心理学报 2012, Vol. 44, No.2, 263−275
Acta Psychologica Sinica DOI: 10.3724/SP.J.1041.2012.00263

贝叶斯题组随机效应模型的必要性及影响因素

刘 玥 刘红云

(北京师范大学心理学院, 北京 100875)

摘 要 题组模型可以解决传统 IRT 模型由于题目间局部独立性假设违背时所导致的参数估计偏差。为探讨题组随机效应模型的适用范围, 采用 Monte Carlo 模拟研究, 分别使用 2-PL 贝叶斯题组随机效应模型(BTRM)和 2-PL 贝叶斯模型(BM)对数据进行拟合, 考虑了题组效应、题组长度、题目数量和局部独立题目比例的影响。结果显示: (1) BTRM 不受题组效应和题组长度影响, BM 对参数估计的误差随题组效应和题组长度增加而增加。(2) BTRM 具有一定的普遍性, 且当题组效应大, 题组长, 题目数量大时使用该模型能减少估计误差, 但是当题目数量较小时, 两个模型得到的能力估计误差都较大。(3)当局部独立题目的比例较大时, 两种模型得到的参数估计差异不大。

关键词 题组; 2-PL 贝叶斯题组随机效应模型; 2-PL 贝叶斯模型; MCMC 算法

分类号 B841

1 引言

项目间的局部独立性是项目反应理论的一个重要前提。然而在许多情况下, 这一假设可能会被质疑。例如, 考生可能在接受考试的过程中根据不同题目之间的关系进行学习; 测验可能要求考生解决一些内部相关联的题目(如同一阅读材料后面不同的题目)等。这种受共同因素影响和制约的一组题目, 就叫题组(testlet)。当测验中存在题组时, 题组内项目间存在局部依赖性, 两个项目反应的联合概率不再是特质水平下两个项目独立反应概率的乘积。因此, 没有考虑局部依赖的似然方程是错误的。如果仍然使用惯常的IRT模型进行估计, 会产生有偏的项目参数估计, 并且可能高估测验的信度(Wang & Wilson, 2005a)。

为了正确处理题组的影响, 有许多研究从模型建构的角度处理局部依赖性的项目, 建立了包含题组效应的模型, 这些研究主要从统计模型和测量模型方面对题组数据的建模进行了探讨。

统计模型主要是基于结构方程模型和因素分析的原理, 包括: 误差相关的单因素模型(Lee, Dunbar, & Frisbie, 2001), 多因素模型(包括两因子模型) (Lee et al., 2001)和二阶模型(Rijmen, 2009)。Lee 等人(2001)的研究中对误差相关模型, 多因素同质模型与传统 IRT 模型进行了比较, 结果显示当测验含有题组时, 误差相关模型和多因素模型要优于传统的误差无关的单因素测量模型。从测量的角度建立题组模型, 一般是通过对传统的 IRT 模型进行变形, 抽取出可能遗失的关于题组效应的信息。在测量模型中, 根据假设不同, 建立了两种不同的模型: 固定效应模型和随机效应模型。固定效应的方法是指对题组的分数模式建立模型, 利用分步计分 IRT 模型的原理, 将整个题组看做一个“超级项目”得以实现的(Chen & Wang, 2007)。题组的随机效应模型主要分为: 题组随机效应模型(Wainer & Wang, 2000; Wang & Wilson, 2005b; Bradlow, Wainer, & Wang, 1999), general 模型(Wang, Bradlow, & Wainer, 2002)和两因子模型(DeMars, 2006)。从模型比较的角度来看, 传统的 IRT 模型, 题组随机效应模型和两因子模型具有相互嵌套的关系。有研究者对它们的适用情况进行了对比研究(DeMars, 2006), 结果显示, 两因子模型更适用于依赖性较强的题组, 题组随机效应模型相比两因子模型具有更好的普遍性。

题组随机效应模型(Testlet Random-effects Model, TRM)最先由 Bradlow 等(1999)提出, 并建立

收稿日期: 2011−01−25

通讯作者: 刘红云, E-mail: hyliu@bnu.edu.cn

在贝叶斯估计的框架下，称之为“贝叶斯题组随机效应模型(Bayesian Testlet Random-effects Model, BTRM)”。由于该模型相比题组的统计模型更加灵活，且与传统 IRT 模型的参数更具有可比性，因此，在研究和实践中受到了越来越多的关注。以往关于 TRM 的研究，主要考察了题组效应大小(Bradlow et al., 1999; Wang et al., 2002; Wang & Wilson, 2005b)，题组长度(Bradlow et al., 1999)对 TRM 和传统 IRT 模型的影响。但是，这些研究对于影响 TRM 的因子设计并不充分，很少有研究同时涵盖了题组效应和题组长度两个方面(Bradlow et al., 1999; Wang et al., 2002)，即使同时包含了这两个因素，但有的研究对题组长度设置较简单，不能很好考察其变化的情况(Bradlow et al., 1999)；有的将这两个因素转换为拉丁方设计(Wang et al., 2002)，不能考察两个因素对题组随机效应模型和传统IRT模型影响的交互作用。另外，题目数量也是影响参数估计准确性的重要因素，很可能也是影响题组随机效应模型对数据拟合程度的一个方面，而大多数模拟研究中，题目数量都是固定的(Bradlow et al., 1999; Wang et al., 2002; Wang & Wilson, 2005b)。

综上，为了充分考察在题组效应，题组长度和题目数量这些因素变化的条件下，TRM 和传统 IRT 模型对数据拟合的情况，从而找到 TRM 的适用范围，有必要基于一个较为全面的因子设计，对两种模型进行模拟研究。研究一完善了对 TRM 模拟研究的因子设计，增加了必要的变化水平，期望能够对题组模型有更加深入和细致的认识。同时也能为实际使用者提供应用该模型的一些建议。

另外，在含有题组的试卷中同时包含局部独立的题目，不仅是实际测验题型的需要，也能保证题组模型的收敛。但是，在所有题组模型的模拟研究中，局部独立的题目占全卷的比例都是固定的(Bradlow et al., 1999; Wang et al., 2002; Wang & Wilson, 2005b)，一般为 1/2。因此，有必要设计一个模拟研究二，探讨局部独立题目占全卷比例不同的情况下，两种模型对数据的拟合情况。从而为题组模型的适用条件提出更多丰富的结论，也为实际中测验编制和模型选择提供一定的参考。

2 贝叶斯题组随机效应模型(BTRM)及其参数估计

2.1 模型

假设有 I 名被试，参加一个由 J 道题目组成的测验，所有题目均为 0, 1 计分。这些测验题目中包含 K 个题组($1 \leqslant K < J$)。用 $d(j)$表示题目 j 的题组，用 n_k 表示题组的大小($1 \leqslant k \leqslant K$)，并且 $d(1)=1$，$d(j)=K$。如果一个题组中 n_k=1，那么它就是一个局部独立的题目，类似的，如果一个题组中 K=J，那么所有题目都各自形成一个题组，它们也是局部独立的。通过 I 名被试对这 J 道题目的作答反应，就可以形成一个有 I*J 个元素的作答反应矩阵 Y=(y_{ij})。

BTRM 以两参数 logistic 模型为基础(Lord, Novick, & Birnbaum, 1968)，加入表示题组效应的题组效应参数，假设 t_{ij} 表示潜在能力特质，则有(Bradlow et al., 1999)：

$$\log it^{-1}(t_{ij}) = \exp(t_{ij}) / (1+\exp(t_{ij}))$$

$$t_{ij} = a_j(\theta_i - b_j - \gamma_{id(j)}) + \varepsilon_{ij}$$

$$\text{并且} \quad y = \begin{cases} 1 & \text{如果} t_{ij} > 0 \\ 0 & \text{其他} \end{cases} \tag{1}$$

其中，ε_{ij}服从独立正态分布，表示被试 i 在重复的作答项目 j 时，其作答正确与否的随机部分。参数 a_j，b_j，θ_j 和传统 IRT 模型中的含义一样，分别表示题目的区分度，难度和被试能力参数。$\gamma_{id(j)}$表示两个项目 j 和 j’在同一题组中，即 $d(j)$=$d(j')$时的局部依赖效应。

由于该模型相比一般的 IRT 模型，增加了题组效应参数$\gamma_{id(j)}$，为题目参数和被试参数的估计带来了困难。为保证参数估计的收敛，并得到较为精确的结果，一般采取贝叶斯估计的方法利用参数的先验信息进行估计。

2.2 参数估计

下面，以本研究使用的 SCORIGHT 软件所采用的参数估计方法为例，介绍对题组模型进行贝叶斯估计的方法。根据两参数 BTRM 的定义，令 $\Lambda_1 = \left\{\vec{\theta}_i, \vec{h}_j, \vec{b}_j, \vec{\gamma}_{id(j)}\right\}$。在贝叶斯框架下，从被试，题目和已有研究或者实证数据中借用先验信息，对 Λ_1 设定以下先验分布(Wang, Bradlow, & Wainer, 2004)：

$$\begin{pmatrix} h_j \\ b_j \end{pmatrix} \sim N_2\left(\begin{pmatrix} \beta_h^{(2)} \\ \beta_b^{(2)} \end{pmatrix}, \begin{pmatrix} (\sigma_h^{(2)})^2 & \rho_{hb}^{(2)}\sigma_h^{(2)}\sigma_b^{(2)} \\ \rho_{hb}^{(2)}\sigma_h^{(2)}\sigma_b^{(2)} & (\sigma_b^{(2)})^2 \end{pmatrix}\right) = N(\mu_{2PL}, \Sigma_{2PL}) \tag{2}$$

其中 h_j=log(a_j)这些参数假定服从正态分布。为了模型识别，将能力参数分布的均值和方差固定为 0 和 1，同时题组效应的分布也被固定。另外，特别强调，题组效应的方差 $\sigma^2_{d(j)}$ 是每个题组特定的，允许

在不同的题组之间随着局部依赖性的强度而变化。令表示先验分布的参数为 $\Lambda_2=\left\{\beta_h^{(2)},\beta_b^{(2)},\Sigma_{2PL},\sigma_{d(j)}^2\right\}$。这些参数是从 Λ_1 的共轭先验分布中选出来的, 加入它们是为了反映这些分布不确定的随机部分。系数的分布为:

$$\begin{aligned}\beta_h^{(2)} &\sim MVN(0,V_a)\\ \beta_b^{(2)} &\sim MVN(0,V_b)\end{aligned} \tag{3}$$

其中, $|V_a|^{-1}=|V_b|^{-1}=0$。同时, 对 Σ_{2PL} 使用一些超先验分布的信息。即:

$$\Sigma_{2PL}\sim Inv-Wishart(2,M_2^{-1})\text{ 和 }M_2=\begin{pmatrix}\frac{1}{100} & 0\\ 0 & \frac{1}{100}\end{pmatrix} \tag{4}$$

每个题组 $\sigma_{d(j)}^2$ 的分布为: $\sigma_{d(j)}^2\sim Inv-\chi_{g_\tau}^2$。且有: $g_\tau=\frac{1}{2}$。由于固定能力分布的方差限定了 IRT 模型的量尺, 所以题组的方差与前人文献中得出的推论的条件是能够进行比较的(Bradlow et al., 1999)。

为了得到未知参数的估计值, 需要通过一系列复杂的计算得到参数的后验分布, 进而抽取特征值作为其估计值。马尔可夫链蒙特卡尔(Markov chain Monte Carlo)方法是一种越来越受到广泛应用的贝叶斯计算方法。

使用这种 MCMC 方法, 其优势在于一旦得到了后验样本, 就能方便的从中做出推论, 因此不同的使用者可以根据自己研究的需要选择所关心的推论。另外, 由于近年来计算机的发展, 使得这种方法的计算越来越可行和普及。进而出现了很多人性化的贝叶斯估计软件, 例如 WinBUS, 这使得 MCMC 方法得到了广泛的应用。

本研究就是基于贝叶斯估计的 MCMC 方法, 通过 SCORIGHT 3.0 软件(Wang et al., 2004)分别估计两种模型参数。

3　研究一: 题组模型必要性的模拟研究

3.1　模拟设计

本研究模拟产生了 2000 名被试的作答反应数据。自变量有 3 个: (a)题组效应: 在 BTRM 中, 题组参数的方差 σ_γ^2 取四个水平, 0, 0.5, 1 和 2, 分别表示题组效应不存在、小、中和大四种情况。(b)题组长度: 指同一个题组内题目的数量, 分别取 2, 5 和 10 三种情况; (c)题目数量: 测验中所有的题目个数(包括局部依赖的题目和局部独立的题目), 分别取 20,40 和 60 三种情况。4×3×3=36 种条件下, 分别使用 BTRM 和不含题组的两参数贝叶斯模型(Bayesian Model, BM)[1] 进行模型拟合。考虑到模型使用的广泛性, 本研究只考虑两参数 BM (简称 BM)和在此基础上推广的 BTRM。

3.2　数据生成

使用 R 语言自编程序产生每种条件下的被试反应数据。每种条件下包括 2000 名被试对题目的反应, 所有的题目均为 0,1 计分的题目。为了模型的识别, 所有题目中一半为局部独立的题目, 另一半为包含题组的题目。例如, 在题组效应最大, 题组长度为 10, 题目数量为 60 的条件下, 这 60 道题目中有 30 道是局部独立的题目, 另外 30 道题目中, 10 道题目为一个题组, 共有 3 个题组。每种设计条件下数据重复模拟 30 次。

在各条件下, 模拟数据的参数分布参照前人文献及实证研究经验。其中 $\theta\sim N(0,1^2)$, $a_j\sim N(0.8,0.2^2)$, $b\sim N(0,1^2)$, 该分布与 ETS 进行 SAT 测验的观察分数的边缘分布匹配, 故可以与实际数据相比较(Bradlow et al., 1999)。注意由于模型识别的需要, 能力参数服从分布 $\theta\sim N(0,1^2)$, 因此, 所有题组效应大小的参数 σ_γ^2 都是相对于能力参数的方差 1 而言的。对于含有题组效应的模型, 其数据产生参照公式(1), 根据不同的条件, 对 γ 分布的方差进行设置。对于满足局部独立性的题目, 其数据产生参照两参数项目反应模型。

3.3　MCMC 估计

本研究应用 SCORIGHT 软件(Wang et al., 2004), 使用 MCMC 方法估计模型参数。MCMC 链数设置为 5, 后验抽取的间隔数为 10。各参数估计结果为每条 MCMC 链 4000 次迭代中后 1000 次的后验均值, 前 3000 次作为最初的值被舍弃。

3.4　评价标准

本研究从六个方面评价参数估计的精度以及模型与数据的拟合程度: (1)偏差(*Bias*), 绝对偏差(*MAE*), 误差均方根(*RMSE*), 估计值和真值的相关系数, 95%置信区间对真值的覆盖比例(*CP*), 95%置信区间长度(*MPIW*)。

*本文所考虑的随机效应题组模型采用贝叶斯方法估计参数，因此又称为贝叶斯随机效应题组模型，简称为 BTRM；为了便于参数估计结果的比较，传统 IRT 模型也采用贝叶斯估计，因此本文称传统的不含题组的 IRT 模型为贝叶斯模型，简称 BM。

偏差的意义是总体考察各条件下, 两种模型各参数的估计有没有定向的偏差。其计算公式如下:

$$bias = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{R}\sum_{r=1}^{R}(\hat{\varsigma}_r - \varsigma_r) \quad (5)$$

其中, $\hat{\varsigma}_r$ 表示参数的估计值, ς_r 表示参数的真实值, r 表示每次重复的题目数量, n 表示重复的次数。

绝对偏差考察各条件下, 两种模型各参数的估计值与真实值的差异绝对值大小。该值一般与误差均方根具有很高的一致性。因此在后文的结果中, 一般都参照误差均方根加以比较。其计算公式如下:

$$MAE = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{R}\sum_{r=1}^{R}\left|\hat{\varsigma}_r - \varsigma_r\right| \quad (6)$$

公式中各参数表示的意义与偏差的公式相同。

误差均方根是比较模型参考的最主要的标准, 它的大小反应了模型的参数估计值还原到真实值的能力。误差均方根越小, 说明模型参数估计结果与模拟的真实值越接近。前人关于模型比较的研究大多参考这一评价标准。其定义如下:

$$RMSE = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\varsigma}_r - \varsigma_r)^2} \quad (7)$$

公式中各参数表示的意义与偏差的公式相同。

估计值和真值的相关系数是计算各条件下, 两种模型各参数估计值与真实值的积差相关。它的意义是考察参数估计结果与真实值的一致性水平。

各参数 95%置信区间对真值的覆盖比例主要考察各条件下, 两种模型各参数估计值的 95%区间估计范围是否能包括真实值。该指标定义如下:

$$95\%CP = \frac{1}{n}\sum_{n=1}^{N}\frac{1}{r}\sum_{r=1}^{R}\delta$$

$$\delta = \begin{cases} 1 & (\text{真实值} \in \hat{F}_{0.025}, \hat{F}_{0.975}) \\ 0 & (\text{真实值} \notin \hat{F}_{0.025}, \hat{F}_{0.975}) \end{cases} \quad (8)$$

其中, $\hat{F}_{0.025}$ 表示经验累积分布函数中的 0.025 百分位数, $\hat{F}_{0.975}$ 表示经验累积分布函数中的 0.975 百分位数。如果真实值落在了估计值上下 1.96 个标准误的区间范围内, 则认为该参数的 95%置信区间覆盖了真实值。

各参数 95%置信区间长度(MPIW)是结合各参数 95%置信区间对真值的覆盖比例参考的一个指标。它的定义如下:

$$MPIW = \frac{1}{n}\sum_{n=1}^{N}\frac{1}{r}\sum_{r=1}^{R}(\hat{F}_{0.975} - \hat{F}_{0.025}) \quad (9)$$

公式中各参数表示的意义与各参数 95%置信区间对真值的覆盖比例的公式相同。

3.5 研究结果

根据软件提供的收敛指标, 研究中所有设计因子的条件下, 两种模型均成功收敛。

3.5.1 估计的偏差、绝对偏差和误差均方根 从参数估计值的偏差结果可以看出, 在各种条件下, 两种模型的区分度参数, 难度参数, 能力参数估计偏差均较小, 且在 0 上下有较小幅度的波动, 说明几种条件下两种模型对这些参数的估计基本上是无偏的。在 BTRM 中, 题组效应参数方差 σ_γ^2 在各种条件下估计值的偏差均大于 0, 说明该模型可能被高估了题组效应。

各条件下随着题组效应大小变化, 各参数估计的误差均方根见图 1~图 4。

(1) 区分度参数

在题组效应存在时, BTRM 得到的区分度参数估计精度均明显高于 BM, 且 BTRM 对区分度参数估计的误差基本不受题组效应, 题组长度和题目数量的影响, 在各条件下估计结果相对稳定。但 BM 的估计结果会受到这三个因素的影响。随着题组效应的增加, BM 对区分度参数估计的误差有较大增加。并且两种模型对区分度参数估计准确性的差异随着题组效应的增加而增大。

题组长度方面, 当题目数量较多时(60 道), 随着题组长度增加, BM 对区分度参数估计的误差减小; 当题目数量较少时(20 道), 随着题组长度增加, BM 估计的误差增加。因此, 当测验较长时, 随着题组长度的增加, 两种模型的对区分度参数估计的准确性的差异减小; 当测验较短时, 随着题组长度的增加, 两种模型的对区分度参数估计准确性的差异增大。

随着题目数量的减小, BM 对区分度参数估计的误差增大。特别注意, 当题目数量较少(20 道), 且存在较大的题组效应, 如仍使用传统 BM 进行分析, 将导致对区分度参数估计的误差远远大于题组模型, 且随着题组长度和题组效应的增加, 传统 BM 得到的区分度参数误差明显增大, 说明当测验中题目较少、题组长度较长、题组效应较大时, 传统 BM 无法得到可靠的区分度参数估计结果。

(2) 难度参数

BTRM 对难度参数估计的误差不受各因素影响, 在各条件下没有明显差异。但对于 BM, 分别随着题组效应和题组长度的增加, 难度参数估计的误差有较大增加。因此, 两种模型对难度参数估计的准确性的差异分别随着题组效应和题组长度的

增加而增大。另外同时考虑两个因素可以发现, 题组效应越大, 题组长度越长, 两种模型对难度参数估计的误差差异也越大, 使用 BTRM 对难度参数进行估计能得到更加准确的结果。

题目数量越少, BM 对难度参数估计的误差越大。尤其在题目少的情况下(20 道), 如果存在较大的题组效应, 且题组长度较长, 使用传统 BM 得到的难度参数估计值, 其误差明显大于 BTRM。这是与区分度参数的结果相一致的。

(3)题组效应参数

题组效应参数的方差主要受题组效应和题组长度的影响, 题目数量对该参数的估计没有影响。具体来说, 随着题组效应的增加, 题组效应参数方差估计的误差增大; 随着题组长度的增加, 题组效应参数方差估计的误差减小, 特别的, 当题组长度达到 5 以上时, 其估计的准确性明显优于题组长度为 2 的情况。可以推测, 如果要求 BTRM 对题组效应做出较为准确的估计, 需要一个题组内至少包含 5 道题目。

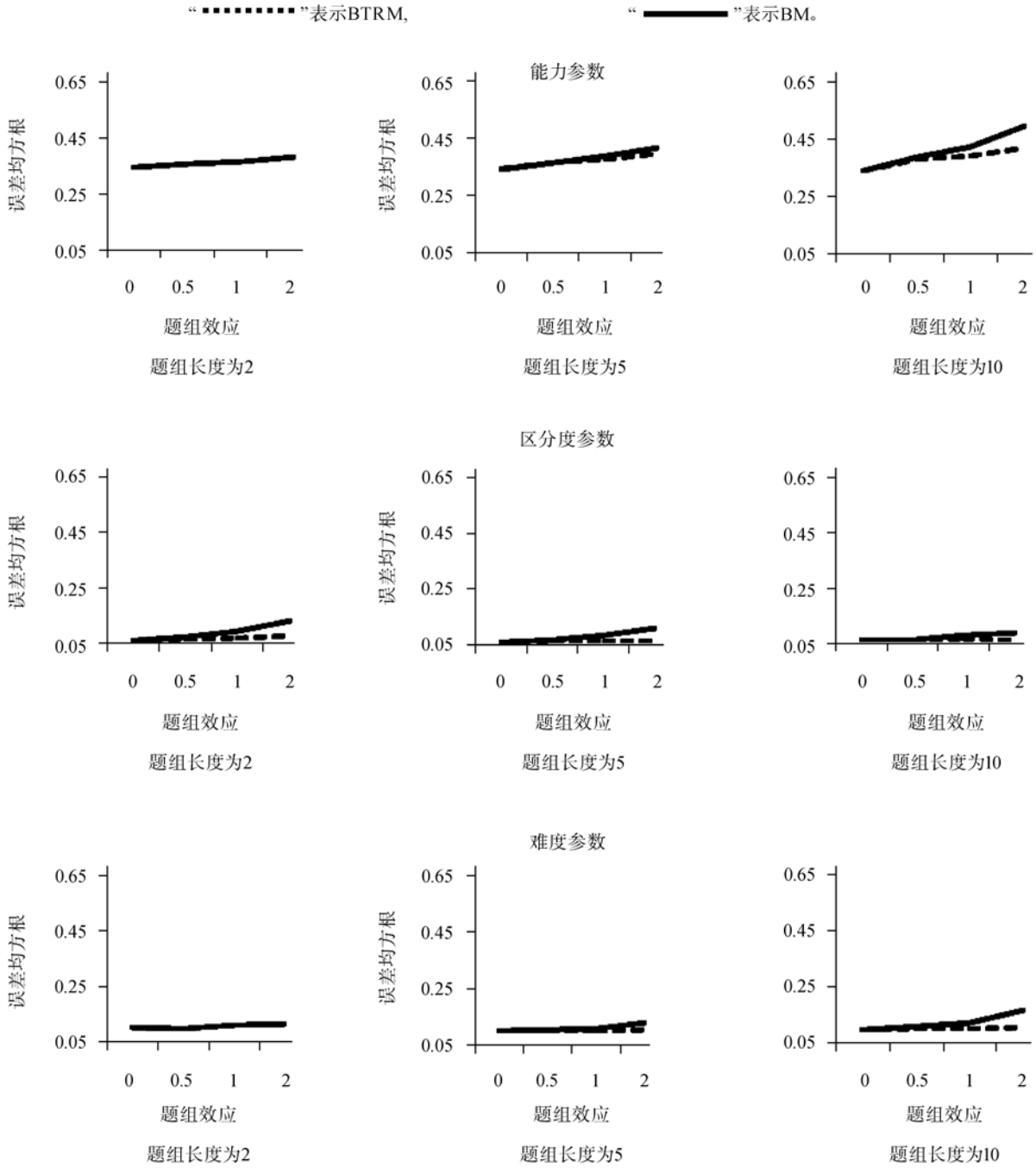

图 1　题目数量 60 条件下两种模型误差均方根(*RMSE*)

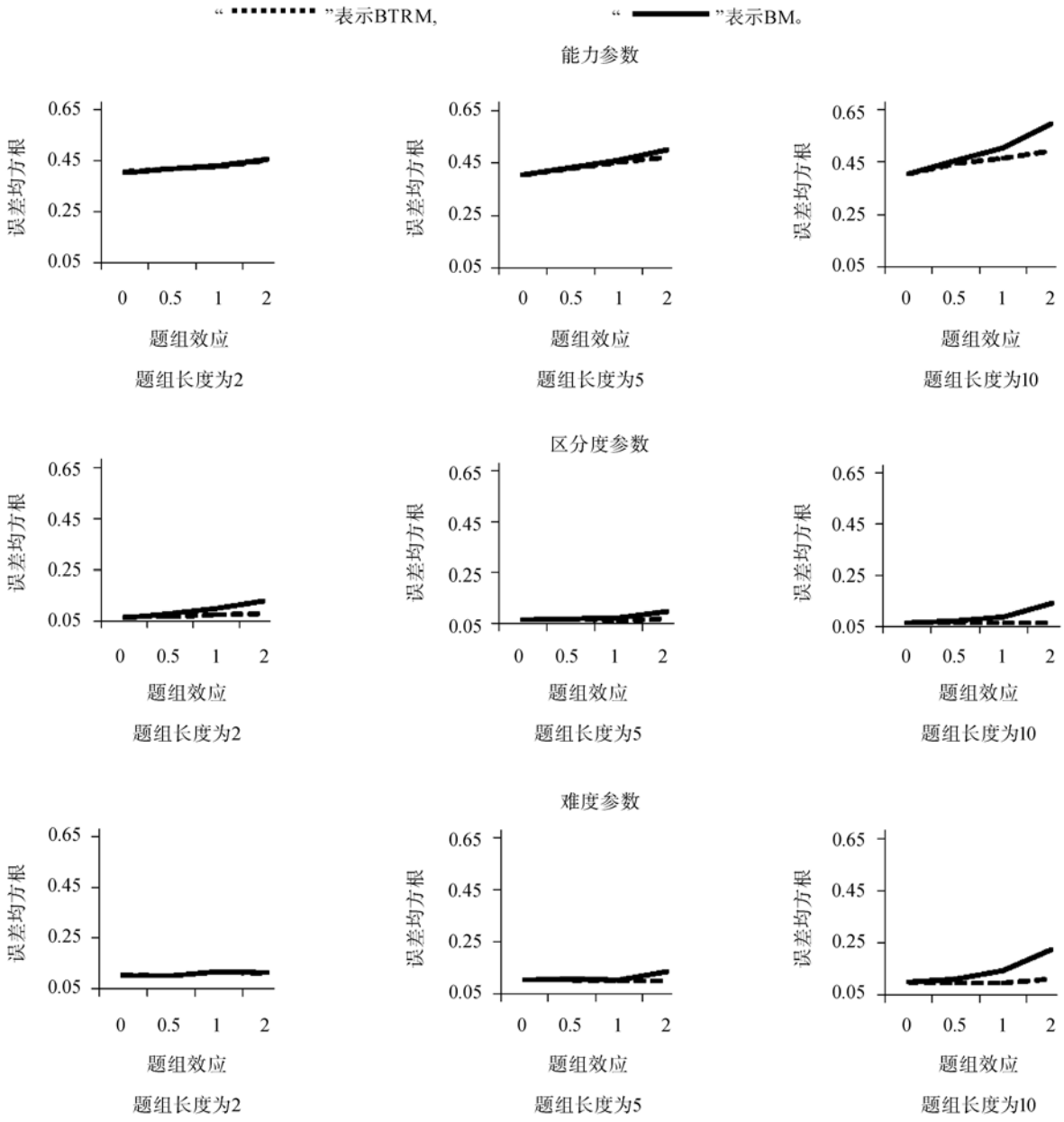

图 2 题目数量 40 条件下两种模型误差均方根(*RMSE*)

(4)能力参数

两种模型对能力参数估计准确性的差异较大，且同时受题组效应，题组长度和题目数量三个因素的影响，具体为：

与难度参数类似，分别和同时增大题组效应，题组长度两个因素，两种模型对能力参数估计的误差差异均增大，使用 BTRM 对能力参数进行估计能得到更加准确的结果。

题目数量的减小会导致两种模型对能力参数估计的准确性同时降低。特别的，当题目数量较少时(20 道)，会出现两个需要注意的结果：第一，当题组效应不存在，即题目之间满足局部独立性时，如果使用存在题组长度为 2 或 5 的题组模型对能力参数进行估计，其估计结果的误差将大于 BM；第二，相对于题目参数，两种模型对能力参数的估计受题目数量的影响更大。因此当题目数量仅为 20 时，两个模型对能力参数的估计均不准确。该结果说明，若想在题组存在的条件下使用 BTRM 得到对能力参数更准确的估计值，需要测验的长度达到一定要求，否则，即便使用 BTRM，也不能得到满意的估计结果。

3.5.2 估计值与真实值的相关 表 1 列出了各单一水平下，两种模型各参数估计值和真实值的相关系数。可以看出，在各种条件下，各参数估计值与

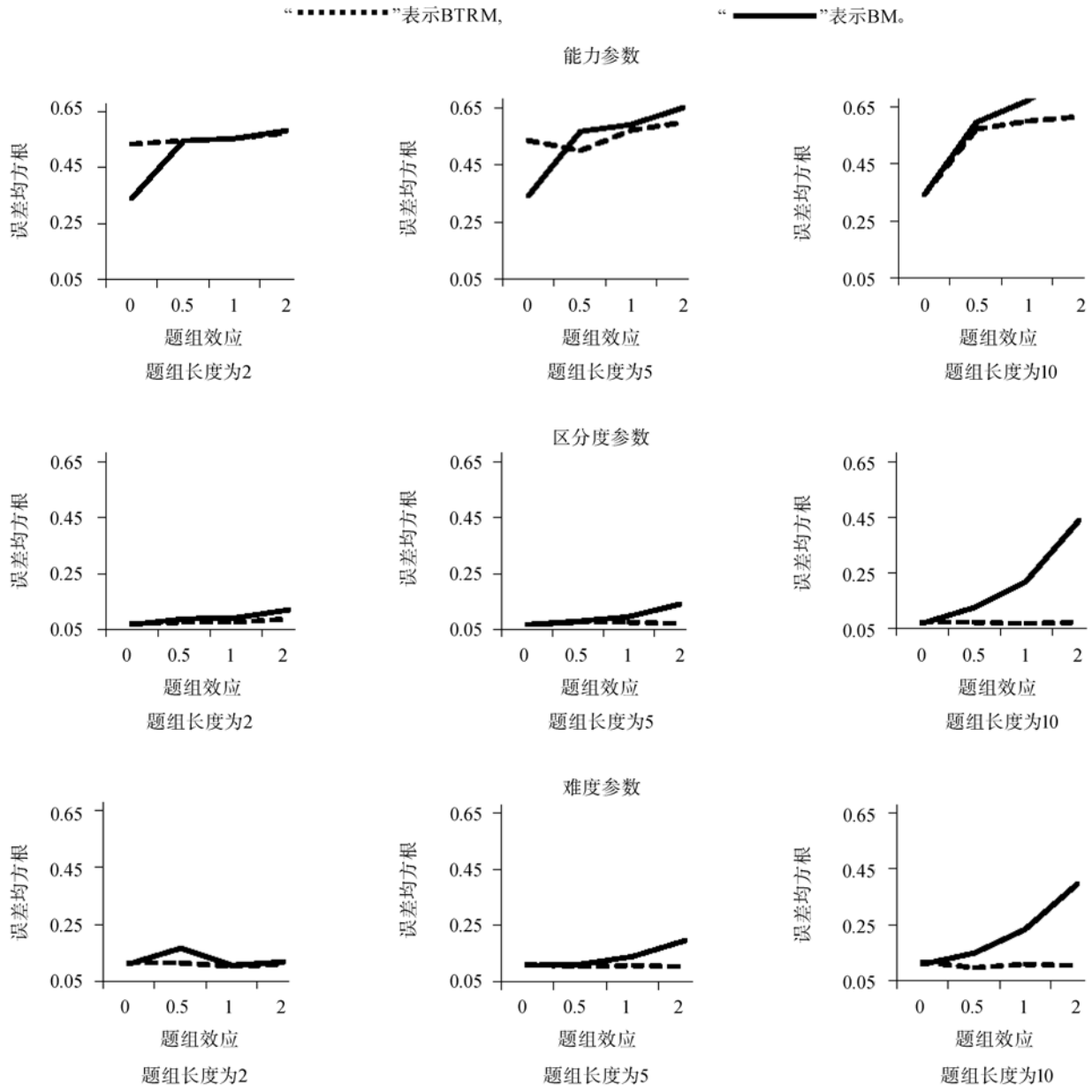

图 3　题目数量 20 条件下两种模型误差均方根(*RMSE*)

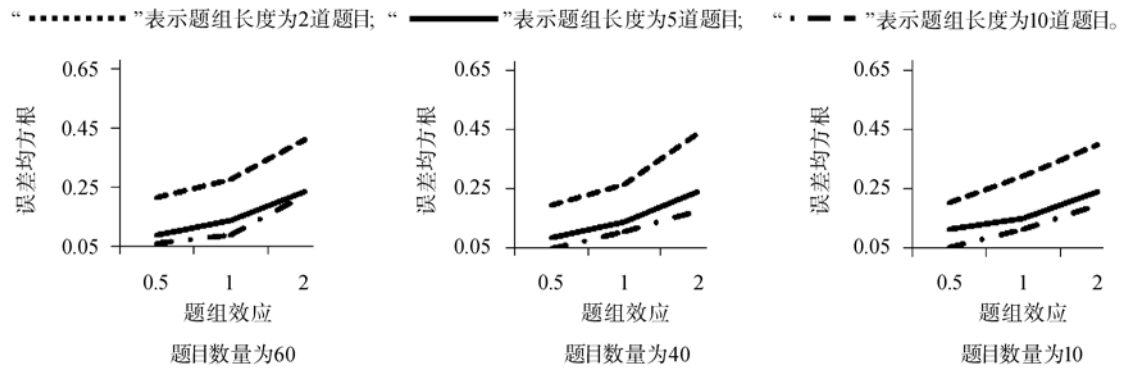

图 4　各条件下 BTRM 题组效应误差均方根(*RMSE*)

表 1 研究一各单一水平下两种模型各参数估计值和真实值的相关系数

模拟因子	因子水平	BTRM			BM		
		θ	a	b	θ	a	b
题组效应 (σ_γ^2)	0	0.917	0.947	0.995	0.933	0.949	0.996
	0.5	0.895	0.945	0.996	0.892	0.930	0.995
	1	0.887	0.944	0.995	0.877	0.902	0.993
	2	0.875	0.940	0.995	0.850	0.837	0.989
题组长度	2	0.898	0.937	0.995	0.905	0.918	0.995
	5	0.892	0.944	0.995	0.895	0.923	0.994
	10	0.894	0.951	0.995	0.874	0.895	0.990
	20	0.837	0.937	0.995	0.846	0.879	0.991
题目数量	40	0.899	0.946	0.995	0.892	0.922	0.994
	60	0.929	0.949	0.995	0.925	0.931	0.995

真实值的相关均较大(在 0.837~0.996 之间), 说明各条件下题目参数和被试参数估计值与真实值存在较高的一致性。BTRM 区分度参数估计值与真实值的相关要大于 BM, 且随着题组效应增大, 题组长度增加和测验长度减小, 两个模型得到的区分度参数的差异越大。(题目数量为 20, 题组长度为 10, 题组效应最大时, BTRM 为 0.948, BM 为 0.652)。能力参数和难度参数两个模型得到的相关系数差异相对较小, 尤其是难度参数(两种模型均在 0.990 附近)。

3.5.3 置信区间对真值的覆盖率以及区间长度

表 2 表示各条件下两种模型各参数 95%置信区间长度和对真值的覆盖比例。

从估计值 95%置信区间长度的结果可以看出, 各条件下, BTRM 各参数估计值 95%置信区间长度均大于 BM。且两种模型 95%置信区间长度的差异随着题组效应的增大而增大, 但是, 题目数量和题组长度对两种模型 95%置信区间长度的差异几乎没有影响。

结合 95%置信区间对真值的覆盖比例, 可以看出, 各条件下(不包括题组效应为零的情况), BTRM 各参数估计值 95%置信区间对真值的覆盖比例均大于 BM。这与前人研究结果一致(Bradlow et al., 1999)。参考他们的结论, 可以认为, BM 的 95%置信区间长度较 BTRM 窄, 但是其对真值的覆盖比例小于 BTRM, 说明在这些条件下, BM 可能错误的高估了后验信息。特别的, 该指标受题目数量变化的影响较大, 两种模型的差异随着题目数量的减小而增大, 且这种趋势对于区分度参数最明显。

表 2 研究一各单一水平下两种模型各参数估计值 95%置信区间长度和对真值的覆盖比例

模拟因子	因子水平	BTRM						BM					
		θ		a		b		θ		a		b	
		长度	比例	长度	比例	长度	比例	长度	比例	长度	比例	长度	比例
题组效应 (σ_γ^2)	0	1.624	0.954	0.252	0.943	0.375	0.940	1.422	0.949	0.248	0.945	0.368	0.940
	0.5	1.777	0.950	0.259	0.944	0.373	0.949	1.671	0.931	0.247	0.879	0.364	0.914
	1	1.834	0.951	0.263	0.941	0.376	0.945	1.681	0.915	0.245	0.777	0.367	0.868
	2	1.909	0.950	0.274	0.946	0.390	0.950	1.688	0.878	0.247	0.649	0.377	0.793
题组长度	2	1.778	0.953	0.283	0.939	0.389	0.948	1.726	0.945	0.242	0.801	0.379	0.934
	5	1.836	0.949	0.260	0.940	0.375	0.943	1.691	0.918	0.243	0.829	0.363	0.879
	10	1.906	0.948	0.253	0.951	0.376	0.953	1.624	0.860	0.253	0.676	0.366	0.761
题目数量	20	2.166	0.950	0.281	0.942	0.392	0.941	1.869	0.910	0.268	0.729	0.380	0.811
	40	1.729	0.952	0.256	0.944	0.374	0.948	1.611	0.922	0.240	0.855	0.364	0.908
	60	1.463	0.952	0.250	0.944	0.371	0.950	1.367	0.923	0.232	0.854	0.363	0.918

4 研究二：局部独立题目比例对题组模型影响的模拟研究

4.1 模拟设计

本研究模拟了 2000 名被试的作答反应数据。题组效应为研究一中最大的情况。题组长度限定为 5 道题目。自变量有两个：(a)题目数量：分别取 20,30,40,60 四种情况。(b)局部独立题目占全卷的比例：分别取 1/3, 1/2, 2/3 三种情况。4×3 种条件下, 分别使用 BTRM 和 BM 进行模型拟合。研究二的数

据生成, MCMC 估计和评价标准均与研究一相同。

4.2　研究结果

根据软件提供的收敛指标, 研究中所有设计因子的条件下, 两种模型均成功收敛。

4.2.1　估计的偏差、绝对偏差和误差均方根　与研究一类似, 从参数估计值的偏差结果可以看出, 在各种条件下, 两种模型对区分度参数, 难度参数, 能力参数的估计基本上是无偏的。同样, 根据对题组效应参数方差的估计偏差可知, BTRM 模型可能在多数情况下高估了题组效应。

各条件下两种模型各参数误差均方根结果见表 3。

表 3　研究二各条件下两种模型各参数估计值的误差均方根

题目数量	局部独立题目比例	BTRM				BM		
		θ	a	b	σ_γ^2	θ	a	b
60	1/3	0.422	0.066	0.103	0.256	0.445	0.129	0.126
	1/2	0.397	0.067	0.103	0.237	0.418	0.112	0.130
	2/3	0.381	0.064	0.105	0.238	0.398	0.093	0.122
40	1/3	0.490	0.073	0.109	0.275	0.518	0.107	0.150
	1/2	0.472	0.068	0.101	0.241	0.499	0.097	0.137
	2/3	0.450	0.067	0.106	0.232	0.474	0.084	0.139
30	1/3	0.552	0.071	0.105	0.276	0.588	0.102	0.175
	1/2	0.525	0.068	0.105	0.239	0.562	0.095	0.171
	2/3	0.506	0.069	0.104	0.277	0.535	0.088	0.152
20	1/3	0.657	0.080	0.127	0.440	0.702	0.131	0.228
	1/2	0.598	0.072	0.106	0.242	0.652	0.139	0.197
	2/3	0.566	0.071	0.097	0.224	0.610	0.134	0.180

(1) 区分度参数

在该研究的所有条件下, BTRM 得到的区分度参数估计精度均高于 BM。且 BTRM 对区分度参数估计的误差不受局部独立题目比例的影响, 而 BM 的估计结果在不同题目数量的条件下呈现不同的趋势。在题目数量为 30-60 的条件下, 随着局部独立题目比例增加, BM 对区分度参数估计的准确性增加, 即两种模型的对区分度参数估计准确性的差异减小。并且, 在这些条件下, 随着题目数量的增加, 在局部独立题目比例各水平, 两种模型对区分度参数估计准确性的差异增大。在题目数量为 20 道的条件下, 随着局部独立题目比例增加, BM 对区分度参数估计的准确性不变, 且两种模型估计准确性的差异最大。

另外, 在该研究限定的条件下, 两种模型对区分度参数估计的误差不受题目数量影响。

(2) 难度参数

各条件下, 两种模型对难度参数估计的结果与区分度参数呈现出基本相同的趋势。区别在于, 当题目数量为 30~60 时, 随着题目数量的增加, 在局部独立题目比例各水平, 两种模型对难度参数估计准确性的差异减小。即当题目数量较多时(60 道), 如果局部独立的题目比例较大, 两种模型对难度参数估计的准确性没有显著差异。

(3)题组效应参数

当题目数量为 30~60 时, 题组效应参数方差估计的准确性基本不受局部独立题目比例的影响。但当题目数量为 20 时, 如果局部独立题目的比例仅为 1/3, 题组效应参数方差的误差远大于其他情况。可以推测, 如果题目数量较少(20 道), 局部独立性的比例为 1/3 时, BTRM 对题组效应不能得到准确的估计。

(4)能力参数

所有条件下, BTRM 得到的能力参数估计精度均高于 BM。两种模型对能力参数估计准确性同时受局部独立题目比例和题目数量两个因素的影响, 具体为:

随着局部独立题目比例增加, 两个模型对能力参数估计的误差均减小。且当题目数量为 30~60 时, 两个模型对能力参数估计误差的差异也减小。说明局部独立题目所占比例越多, 即使存在较大的题组效应, 使用 BM 得到的能力估计值误差也不会远大于 BTRM。当题目数量为 20 时, 两个模型对能力参数估计误差的差异不受局部独立题目比例影响。

随着题目数量的减少, 两个模型对能力参数估计的误差均增加, 且其差异也增加。结合两个因素考虑可以发现, 随着题目数量减少, 在局部独立题目比例各水平, 两个模型对能力参数估计精度的差异增加。特别的当题目数量为 20 时, 两个模型对能力参数估计的误差明显变大, 且它们的差异也达到最大, 在局部独立题目比例的各个水平基本稳定不变。

4.2.2 估计值与真实值的相关 对两种模型各参数估计值和真实值的相关系数进行考察, 可以看出, 在各种条件下, 各参数估计值与真实值的相关均较大(在 0.721~0.996 之间), 说明各条件下题目参数和被试参数估计值与真实值存在较高的一致性。BTRM 各参数估计值与真实值的相关要大于 BM。两个模型得到的区分度参数的差异最大。两个模型得到的能力参数和难度参数相关系数差异相对较小, 尤其是难度参数差异最小(两种模型均在 0.990 附近)。

4.2.3 置信区间对真值的覆盖率以及区间长度 根据两种模型各参数 95%置信区间长度和对真值的覆盖比例的结果, 可以看出:

与研究一类似, 各条件下 BTRM 各参数估计值 95%置信区间长度均大于 BM, 而 BTRM 各参数估计值 95%置信区间对真值的覆盖比例均大于 BM。这印证了研究一中 BM 高估了后验信息的结论。且随着局部独立题目比例的增加, 两种模型各参数估计值 95%置信区间长度的与 95%置信区间对真值的覆盖比例的差异均减小, 题目数量对这些指标无显著影响。

5 讨论

研究一模拟设计中三个影响因子的设计参考了相关研究, 因此, 它们对两种模型参数估计结果的影响, 可以与前人研究进行比较。研究二中新加入了局部独立题目比例这一因子, 可以为两种模型的比较提供新的信息。

首先, BTRM 对各参数的估计结果几乎不受题组效应的影响, 而 BM 参数估计误差受题组效应影响更大。这与前人的模拟研究结果一致(Bradlow et al., 1999; Wang et al., 2002)。其中, BM 中区分度参数的变化可以由模型的定义和参数的意义来解释。区分度参数表示了潜在能力特质与题目之间的相关, 未加入题组效应参数的模型将题目之间的相互依赖性视作“干扰”。如果题组效应越大, 则“干扰”越大, 那么能力特质与题目的相关就越小, 即区分度参数估计的误差越大。而在 BTRM 中, 题组效应参数控制了局部依赖性产生的“干扰”, 提高了能力特质与题目的相关, 相对减小了区分度参数估计的误差。从而, 提高了整个模型对数据的拟合程度。

其次, 对于题组长度这一因素, BTRM 对各参数的估计结果基本稳定, 几乎不受它的影响。而对 BM, 除了区分度参数的估计误差在不同的题目数量水平上的变化趋势有所不同之外, 其他参数估计结果的准确性随着题组长度的增加而减小。因此, 当题组较长时, 使用 BTRM 对各参数估计的误差要大大小于 BM。这也与前人研究结果有相似之处(Bradlow et al., 1999)。特别的, 随着题组长度的增加, BTRM 题组效应参数估计值的误差减小。说明题组长度越长, 同一个题组内用于估计题组效应参数的信息量就越大, 模型对题组内题目相依性的信息把握的越准确, 参数估计的效果就越好。随着题组效应参数估计准确性的增加, 整个模型对局部依赖性产生的“干扰”的控制力就越强, 因此, 反映题目和潜在能力特质关系的区分度参数估计的误差也较稳定。这也是造成 BTRM 区分度参数估计的误差基本不受题组长度影响的原因。另外, 当题组长度为 5 和 10 时, 题组效应参数估计的准确性明显大于题组长度为 2 的情况。可以推测, 只有当题组长度为 5 以上时, 使用 BTRM 才能得到较为准确的题组效应参数估计。

第三, 相对于以往研究, 题目数量是本研究新增的一个研究因子。对于题目参数, BTRM 的估计误差几乎不受题目数量影响, 而 BM 的估计误差随着题目数量的减小有明显的增大。因此, 随着题目数量的减少, 在其他各水平上两种模型对题目参数估计值误差的差异增加。对于被试参数, 与多数关于项目反应模型的模拟研究结果一致, 题目数量越少, 估计的准确性就相应降低。针对研究一中题目数量从 40 降为 20 后, 能力参数估计误差有明显增加的现象, 在研究二中加入了题目数量为 30 的情况, 期望找到误差急剧增加的关键点。结果发现, 从题目数量 30 到 20, 能力参数误差增加的幅度大于其他水平。因此可以认为, 题目数量为 20 道的条件下, 两种模型均不能对被试能力作出准确的估计。在该条件下, 如果存在题组效应, 即使选用了 BTRM, 也不能得到较为准确的估计结果。并且, 如果题目间相互独立, 却错误的使用了长度为 2 或者 5 的题组模型, BTRM 对被试能力参数估计反而不如传统 BM, 即 BTRM 的普遍性受到破坏。

第四, 局部独立题目的比例也是本研究新增的研究因子。相对于题目参数, 能力参数受该因子的影响较大。总的来说, 当题目数量为 30~60 时, 随着局部独立题目比例的增加, 两种模型对参数估计准确性都增加, 且它们的差异减小。说明当试卷中局部独立题目的比例较大时, 即使存在较大的题组效应, 模型的选取对各参数估计结果的准确性不会有显著的影响。当题目数量为 20 时, 随着局部独立题目比例的增加, 两种对参数估计准确性的差异不变, 且达到较大水平。说明题目数量较小时(20 道), 如果题组效应较大, 无论局部独立题目的比例增加多少, 使用 BM 仍会对参数估计带来较大误差。

最后, 本研究使用 SCORIGHT 软件实现贝叶斯方法的 MCMC 估计, 每条 MCMC 链设置迭代次数为 4000。这一设置参照了 SCORIGHT 手册中的举例说明。SCORIGHT 在 output 的 convergence 文件中, 提供了模型收敛的诊断结果。该软件使用的是 Gelman 和 Rubin (1993)的 F 检验作为收敛标准。根据该标准, 本研究的所有估计均成功收敛。但是, 在使用 MCMC 方法的研究中, 可能由于估计软件先验设置或者程序算法的差异, 导致对迭代次数的要求不尽相同。因此, 本研究的所有结果都基于 SCORIGHT 软件下的这种设置, 如果改变 MCMC 估计的设置, 或者改变估计软件, 可能会得到不同的结果。

本研究探讨了题组模型的适用范围和影响因子, 对于测验的编制和分析具有一定的实践意义。一方面, 对于测验编制来说, 首先应当尽量保证题目之间局部独立。其次, 如果测验必须含有题组, 为了避免局部依赖情况下, 使用传统 IRT 模型带来的误差, 最好采用较短的题组。另外, 应尽量加大局部独立题目的比例。因为当题组长度较短, 局部独立题目比例较大时, 题组模型和局部独立模型之间的差异相对较小; 另一方面, 对于测验数据分析来说, 应当建立一套科学的操作流程。首先对题目之间是否存在局部依赖性进行检验, 检验方法可以采用标准曲线拟合方法, *Q3* 方法和 *p* 方法(Chen & Wang, 2007; Li, Bolt, & Fu, 2006)。其中 *Q3* 方法由于计算简单, 因此通常被作为识别局部依赖性的指标。然后根据检验结果, 选用适当的模型。如果检验发现题目之间存在较强的局部依赖性, 则应当选用题组模型以避免传统 IRT 模型对参数估计的偏差。

本研究还存在一定的局限性。首先, 本研究采用模拟数据的方法, 为控制无关因素加入了人为限制, 例如本研究将同一测验中每个题组内的题组效应限制为相等, 而实际中各题组的题组效应可能不同。因此, 本研究结果不一定能直接适用于实际情况, 建议实际应用者对结论的推广应当更加谨慎。其次, 本研究只关注两参数 Logistic 模型扩展的题组随机效应模型, 没有对 Rasch 或者三参数 Logistic 模型的题组随机效应模型进行讨论。有必要在以后的研究中对这些模型的适用条件进行探讨。然后, 实际测验中一套试卷可能同时包含了 0,1 计分的题目和分步计分的题目, 而本研究只考虑了 0,1 计分的题目。可以参照 Wang 等(2002)提出的 General 贝叶斯题组模型对含有题组的混合题目进行参数估计。最后, 本研究通过模拟实验讨论了贝叶斯题组随机效应模型在不同条件下的表现, 但是没有对题组效应产生的原因及其影响因素进行考察。Wainer 等人(2007)提出的含协变量的题组模型就能很好的回答上一问题。今后的研究可以结合这一模型对题组效应的成因进行探讨。

6　结论

本研究设置了多种可能影响题组模型表现的条件, 这些条件与现实中教育考试的情况有很大的相似性。对这些条件下 BTRM 和 BM 进行对比, 有助于探讨有必要使用 BTRM 的具体范围, 为实践中测验的编制和数据分析中模型的选择提供了初步的参考和依据。本研究的主要结论如下:

第一, 在本研究设置的各个变化条件下, BTRM 相对于 BM 表现出更大的优势。即使在题组效应不存在的情况下 BTRM 参数估计结果的准确性几乎与 BM 相同, 说明该模型具有很好的普遍性。

第二, BTRM 不受题组效应和题组长度因素的影响, 而 BM 对参数估计的误差随之增加。因此当局部依赖性严重, 尤其测验结构中存在较长题组时, 应当选用 BTRM 以得到更加准确的参数估计结果。

第三, 题目数量对两种模型各参数估计准确性的差异影响很大。当题目数量少时(20 道), BTRM 不仅其普遍性受到破坏, 并且即使存在题组效应, 其参数估计的准确性也不理想。说明要使用 BTRM 得到准确的估计结果, 题目数量至少要大于 20 道。

第四, 局部独立题目的比例越大, 两种模型估计结果越好, 且二者的差异越小。当题目数量达到 30 以上时, 如果保证局部独立题目比例较大, 即使存在题组效应, 也可以选用 BM。

第五, 在测验编制阶段, 最好保证题目的局部

独立性, 否则最好采用较短的题组并且增加局部独立题目的比例。在测验数据分析阶段, 应当首先检验题目是否存在局部依赖性, 再根据检验结果选择适合的模型, 减少对各参数估计的误差。

参 考 文 献

Bradlow, E. T., Wainer, H., & Wang, X. H. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153–168.

Chen, C. T., & Wang, W. C. (2007). Effects of ignoring item interaction on item parameter estimation and detection of interacting items. *Applied Psychological Measurement, 31*(5), 388–411.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43*(2), 145–168.

Gelman, A., & Rubin, D. B. (1993). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–472.

Lee, G., Dunbar, S. B., & Frisbie, D. A. (2001). The relative appropriateness of eight measurement models for analyzing scores from tests composed of testlets. *Educational and Psychological Measurement, 61*(6), 958–975.

Li, Y. M., Bolt, D. M., & Fu, J. B. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*(1), 3–21.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Reeding, MA: Addison-Wesley.

Rijmen, F. (2009). Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison. *Tech. Rep*. No. RR-09-37). Educational Testing Service.

Wainer, H., Bradlow, E. T., & Wang, X. H. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Wainer, H., & Wang, X. H. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203–220.

Wang, W. C., & Wilson, M. (2005a). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*(4), 296–318.

Wang, W. C., & Wilson, M. (2005b). The Rasch testlet model. *Applied Psychological Measurement, 29*(2), 126–149.

Wang, X. H., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*(1), 109–128.

Wang, X., Bradlow, E. T., & Wainer, H. (2004). User's Guide for SCORIGHT (Version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis. *Research report-educational testing service PRINCETON RR, 4*.

When Should We Use Testlet Model? A Comparison Study of Bayesian Testlet Random-Effects Model and Standard 2-PL Bayesian Model

LIU Yue; LIU Hong-Yun

(*School of Psychology*, *Beijing Normal University*, *Beijing* 100875, *China*)

Abstract

A testlet is comprised of a group of multiple choice items based on a common stimuli. When a testlet is used, the traditional item response models may not be appropriate due to the violation of the assumption of local independence (LI). A variety of new models have been proposed to analyze response data sets for testlets. Among them, the Bayesian random effects model proposed by Bradlow, Wainer and Wang (1999) is one of the most promising. However, in many situations it is not clear to practitioners whether the traditional IRT methods should still be used instead of a newly proposed testlet model.

The objective of the current study is to investigate the effects of model selection in various situations. In simulation 1, simulated response data sets were generated under three simulation factors, which were: testlet variance (0, 0.5, 1, 2); testlet size (2, 5, 10); and test length (20, 40, 60). For each simulation condition, the test structure was determined by fixing the number of examinees as I =2000, and the percentage of testlet items in a test as 50%. Under each condition, 30 replications were generated. Both two-parameter Bayesian testlet random effect model and standard two-parameter Bayesian model were fitted to every dataset using MCMC method. The computer program SCORIGHT was used to conduct all the analysis across different conditions.

Two models were compared corresponding to seven criteria: bias, mean absolute error, root mean square error, correlation between estimated and true values, 95% posterior interval width, 95% coverage probability.

These indexes were computed for all parameters separately.

Simulation 2 compared the two models under two factors: the proportion of independent items (1/3, 1/2, 2/3); test length (20, 30, 40, 60). The data generation, analyze process and criteria mimicked those of simulation 1.

The results showed that: (1) The accuracy of the estimation of all parameters under 2-PL Bayesian testlet random-effect model remained stable with varying levels of testlet effect and testlet size. However, the estimate errors of all the parameters under 2-PL Bayesian model increased dramatically as the testlet effect and testlet size became larger. Besides, using Bayesian testlet random-effect model, the error for every parameter was always less than that for 2-PL Bayesian model. It was especially necessary to choose 2-PL Bayesian testlet random-effects model when testlet variance and testlet size were large. (2) Even though, the accuracy of estimation of item parameters in Bayesian testlet random-effect model wasn't affected by test length, the accuracy of ability parameter was. Moreover, as the test got shorter, the errors of all parameters under 2-PL Bayesian model increased dramatically. In all, under short test conditions, even if there was large testlet effect, Bayesian testlet random-effect model couldn't work well, meanwhile, if items were all independent, using Bayesian testlet random-effect model would result in much worse ability estimations than 2-PL Bayesian model. (3) When the proportion of independent items was large, and the test length was larger than 20 items, the estimations of two models didn't show significant differences.

In conclusion, 2-PL Bayesian testlet random-effect model is more general. Using the more complex testlet model when items are all independent, will lead almost the same accuracy of the parameter estimations as using the 2-PL Bayesian model. It is better to choose 2-PL Bayesian testlet random-effect model when testlet variance, testlet size, and test length are large. However, when test length is short, even the Bayesian testlet random-effects model couldn't provide accurate estimations of parameters when local dependence happened. So it is important to make sure the test was comprised of enough items before applying a testlet model. We also give some suggestions for practitioners. In the test construction period, first it is better for items to be independent, if not, shorter testlets and larger proportion of independent items should be included. While in the test analysis period, local dependence should be detected first. If evidence shows that there is dependence structure, then an appropriate model should be chosen to avoid estimation errors.

Key words　testlet; 2-PL Bayesian testlet random-effect model; 2-PL Bayesian model; MCMC method