(5x)	1	0	
			(5x)	0	1	
2	30	2	(10x)	1	1	
			(10x)	1	0	
			(10x)	0	1	
3	20	3	(5x)	1	1	1
			(5x)	1	0	0
			(5x)	0	1	0
			(5x)	0	0	1
4	40	3	(10x)	1	1	1
			(10x)	1	0	0
			(10x)	0	1	0
			(10x)	0	0	1

3.3 分析软件

采用 WinGen2 (Han, 2007)软件, 模拟生成满足总体参数的被试反应数据。采用 MPLUS 6.0 (Muthén, L. K. & Muthén, B. O., 2008)软件, 用 WLS_c 和 WLS_{mv} 方法估计 SEM 框架下的项目参数; 用 MPLUS 6.0 中修正的数值积分的 MLR 方法和 BMIRT (Yao, 2008)软件中 MCMC 方法估计

MIRT 框架下的项目参数。

3.4 评价标准

对于模拟研究, 因为事先知道参数的真值, 采用偏度(BIAS)和均方根差(Root Mean Square Error, RMSE)作为检验参数估计精度的指标。BIAS 的定义如下: $BIAS = \frac{1}{n} \sum_{j=1}^n (\hat{\theta} - \theta)$; RMSE 的定义如下:

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (\hat{\theta} - \theta)^2}, \text{ 其中 } \theta \text{ 为参数真值, } \hat{\theta} \text{ 为}$$

参数估计值。BIAS 的值描述了参数估计的偏度, 越接近于零偏度越小, 正值表示该方法有高估参数的趋势, 负值表示该方法有低估参数的趋势。RMSE 的值描述了参数估计的精度, 其绝对值越小表示参数估计值与真值的差异越小。

所有的模型对 MCMC, MLR, WLS_c 和 WLS_{mv} 四种估计方法得到的结果进行比较。比较的参数包括项目水平的参数: 区分度、难度、阈值和因素载荷以及能力之间的相关。

4 主要结果

参数估计过程中, 对 WLS_c, WLS_{mv} 和 MLR 方法, 同时考虑迭代次数的限制(不超过 1000 次)和参数估计精度的限制条件(两次迭代估计值的差异小于 0.001)的标准。对于四类数据集, 除了 WLS_c 估计方法, WLS_{mv} 和 MLR 方法在所有情况下均可得到收敛的参数估计结果。对于 WLS_c 方法, 数据集 1 样本量为 250 时, 30 次重复中得到 23 次收敛的结果; 数据集 2 样本量为 250 时, 无法得到收敛的参数估计结果, 样本量为 500 时, 30 次重复中得到 29 次收敛的结果; 数据集 3 样本量为 250 时, 30 次重复中得到 15 次收敛的结果; 数据集 4 样本量为 250 和 500 时, 均无法得到收敛的参数估计结果。而对于 MCMC 方法, 估计过程中不同链条得到参数估计结果的稳定性描述了参数估计的收敛程度, 为了得到稳定的参数估计结果, 在每一种情况, 可采用 R 语言中检验 MCMC 估计结果稳定性的程序对参数估计结果的稳定性进行检验, 确定抽取的链条数和计算参数估计值所采用的链条数, 本研究中每次抽取 10000 次, 采用后 5000 次链条结果的均值作为估计参数值。

4.1 项目区分度估计结果

表 3 给出了不同估计方法得到的区分度参数的估计结果。

表 3 项目区分度估计偏度和均方根差

样本量	估计方法	2 维				3 维			
		15 题		30 题		20 题		40 题	
		BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
250	MCMC	0.135	0.389	0.159	0.380	0.152	0.417	0.170	0.423
	MLR	0.020	0.991	0.021	0.331	0.060	1.757	-0.007	0.377
	WLS _c	0.196	1.043	NC	NC	0.522	1.837	NC	NC
	WLSMV	0.006	0.509	0.112	0.297	0.008	1.554	-0.027	0.415
500	MCMC	0.067	0.280	0.069	0.244	0.080	0.361	0.072	0.245
	MLR	0.054	0.276	0.030	0.206	0.018	0.345	0.020	0.227
	WLS _c	0.158	0.413	1.331	2.723	0.206	0.725	NC	NC
	WLSMV	0.042	0.273	0.096	0.181	-0.011	0.335	0.001	0.229
1000	MCMC	0.055	0.223	0.041	0.161	0.036	0.240	0.043	0.193
	MLR	0.051	0.208	0.014	0.139	0.031	0.210	0.015	0.165
	WLS _c	0.086	0.236	0.265	0.280	0.100	0.284	0.018	1.012
	WLSMV	0.032	0.202	0.065	0.126	0.008	0.207	0.003	0.167
2000	MCMC	0.029	0.174	0.029	0.124	0.029	0.187	0.033	0.127
	MLR	-0.006	0.136	-0.015	0.112	0.031	0.157	0.025	0.112
	WLS _c	-0.001	0.141	0.074	0.141	0.049	0.179	0.339	0.342
	WLSMV	-0.025	0.137	0.005	0.103	0.006	0.156	0.019	0.162

注：NC 表示在此条件下，没有得到收敛的参数估计结果。(下同)

在样本量较小时, MCMC 倾向于高估项目的区分度, 而 WLSMV 和 MLR 方法估计的偏差相对较小。随着样本量的增加, MCMC、WLSMV 和 MLR 对区分度参数估计的偏差均很小, 参数估计的精度 (RMSE) 随着样本量的增加越来越高。但是当每个测验维度仅有 10 个题目、且样本容量很小 ($N=250$) 的情况下, MCMC 估计法得到的参数估计精度高于 MLR 和 WLSMV 方法, 此时, MLR 和 WLSMV 的误差很大, 无法得到稳定的参数估计结果, MCMC 似乎受样本容量的影响较其他方法小, 但是关于这一点需要进一步的研究来证实。在其它情况下, MLR 和 WLSMV 估计方法得到的参数估计精度 (RMSE) 的值都略高于 MCMC, 但是三种估计方法精度的差异很小。另外, 测验长度对参数估计精度有一定的影响, 除了 WLS_c 估计法, 其余三种估计方法结果显示, 测验越长参数估计精度越高。测验维度个数对项目区分度的估计精度也有的影响, 随着测验维度的增加, 参数估计的精度变低。

4.2 项目难度估计结果

表 4 给出了不同估计方法得到的难度参数估计结果。

对于难度参数, 四种估计方法中, WLS_c 仍然是估计精度最差的方法。在某些情况下, WLS_c 可能会略微高估难度参数 (如 3 维度样本容量为 1000

时)。MCMC 在所有情况下, 对难度参数估计的偏差都很小, WLSMV 和 MLR 对难度估计的偏差除了在大样本 3 维度情境下偏差稍微高一些外, 其他情况难度参数估计的偏差也很小。对于难度参数, 随着样本量的增加, 参数估计的精度越来越高。MCMC, MLR 和 WLSMV 得到的难度参数的精度基本没有差异。测验长度和维度对难度参数估计精度的影响与对区分度参数的影响相似, 除了 WLS_c 估计法, 其余三种估计方法结果显示, 测验长度越长, 参数估计的精度越高。三维测验难度参数估计的精度略低于二维测验。

4.3 项目因素载荷估计结果

表 5 给出了不同估计方法得到的因素载荷参数估计结果。

从表 5 的结果可以看出, WLS_c 估计方法仍是所有估计方法中估计偏差和精度最差的方法, 但是当样本量较大时, WLS_c 方法可以得到相对于区分度参数来讲偏差和精度都较好的参数估计结果, 且在大样本时, WLS_c 估计与其他方法的偏差相对较小。MCMC、WLSMV 和 MLR 估计方法对项目因素载荷的估计偏差都很小。随着样本量的增加, 每种方法得到的因素载荷的精度都越来越高。除了在每个测验维度 10 个题目 (5 个混合, 5 个独立) 的情况下, MCMC 估计法得到的因素载荷估

表 4 项目难度估计偏度和均方根差

样本量	估计方法	2 维				3 维			
		15 题		30 题		20 题		40 题	
		BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
250	MCMC	0.003	0.201	-0.024	0.186	0.022	0.230	-0.017	0.207
	MLR	-0.056	0.252	0.077	0.234	0.002	0.445	0.006	0.216
	WLS _c	-0.035	0.337	NC	NC	-0.180	0.898	NC	NC
	WLSMV	-0.062	0.220	0.073	0.243	0.033	0.374	0.035	0.250
500	MCMC	-0.006	0.178	0.001	0.142	0.012	0.154	0.011	0.169
	MLR	0.049	0.162	-0.013	0.137	0.117	0.190	0.112	0.170
	WLS _c	0.057	0.191	-0.022	0.662	0.000	0.285	NC	NC
	WLSMV	0.050	0.150	-0.015	0.130	0.002	0.152	0.119	0.177
1000	MCMC	-0.016	0.120	0.017	0.113	-0.021	0.134	0.017	0.112
	MLR	0.038	0.123	-0.018	0.096	-0.006	0.166	-0.009	0.140
	WLS _c	0.037	0.122	-0.021	0.132	0.123	0.203	0.182	0.497
	WLSMV	0.037	0.115	-0.025	0.095	0.122	0.190	-0.003	0.138
2000	MCMC	-0.010	0.094	-0.007	0.086	-0.009	0.097	-0.005	0.087
	MLR	-0.020	0.078	-0.030	0.083	-0.007	0.081	-0.001	0.069
	WLS _c	-0.023	0.078	-0.022	0.075	-0.005	0.082	0.003	0.129
	WLSMV	-0.022	0.076	-0.024	0.068	-0.006	0.076	-0.001	0.098

表 5 项目因素载荷估计偏度和均方根差

样本量	估计方法	2 维				3 维			
		15 题		30 题		20 题		40 题	
		BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
250	MCMC	0.021	0.106	0.033	0.107	0.013	0.094	0.024	0.103
	MLR	-0.011	0.209	-0.002	0.109	-0.014	0.254	-0.010	0.119
	WLS _c	0.034	0.398	NC	NC	0.060	0.478	NC	NC
	WLSMV	-0.013	0.168	-0.004	0.112	-0.018	0.406	-0.015	0.127
500	MCMC	0.009	0.098	0.022	0.081	0.013	0.100	0.015	0.076
	MLR	0.007	0.096	0.005	0.074	-0.003	0.108	0.001	0.072
	WLS _c	0.028	0.126	0.105	0.445	0.026	0.155	NC	NC
	WLSMV	0.003	0.096	0.003	0.074	-0.009	0.109	-0.003	0.074
1000	MCMC	0.008	0.079	0.009	0.057	0.009	0.069	0.009	0.058
	MLR	0.011	0.075	0.002	0.051	0.005	0.066	0.001	0.053
	WLS _c	0.020	0.081	0.049	0.089	0.017	0.080	0.054	0.140
	WLSMV	0.006	0.074	0.000	0.053	0.000	0.067	-0.001	0.055
2000	MCMC	0.005	0.062	0.006	0.043	0.002	0.058	0.006	0.040
	MLR	-0.003	0.051	0.006	0.042	0.005	0.049	0.013	0.035
	WLS _c	-0.001	0.053	0.015	0.048	0.009	0.055	0.059	0.085
	WLSMV	-0.008	0.052	-0.007	0.039	-0.000	0.051	0.006	0.054

计的精度显著高于 MLR 和 WLSMV 方法外, 其余大多数情况下, WLSMV 和 MLR 方法估计的精度略微高一些, 但是相差均很小。测验长度越长, 因素

载荷估计的精度越高。但是从参数估计的结果还可以看出, WLSMV 和 MLR 估计方法受测验长度的影响较 MCMC 方法大, 随着样本量的增加这一影响

逐渐减小。测验维度对项目因素载荷的影响较小。

4.4 项目阈值估计结果

从项目阈值参数的估计结果(表 6)可以看出,对于阈值参数,四种方法参数估计的偏差均很小。对于参数估计的精度,WLS_c 估计方法估计的精度略低。总体来看,MCMC、WLSMV 和 MLR 三种估计方法中,MCMC 估计法得到的阈值参数估计的精度略低于 MLR 和 WLSMV 方法,而 MLR 和 WLSMV 估计方法之间的差异相对不大。总体来看,

测验长度越长,阈值参数估计的精度越高。但是从参数估计的结果可以看出,测验长度对阈值参数的影响相对其他参数较小,且随着样本量的增加这一影响越来越不明显。

4.5 维度相关系数估计结果

维度间相关系数的估计结果(表 7)表明,在题目数较少样本量也较小时,MCMC 估计方法得到的相关系数估计值的偏度最小,其它方法倾向于高估维度之间的相关,WLS_c 方法几乎在所有条件下都

表 6 项目阈值估计偏度和均方根差

样本量	估计方法	2 维				3 维			
		15 题		30 题		20 题		40 题	
		BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
250	MCMC	0.005	0.091	-0.010	0.083	0.002	0.092	-0.008	0.083
	MLR	-0.021	0.081	0.014	0.054	-0.009	0.082	0.001	0.079
	WLS _c	-0.004	0.104	NC	NC	-0.066	0.175	NC	NC
	WLSMV	-0.021	0.079	0.034	0.087	-0.007	0.081	0.005	0.079
500	MCMC	-0.004	0.081	-0.001	0.070	0.002	0.077	0.002	0.066
	MLR	0.019	0.062	-0.002	0.037	-0.006	0.058	0.045	0.063
	WLS _c	0.019	0.073	-0.001	0.151	-0.006	0.075	NC	NC
	WLSMV	0.019	0.062	-0.008	0.055	-0.005	0.057	0.046	0.065
1000	MCMC	-0.007	0.055	0.008	0.051	-0.012	0.060	0.005	0.052
	MLR	0.017	0.045	-0.005	0.028	0.046	0.065	-0.006	0.054
	WLS _c	0.016	0.048	-0.003	0.062	0.043	0.065	0.045	0.106
	WLSMV	0.017	0.045	-0.009	0.040	0.046	0.065	-0.005	0.053
2000	MCMC	-0.002	0.042	-0.002	0.039	-0.003	0.040	-0.002	0.036
	MLR	-0.007	0.029	-0.007	0.029	-0.002	0.028	-0.000	0.027
	WLS _c	-0.007	0.029	-0.007	0.032	-0.002	0.030	0.002	0.062
	WLSMV	-0.007	0.029	-0.008	0.029	-0.002	0.028	-0.000	0.038

表 7 维度之间相关系数估计的偏度和均方根差

样本量	估计方法	2 维				3 维			
		15 题		30 题		20 题		40 题	
		BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
250	MCMC	-0.023	0.089	-0.025	0.074	0.005	0.163	-0.018	0.073
	MLR	0.150	0.206	-0.001	0.083	0.058	0.175	0.017	0.084
	WLS _c	0.155	0.241	NC	NC	0.109	0.174	NC	NC
	WLSMV	0.055	0.137	0.014	0.085	0.068	0.173	0.035	0.089
500	MCMC	-0.013	0.079	-0.018	0.052	-0.012	0.090	-0.016	0.055
	MLR	0.087	0.094	-0.001	0.054	0.006	0.087	-0.008	0.057
	WLS _c	0.080	0.132	0.246	0.271	0.071	0.124	NC	NC
	WLSMV	0.041	0.089	0.010	0.053	0.018	0.086	0.001	0.054
1000	MCMC	-0.018	0.070	-0.009	0.042	-0.022	0.063	-0.011	0.037
	MLR	0.074	0.084	0.011	0.043	0.001	0.061	0.005	0.034
	WLS _c	0.021	0.075	0.095	0.107	0.036	0.076	0.184	0.193
	WLSMV	0.015	0.073	0.016	0.043	0.008	0.061	0.012	0.036
2000	MCMC	0.005	0.046	0.017	0.032	0.020	0.049	0.045	0.029
	MLR	-0.014	0.046	0.008	0.028	0.002	0.042	-0.004	0.028
	WLS _c	-0.002	0.044	0.049	0.057	0.017	0.047	0.069	0.077
	WLSMV	-0.009	0.044	0.010	0.029	0.006	0.042	-0.001	0.028

高估了维度间的相关,是偏差和精度最差的方法。其它方法高估的程度随样本量的增大而减小,在大样本时 MCMC、MLR 和 WLSMV 间的差异很小。测验中题目越多,维度间相关系数估计的精度越高,测验维度个数对相关系数估计的精度影响不大。

4.6 题目测查维度数对参数估计精度的影响

在模拟设计中,有的项目同时测量两个或两个以上的维度,这样的项目被称为混合的维度内的项目,有的项目只测量一个维度,称之为独立的维度间的项目。为了考虑在多维测验情境下,测验题目考察的维度个数对参数估计结果的影响,首先将测

验题目分成混合型和独立型两种类型,分别计算两类项目对应的参数估计精度,四种方法所得到的趋势一致,为了叙述简练,将四种估计方法所得到的参数估计精度合并。表 8 给出了独立和混合情况下,区分度、难度、因素载荷和项目阈值参数估计的精度。结果表明,项目参数估计的精度受项目测量维度多少的影响,如果项目只测试一个维度,那么该项目参数估计的精度较高,如果项目同时测量几个维度,则项目参数估计的精度相对较差。同时,项目测量维度个数对项目区分度和因素载荷的影响较大,而对项目难度和阈值参数的影响相对较小。

表 8 项目测量维度数量对参数估计精度的影响

维度	项目数	样本量	区分度		难度		因素载荷		阈值	
			独立	混合	独立	混合	独立	混合	独立	混合
2	15	250	0.381	0.578	0.200	0.258	0.115	0.136	0.117	0.128
		500	0.272	0.403	0.183	0.257	0.103	0.120	0.083	0.104
		1000	0.205	0.311	0.115	0.157	0.069	0.105	0.056	0.065
		2000	0.125	0.197	0.068	0.100	0.049	0.061	0.030	0.032
	30	250	0.468	0.563	0.232	0.330	0.124	0.104	0.130	0.126
		500	0.326	0.495	0.161	0.240	0.097	0.114	0.089	0.084
		1000	0.188	0.316	0.122	0.194	0.064	0.086	0.059	0.062
		2000	0.109	0.164	0.071	0.099	0.037	0.068	0.031	0.033
3	20	250	0.421	0.534	0.208	0.237	0.121	0.133	0.114	0.111
		500	0.254	0.330	0.157	0.197	0.083	0.099	0.079	0.082
		1000	0.155	0.211	0.118	0.145	0.054	0.072	0.059	0.061
		2000	0.138	0.182	0.069	0.101	0.047	0.060	0.030	0.033
	40	250	0.434	0.600	0.223	0.273	0.123	0.123	0.116	0.106
		500	0.244	0.369	0.150	0.246	0.080	0.091	0.075	0.080
		1000	0.163	0.254	0.105	0.161	0.057	0.069	0.052	0.051
		2000	0.142	0.221	0.072	0.120	0.039	0.058	0.041	0.043

注:表中数据为四种估计方法合并后的结果,如某一条件下 WLS_c 没有得到收敛的结果,则表中数据为其余三种方法估计结果的合并。

5 讨论

5.1 估计方法的选择

本文基于 SEM 框架下的二分分类指标的因素分析模型与 IRT 框架下的 2-PL Logistic 模型,就几种不同估计方法得到的参数估计的精度进行了比较。对于二分分类数据因素分析,由于分类数据从假设上违背了传统因素分析对数据的要求,传统的用于连续变量因素分析的极大似然估计等方法就不再适用,因而在估计方法的选择上应该特别注意。在实际应用中,如果确信数据违背了传统因素分析的假设条件,那么在选择参数估计方法上就要

特别注意。根据以往的研究和本模拟研究的结果,分类数据的 WLS_c 估计由于在估计方法上需要计算完整的加权矩阵,因此当测验中题目较多时,有可能无法得到参数估计的结果,且较其它方法对样本量有更高的要求。同时, WLS_c 方法对参数估计的偏差较大,精度也相对较差,因此在实际应用中不建议采用。均值方差修正的 WLSMV 估计可以很好地估计出模型中的参数,因此建议在 SEM 框架下,对于分类数据因素分析采用修正的 WLSMV 方法或 WLSM 方法估计参数(在本研究的情境下 WLSMV 和 WLSM 方法没有差异)。

SEM 框架下的 WLSMV 和 IRT 框架下改进的

数值积分的 MLR 方法和 MCMC 估计方法得到了几乎相同的参数估计结果,从实证的角度说明了两种理论框架下参数估计方法的可比性。这一结果说明,如果同时采用 SEM 框架下和 MIRT 框架下的分析方法对工具的质量进行分析,可以从不同的角度对工具的质量进行评价,可为测验工具提供更加丰富的信息。IRT 下的项目特征曲线、项目信息量和测验信息量等概念可以了解测验的特点提供更多的信息;而 SEM 下的模型拟合指标可从模型选取和修正的角度提供更丰富的信息。本研究的另一个意义在于,SEM 框架下的因素分析可以很好地处理测验多维的情境,参数的等价性就为解决多维 IRT 的参数估计等问题提供了一个思路。从本文模拟研究的结果还可以看出,在大多数情况下,SEM 框架下 WLSMV 方法得到的项目参数估计的精度都与 IRT 框架下项目参数估计的精度接近,但在计算速度上,两种框架下估计方法的差异却很大,MCMC 方法和修正的基于数值积分的 MLR 方法计算时间很长,特别是在三维的情况下,其运行速度是 SEM 框架下 WLSMV 方法的 4-5 倍,样本量为 2000 时,甚至超过 10 倍。

MCMC 估计方法是目前用于 MIRT 的一种相对较新的方法,本研究的模拟结果表明,在大多数情况下,MCMC 估计法都可以得到较为理想的参数估计结果,尤其是在样本量很小时,其估计精度往往较其它方法更高,较少受样本量的影响。但是 MCMC 估计方法受先验分布的影响,在实际应用中如何定义合适的先验分布是研究者需要考虑的问题。另外估计过程中抽取链条的数量和每次估计采用的链条数量(或 Burn-in 的数量)都会对参数估计结果带来一定的影响。本研究采用的是抽取 10000 次,采用后 5000 次的结果估计参数。

对于实际应用者来说,究竟是选择在 SEM 框架下还是在 MIRT 框架下对多维分类数据进行分析,有许多需要考虑的问题。研究的目的往往决定了研究者所采用的分析方法。SEM 框架下的因素分析和 IRT 框架下的项目分析方法是在不同专业不同领域中广泛采用的两种关于测量模型的分析技术。两种方法在使用上的趋势可以粗略地概括为:如果一项研究主要关注观测变量和潜变量之间的关系,则 SEM 的方法较常用;但如果一项研究或应用者更关注个体能力的估计,那么 IRT 的方法则较常用。从结果解释的角度来看,SEM 的方法更关注项目因素载荷,实质上是项目的区分度,而 IRT 的方法似

乎更加关注项目的难度(有时也关注区分度),不过这些看似不同的参数实质上存在一定的联系。

对于分类数据的因素分析,除了参数估计方法的选择,还应该关注可能影响参数估计结果的一些因素。为了获得稳定可靠的参数估计结果,需要的样本量往往要比连续变量时更大(Wirth & Edwards, 2007)。虽然关于样本量并没有一个绝对的标准,但是从本文模拟研究的结果可以看出,如果被试能力满足正态分布,那么对于二分分类的指标变量的因素分析,随着模型的复杂程度(维度的个数)的增大,要达到同样的精度所需的样本量也需要增加。

另外,每个维度项目的个数也是影响参数估计精度的因素。在实际测验中,题目越多,对项目参数估计的精度有提高的趋势。同时,参数估计的精度还受题目测量维度个数的影响。单独的只测量一个维度的项目参数估计精度较高,而同时测量几个维度的混合型的项目参数估计精度相对较低,尤其是对于项目区分度和因素载荷的估计。这表明在测验编制过程中尽量做到一个题目只测量一个潜在特质。

5.2 研究的局限性

本研究只考察了在高维情境下,0-1 计分的模型不同参数估计方法之间的比较和影响参数估计精度的因素,对于多级计分的模型有必要进一步探讨;同时对于三参数的模型,几种方法之间的比较也是本研究没有涉及到的问题。另外,本研究采用模拟研究的方法,在设计上对项目区分度(因素载荷)、难度(阈值)和测验维度之间的相关都做了限定,在实际应用中,测验项目之间、测验项目与维度之间以及测验维度之间的关系会更加复杂,本研究得到的参数估计的精度在实际应用中有可能会有变动,有必要在后续研究中探讨更为复杂的设计。同时,本研究的结论能否推广到一般的探索性多维分析模型的情境或其他限定条件的验证性模型的情境,也是值得进一步探讨的问题。

6 研究结论

通过本研究可以得到以下主要结论:

(1)在多维模型下,SEM 框架下的 WLS_c 和 WLSMV 估计方法和 MIRT 框架下 MLR 和 MCMC 四种估计方法中,WLS_c 估计方法的误差最大,且存在参数估计收敛上的问题,当题目数较多、样本量较少时,可能无法得到收敛的参数估计结果。

(2)WLSMV、MLR 和 MCMC 估计方法相比,在

大多数情况下, MCMC 估计方法的偏差略大于 MLR 和 WLSMV 估计法。但是当样本量较大时, 这三种估计方法均可以得到较准确的参数估计结果, 且随着样本量的增大, 参数估计精度越来越高, 样本量超过 1000 时, 三种估计方法的差异很小。

(3) 无论是两个维度还是三个维度, 随着每个维度测验题目的增多, 除了 WLS_c 方法, 其余三种方法参数估计的精度逐渐增高。

(4) 测验维度对区分度参数和难度参数的影响较大, 维度增加参数估计的精度随之下降, 而测验维度对项目因素载荷和阈值的影响则相对较小。

(5) 所有参数估计方法对项目参数的估计精度受项目测量维度个数的影响, 只测量一个维度的项目参数估计的精度较高, 尤其是对于区分度和因素载荷。

参 考 文 献

- Ackerman, T. A. (1994). Using Multidimensional Item Response Theory to Understand What Items and Tests Are Measuring. *Applied Measurement in Education*, 7(4): 255-278.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (Chapters 17-20). Allan Oxford, England: Addison-Wesley.
- Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, 12, 261-280.
- Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.
- Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal*, 8, 1-24.
- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37, 62-83.
- Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40(1), 5-32.
- DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, 9, 327-346.
- Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis based models. *Applied Psychological Measurement*, 34(1), 10-26.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466-491.
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57, 423-436.
- Gierl, M. J., & Mulvenon, S. (1995). *Evaluating the application of fit indices to structural equation models in education research: A review of the literature from 1990 through 1994*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, April.
- Gosz, J. K., & Walker, C. M. (2002). *An empirical comparison of multidimensional item response data using TESTFACT and NORHARM*. Paper presented at the annual meeting of the National Council for Measurement in Education, New Orleans, LA.
- Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). The effect of number of scale points on chi square fit indices in confirmatory factor analysis. *Structural Equation Modelling*, 4, 108-120.
- Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement*, 31(5), 457-459.
- Hutchinson, S. R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Structural Equation Modeling*, 5, 344-364.
- Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, 24, 387-404.
- Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, 59, 381-389.
- Jöreskog, K. G., & Sörbom, D. (2001). *LISREL user's guide*. Chicago: Scientific Software International.
- Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, 26(3), 457-477.
- Lee, S. Y., Poon, W. Y., & Bentler, P. M. (1990). Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics and Probability Letters*, 9, 91-97.
- Lee, S. Y., Poon, W. Y., & Bentler, P. M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, 48, 339-358.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2): 149-174.
- Miller, T. R. (1991). Empirical estimation of standard errors of compensatory MIRT model parameters obtained from the NORHARM estimation program. ACT Research Report Series.
- Mislevy, R. J. (1986). Recent Developments in the Factor Analysis of Categorical Variables. *Journal of Educational and Behavioral Statistics*, (35)3: 280-306.
- Muraki, E. A. (1992). generalized partial credit model: Application of an EM algorithm. *Applied psychological measurement*, 16(2): 159.
- Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115-132.
- Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.
- Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal*

- of *Mathematical and Statistical Psychology*, 45, 19–30.
- Muthén, L. K., & Muthén, B. O. (2008). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.
- Oranje, A. (2003). *Comparison of estimation methods in factor analysis with categorical variables: Applications to NAEP data*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer.
- Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. New York: Psychometric Society.
- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications to developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.
- Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discredited variables. *Psychometrika*, 52, 393–408.
- Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, 27, 159–203.
- Wirth, R. J., & Edwards, M. C., (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12(1), 58–79.
- Yao, L. H. (2008). BMIRT: Bayesian Multivariate Item Response Theory. Software. *Presented at the annual meeting of National Council on Measurement in Education*, New York.
- Yuan, K. H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology*, 51, 289–309.

Item Parameter Estimation for Multidimensional Measurement: Comparisons of SEM and MIRT Based Methods

LIU Hong-Yun; LUO Fang; WANG Yue; ZHANG Yu

(School of psychology, Beijing Normal University; Beijing Key Lab of Applied Experimental Psychology, Beijing 100875, China)

Abstract

Traditional factor analysis models and estimation methods for continuous (i.e., interval or ratio scale) data are not appropriate for item-level data that are categorical in nature. The authors provide a brief review and synthesis of the item factor analysis estimation literature for categorical data (e.g., 0-1 type response scales) under the multidimensional response model. Popular categorical item factor analysis models and estimation methods found in the structural equation modeling and item response theory literatures are presented.

The Monte Carlo simulation studies are conducted and revealed: (1) Similar parameter estimates have been obtained of Modified weighted least squares for categorical data method (WLSMV) from the structural equation model (SEM) framework and adoptive Restricted Maximum Likelihood (MLR) and Markov chain Monte Carlo (MCMC) methods from the multidimensional item response theory (MIRT) framework. Even with a small sample and the item response theory (IRT) estimates converted to SEM parameters, the WLSMV, MLR, and MCMC results are strikingly similar. But in small sample size and long test, weighted least squares for categorical data (WLSc) did not obtain the convergence parameter estimations, although in short test, WLSc estimates have been obtained, the estimates are consistently more discrepant than those produced by the other estimation techniques. (2) The precision of the estimators enhances as the quantity of the sample increases, and the differences between WLSMV and MLR are very trivial, and the precisions of WLSMV and MLR methods are not worse than that of the MCMC method in most conditions. (3) The precision of item factor loading and of item difficulty parameter is influenced by the test length, and the precision of item discrimination and of item difficulty parameter is influenced by the number of test dimension. (4) The precision of the estimators decreases as the number of dimensions measured by the item increases, especially for item discrimination and item factor loading parameter.

Both SEM and IRT can be used for factor analysis of dichotomous item responses. In this case, the measurement models of both approaches are formally equivalent. They were refined within and across different disciplines, and make complementary contributions to central measurement problems encountered in almost all empirical social science research fields. The authors conclude with considerations for categorical item factor analysis and give some advice for applied researchers.

Key words Multidimensional Item Response Theory; confirmatory factor analysis; parameter estimation; categorical data